frontiers | Research Topics

Implementation of AI and machine learning technologies in medicine

Edited by
Enrico Capobianco, Pietro Lio', Yuguang Wang, Jingjing You, Chris Hodge and Zhe He

Published in
Frontiers in Medicine
Frontiers in Big Data
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-3263-8
DOI 10.3389/978-2-8325-3263-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Implementation of AI and machine learning technologies in medicine

Topic editors

Enrico Capobianco — Jackson Laboratory, United States
Pietro Lio' — University of Cambridge, United Kingdom
Yuguang Wang — Shanghai Jiao Tong University, China
Jingjing You — The University of Sydney, Australia
Chris Hodge — The University of Sydney, Australia
Zhe He — Florida State University, United States

Citation

Capobianco, E., Lio', P., Wang, Y., You, J., Hodge, C., He, Z., eds. (2023). *Implementation of AI and machine learning technologies in medicine*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3263-8

Table of contents

frontiers in Medicine

BRIEF RESEARCH REPORT
published: 08 October 2021
doi: 10.3389/fmed.2021.743476

Heat Shock Proteins in Urine as Cancer Biomarkers

Zarema Albakova[1], Diogo Dubart Norinho[2], Yana Mangasarova[3] and Alexander Sapozhnikov[1,4]*

[1] Department of Biology, Lomonosov Moscow State University, Moscow, Russia, [2] Data Science Department, NOS SGPS, Porto, Portugal, [3] National Research Center for Hematology, Moscow, Russia, [4] Department of Immunology, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry of Russian Academy of Sciences, Moscow, Russia

Heat shock proteins (HSPs) are a large family of molecular chaperones, which have shown to be implicated in various hallmarks of cancer such as resistance to apoptosis, invasion, angiogenesis, induction of immune tolerance, and metastasis. Several studies reported aberrant expression of HSPs in liquid biopsies of cancer patients and this has opened new perspectives on the use of HSPs as biomarkers of cancer. However, no specific diagnostic, predictive, or prognostic HSP chaperone-based urine biomarker has been yet discovered. On the other hand, divergent expression of HSPs has also been observed in other pathologies, including neurodegenerative and cardiovascular diseases, suggesting that new approaches should be employed for the discovery of cancer-specific HSP biomarkers. In this study, we propose a new strategy in identifying cancer-specific HSP-based biomarkers, where HSP networks in urine can be used to predict cancer. By analyzing HSPs present in urine, we could predict cancer with approximately 90% precision by machine learning approach. We aim to show that coupling the machine learning approach and the understanding of how HSPs operate, including their functional cycles, collaboration with and within networks, is effective in defining patients with cancer, which may provide the basis for future discoveries of novel HSP-based biomarkers of cancer.

Keywords: heat shock proteins, biomarkers, cancer, urine, machine learning

OPEN ACCESS

Edited by:
Pietro Lio,
University of Cambridge,
United Kingdom

Reviewed by:
Haihui Pan,
George Washington University,
United States
Wenyue Su,
Western University of Health Sciences, United States

****Correspondence:***
Zarema Albakova
zarema.albakova14@gmail.com

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *18 July 2021*
Accepted: *06 September 2021*
Published: *08 October 2021*

Citation:
Albakova Z, Norinho DD, Mangasarova Y and Sapozhnikov A (2021) Heat Shock Proteins in Urine as Cancer Biomarkers. Front. Med. 8:743476. doi: 10.3389/fmed.2021.743476

INTRODUCTION

Heat shock proteins (HSPs) are molecular chaperones that are classified into families such as HSP70, HSP90, HSP40, HSPB, HSP110, and chaperonins (1). Members of HSP families are located in different cellular compartments such as cytosol, nucleus, lysosome, endoplasmic reticulum, and mitochondria (1–3). Several studies reported high levels of HSP70, HSP90, HSP40, HSPB, and chaperonins in plasma, serum, and plasma-/urine-derived exosomes of the patients in different types of cancer compared to healthy individuals (3–15). This has opened new perspectives on the use of HSPs as biomarkers of cancer. However, abnormal expression of HSPs has also been observed in several other pathologies including cardiovascular and neurodegenerative diseases (16–18). For example, Li and his colleagues showed that high expression of HSP70 in plasma positively correlated with heart failure (19). Therefore, new strategies should be used for the identification of cancer-specific HSP biomarkers. Since HSPs are tightly linked to the stress response, level of individual HSP members in the clinical samples may not be enough for precise prediction of cancer. Herein, we used a machine learning approach for the identification of HSP-based urine biomarkers

of cancer. We show that coupling machine learning approach and the understanding of how HSPs operate in networks may be effective in diagnosing cancer. To the best of our knowledge, this is the first study that explores HSP secreted in urine for prediction of cancer and the primary study to assess the relationships between different HSP networks and cochaperones for the discovery of clinically useful HSP-based biomarkers of cancer.

METHODS

We used publicly available mass spectrometry dataset that contains samples from 231 donors (20). Urine samples were derived from the patients with gastric cancer (GC) ($n = 47$), esophageal cancer (EC) ($n = 14$), lung cancer (LC) ($n = 33$), bladder cancer (BC) ($n = 17$), cervical cancer (CCA) ($n = 25$), colorectal cancer (CRC) ($n = 22$), and benign lung diseases (LDs) such as chronic obstructive pulmonary disease (COPD) ($n = 17$) and pneumonia (PM) ($n = 23$) as well as from the healthy volunteers (Control, CTL) ($n = 33$) (20). Urine samples were centrifuged at 200,000 g for 70 min and absolute protein amounts were measured by liquid chromatography with tandem mass spectrometry (LC-MS/MS) and presented as intensity-based fraction of total (iFOT; displayed in 10^5) representing normalized intensity for each protein (20). HSPs such as HSP70, HSP90, HSP40, HSP27, HSP110, chaperonins, and cochaperones were included in the analysis (**Supplementary Table 1**). Proteins that have > 30% of 0.0099 (missing values) were excluded from the analysis.

The expression level for each protein was measured for CTL and six groups of cancers (LC, BC, CCA, CRC, EC, and GC). Since the data were not normally distributed, nonparametric tests were used. The procedure was divided into two stages such as the Kruskal–Wallis (KW) test for all the proteins followed by a *post-hoc* Dunn's test using CTL as reference (21). Bonferroni multiple comparison test (MCT) correction in its multistep variant, known as Holm–Bonferroni correction, was also used (22).

The cancer prediction model was trained on HSP and their cochaperones to isolate their effects in cancer prediction. Taking into account that HSPs are located in different cellular compartments as well as exist in different forms (constitutive/stress-inducible) and require cochaperones for their functional cycles, while also working in networks, we introduced into the model various combinations of simple ratios and multiplication strategies. For example, to isolate the effect of HSP90 homologs, we used the relationship between the level of cytosolic HSP90 homolog to the level of mitochondrial HSP90 homolog in a simple ratio of HSP90AA1/TRAP1, constitutive HSP90 isoform to stress-inducible HSP90 in a simple ratio of HSP90AB1/HSP90AA1, cochaperone level to the HSP90α level in a simple ratio of FKBP4/HSP90AA1, etc. (**Supplementary Table 2**). As a result, a cancer prediction model was created using XGBoost with a tree booster. A binary classification model was built to discriminate the cancer patients (LC, BC, CCA, CRC, EC, and GC) from the non-cancer group (LD and CTL). The performance of the method was evaluated through 10-fold stratified cross-validation. By splitting the data into 10-fold, iteratively training in 9-fold and testing on the remaining fold, we mimic the effect of 10 distinct datasets. This enables us to estimate the generalization error of our model and prevent overfitting, therefore ensuring that the model would generalize well to new data. Bayesian optimization was used to tune hyperparameters. We computed features importance using the gain metric, which measures the loss reduction of adding a split with that feature. Let ξ_l be the set of features at the l^{th} step tuning:

1. Start the first iteration with all the features (ξ_1).
   a. Initialize the Bayesian optimization:
      i. Randomly, select n_1 points $\{\phi_1, \ldots, \phi_{n_1}\}$ located within user defined boundaries:
         1. Train with hyperparameter set ϕ_i and evaluate the model using K-fold cross-validation with log-loss.
   b. Perform the Bayesian optimization:
      i. Sequentially, select n_2 points:
         1. ϕ_j is the point that maximizes the upper confidence bound of the posterior distribution of the Gaussian process by given the data points $\{\phi_1, \ldots, \phi_{j-1}\}$ for $j > n_1$.
   c. Of the $n_1 + n_2$ combinations tried, select the set of hyperparameters that minimize the log-loss such that $\Theta_1 = argmin_{\{\phi_1, \ldots, \phi_{n_1+n_2}\}}$ *log loss*.
   d. For each of the K models with parameters Θ_1 trained in the K-fold cross-validation, extract the feature importance and then compute the average for each feature.
   e. Remove all the features whose importance is equal to the minimum.
2. For iteration l:
   a. Initialize the Bayesian optimization and randomly select n_1 new points.
   b. Probe all $\{\Theta_1, \ldots, \Theta_{l-1}\}$ the points.
   c. Perform the Bayesian optimization by sequentially selecting n_2 points.
   d. Select $\Theta_l = argmin_{\{\phi_1, \ldots, \phi_{n_1+n_2+l-1}\}}$
   e. Perform feature selection
   f. Stop if there is only one feature left or all the features have the same importance, otherwise, continue
3. Stop when reach zero feature.
4. Select ξ_k, Θ_k corresponding to the minimum log loss across all the iterations.

RESULTS

Heat shock proteins and cochaperones including HSP90AB1, TRAP1, FKBP4, HSPA9, HSPB5, CCT1, and CCT5 were identified as differentially expressed proteins (**Table 1**). CCT1, CCT5, and FKBP4 showed significantly lower expression in the cancer patients compared to the healthy volunteers, whereas

TABLE 1 | Differentially expressed HSPs and cochaperones in the urine of the cancer patients compared to healthy volunteers by Dunn's test with Holm–Bonferroni correction.

Cancer type	CCT1		CCT5		FKBP4		HSPB5		HSP90AB1		HSPA9		TRAP1	
	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value
LC	−3.6	1.69E-03	−4.1	2.41E-04	−5.0	4.04E-06	0.25	1.00	−2.2	0.119	4.3	1.02E-04	3.5	2.28E-03
BC	−1.9	0.125	−2.9	1.02E-02	−4.1	1.36E-04	−0.56	1.00	0.61	1.00	4.3	8.80E-05	3.5	2.76E-03
CCA	−3.9	5.92E-04	−3.4	2.37E-03	−2.7	1.19E-02	2.7	4.07E-02	−0.68	1.00	1.9	7.84E-02	2.8	1.70E-02
CRC	−3.4	2.47E-03	−3.8	6.56E-04	−4.3	8.05E-05	−1.9	0.278	−2.7	3.77E-02	2.1	7.84E-02	1.9	0.119
EC	−2.8	1.53E-02	−2.3	4.77E-02	−0.84	0.402	−1.1	1.00	0.58	1.00	3.1	7.18E-03	3.3	3.69E-03
GC	−1.4	0.164	−1.2	0.216	−3.4	1.91E-03	0.76	1.00	−2.7	3.77E-02	2.2	7.27E-02	3.4E-02	0.973

LC, lung cancer; BC, bladder cancer; CCA, cervical cancer; CRC, colorectal cancer; EC, esophageal cancer; GC, gastric cancer; HSPs, heat shock proteins.

HSPA9 and TRAP1 showed a significantly higher expression in patients with cancer compared to the control group for the most cancer types. HSPB5 showed significantly higher expression only in the CCA patients compared to the healthy volunteers (**Table 1**). HSP90AB1 showed a significantly lower expression in the patients with GC and CRC compared to CTL (**Table 1**).

Remarkably, the cancer prediction model trained on HSPs and cochaperones resulted in 90% precision and a balanced accuracy of 84.61% (accuracy of 87.041%) averaged over the 10 cross-validation test folds (**Figure 1A**). In order to identify proteins, which positively contributed to the cancer prediction model, we have implemented the Shapely Addictive Explanations (SHAP) approach. Low levels of HSP90AB1/TRAP1, HSPA6/TRAP1, and HSP90AA1/TRAP1 in urine increase the probability of the patient having cancer, whereas low levels of CCT2/HSP90AB1 and HSPB1*HSPA9 in urine are strongly associated with non-cancer groups (**Figure 1C**). In order to assess the differences in the level of HSPs across different types of cancer, we constructed a heatmap, representing the z-score of HSPs for each patient (**Figure 1B**). HSP90AA1 and HSPD1 showed to be highly expressed in BC; HSPB1 and HSBP5 in CCA; ST13, DNAJA1, and HSPA8 in LC; FKBP4 and HSPA8 in EC (**Figure 1B**). HSPA2 and HSPA4 did not seem to be affected in different types of cancer (**Figure 1B**).

Higher levels of both constitutive and stress-inducible HSP90 isoforms in relationship to mitochondrial HSP90 isoform TRAP1 are associated with benign lung diseases such as PM and COPD, whereas a higher level of TRAP1 to HSP90AA1 and HSP90AB1 is associated with lung cancer (**Figures 1D,E**). In contrast to patients with PM, a low level of CCT5 and high levels of HSPA9*TRAP1 and CCTs/HSP90AA1 are associated with LC (**Figure 1E**; **Supplementary Figure 1A**). Furthermore, lower expression of HSP90AA1/TRAP1 and HSP90AB1/TRAP1 positively contributed to LC compared to higher expression of HSP90AA1/TRAP1 and HSP90AB1/TRAP1 in the COPD patients (**Figure 1E**; **Supplementary Figure 1B**). Overall, urine samples contain cancer-specific HSP signatures. Therefore, these HSP signatures may be used to distinguish cancer from noncancer patients and patients with benign disease as well as they may be further used to identify specific types of cancer; however, this requires further investigation.

DISCUSSION

Heat shock proteins are ubiquitously expressed as molecular chaperones, which support tumor growth and survival (23). Cells possess various families of HSPs with distinct functions, often working in collaboration to perform proper folding and degradation of client proteins (24, 25). Several studies reported altered expression of HSPs in malignant cells compared to their normal cell counterparts (3–15). Furthermore, overexpression of HSPs has been linked with tumor aggressiveness, metastasis, and poor prognosis (2, 24, 26–29). In this study, we aimed at exploring the potential of HSPs in urine as biomarkers of cancer. We showed that HSP chaperone networks can be used to predict cancer with ∼90% precision in 10-fold cross-validation. We highlighted that understanding of HSP chaperone system and the notion of how HSPs operate are critical for prediction of cancer.

Our approach started with an identification of differentially expressed HSP proteins in different types of cancer compared to healthy volunteers. We showed that different HSP members are up- and down-regulated in different types of cancer, suggesting that a specific type of cancer has distinct HSP signatures (**Table 1**). We then developed a cancer prediction model, which reflected the way how HSP chaperone networks work. The model is based on the notion that HSP networks work in collaboration with each other as well as with cochaperones and that there also may be some shift in the proportion of different HSP homologs in the cancer patients compared to the healthy individuals and the benign patients, leading to all of these changes being captured by machine learning approach. Using this approach, we could predict cancer with 90% precision (**Figure 1A**). Furthermore, our cancer prediction model could discriminate between various types of cancer based on the expression of distinct HSPs in urine samples, which may help in diagnosing specific subtypes of cancer among a heterogeneous group of tumors, such as lymphoma or breast cancer. In this regard, Klimczak et al. (30) used The Cancer Genome Atlas and KM plotter databases to show that expression of six HSPs including *HSPA2, DNAJC20, HSP90AA1, CCT1, CCT2,* and *CCT6A* can be used to predict prognosis in patients with breast cancer (30). Furthermore, upregulation of distinct HSPs was associated with either estrogen receptor-positive, progesterone receptor-positive, or human epidermal growth factor receptor

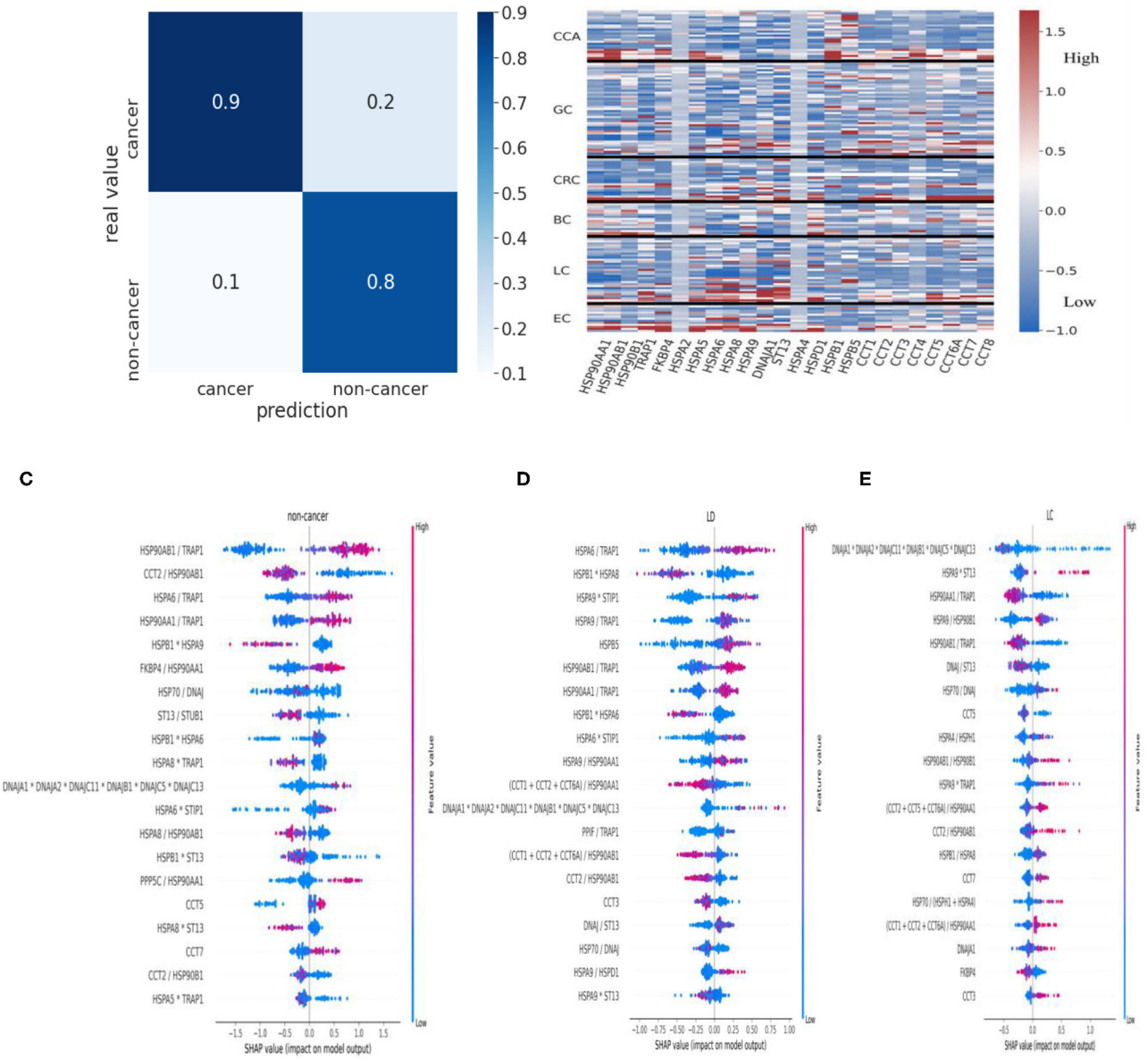

FIGURE 1 | HSPs in urine as biomarkers of cancer. **(A)** Confusion matrix for the cancer prediction model. **(B)** Heatmap of z-score normalized HSP expression levels in the urine of the patients with different types of cancer. Values were clipped to the 1st percentile of the *z*-scores and to the 97th percentile to minimize the effect of outliers. **(C)** HSPs and cochaperones in cancer and non-cancer patients. Negative values indicate a positive contribution of specific proteins to the probability that a patient has cancer. Positive SHAP values indicate that the corresponding values of the proteins are associated with lower chances of the patient having cancer. For simplicity, we presented HSPA2+HSPA6+HSPA8+HSPA12+HSPA5 as "HSP70" and DNAJA1+DNAJA2+DNAJC11+DNAJB1+DNAJC5+DNAJC13 as "DNAJ". **(D,E)** SHAP summary plots for the cancer prediction model. HSPs in urine were used to identify the critical proteins and the protein ratios in patients with benign lung disease (LD) such as PM and COPD **(D)** and LC patients **(E)**. HSPs, heat shock proteins; PM, pneumonia; COPD, chronic obstructive pulmonary disease; LC, lung cancer; SHAP, Shapely Addictive Explanations.

2-positive breast cancers (30). Therefore, the identification of type-specific HSP signatures in a heterogeneous group of tumors warrants further investigation.

It is also interesting to see the changes in HSPs between patients with benign lung disease and lung cancer patients (**Figures 1D,E**). Patients with lung disease have a higher level of cytoplasmic HSP90 homologs (HSP90AA1 and HSP90AB1) in relationship to mitochondrial HSP90 homolog (TRAP1), whereas patients with lung cancer have a higher level of TRAP1 to the level of cytoplasmic HSP90 (**Figures 1D,E**). Furthermore, the

level of HSP70 to its cochaperone DNAJ/HSP40 does not seem to change between benign lung disease and cancer in contrast with a higher level of ST13 to DNAJ associated with lung cancer (**Figures 1D,E**). During the HSP70 functional cycle, ST13, also known as Hsc70-interacting protein (Hip), preferentially binds to the ADP-bound state of HSP70–peptide complexes, slowing the release of ADP from HSP70-nucleotide binding domain, thus, promoting degradation of HSP70 clients (24, 31, 32). This may suggest that HSP70 is predominantly "freezed" in its high-affinity ADP state in lung cancer patients and that the role of Hip should be further investigated in the context of cancer. The levels of CCTs also seem to influence the shift from lung disease to lung cancer (**Figures 1D,E**; **Supplementary Figure 1**). This provides a good example of the specific HSPs that made a positive contribution to shifting a balance from the benign disease state to cancer. Further understanding of HSP changes between benign disease and cancer may potentially provide clues for the discoveries of novel HSP-based biomarkers and therapeutic targets.

In conclusion, coupling the machine learning approach and understanding of how HSPs operate, including their functional cycles as well as collaboration with and within networks, are certainly effective in identifying specific types of cancer, which may form the basis for future discoveries of novel HSP-based biomarkers of cancer.

CONCLUSION

Heat shock proteins are molecular chaperones that are aberrantly expressed in cancer patients and shown to be implicated in the various stages of cancer development. We hypothesized that HSPs in urine can be used to predict cancer. We show that HSPs can be used to identify cancer patients with nearly 90% precision based on HSP signatures in urine. We highlighted that understanding of HSP networks and how HSP operates in cells are crucial for the identification of HSP-based biomarkers of cancer. Further understanding of the HSP chaperone system may help in the development of effective type-specific biomarkers of cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZA collected the resources, contributed to the conceptualization, writing, review and editing of the manuscript, formal analysis and finance acquisition. DDN contributed to methodology, machine learning, review and editing of the manuscript. YM and AS provided administrational support. All the authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the RFBR, project number 20-315-90081.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.743476/full#supplementary-material

REFERENCES

1. Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford EA, et al. Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones.* (2009) 14:105–11. doi: 10.1007/s12192-008-0068-7
2. Albakova Z, Siam MKS, Sacitharan PK, Ziganshin RH, Ryazantsev DY, Sapozhnikov AM. Extracellular heat shock proteins and cancer: new perspectives. *Transl Oncol.* (2021) 14:100995. doi: 10.1016/j.tranon.2020.100995
3. Seiwert TY, Tretiakova M, Ma PC, Khaleque MA, Husain AN, Ladanyi A, et al. Heat shock protein (HSP) overexpression in lung cancer and potential as a therapeutic target. *Cancer Res.* (2005) 65:559.
4. Suzuki K, Ito Y, Wakai K, Kawado M, Hashimoto S, Seki N, et al. Serum heat shock protein 70 levels and lung cancer risk: a case-control study nested in a large cohort study. *Cancer Epidemiol Biomark Prev.* (2006) 15:1733. doi: 10.1158/1055-9965.EPI-06-0005
5. Gráf L, Barabás L, Madaras B, Garam N, Maláti É, Horváth L, et al. High serum Hsp70 level predicts poor survival in colorectal cancer: results obtained in an independent validation cohort. *Cancer Biomark.* (2018) 23:539–47. doi: 10.3233/CBM-181683
6. Gunaldi M, Afsar CU, Okuturlar Y, Gedikbasi A, Kocoglu H, Kural A, et al. Elevated serum levels of heat shock protein 70 are associated with breast cancer. *Tohoku J Exp Medi.* (2015) 236:97–102. doi: 10.1620/tjem.236.97
7. Dakappagari N, Neely L, Tangri S, Lundgren K, Hipolito L, Estrellado A, et al. An investigation into the potential use of serum Hsp70 as a novel tumour biomarker for Hsp90 inhibitors. *Biomarkers.* (2010) 15:31–8. doi: 10.3109/13547500903261347
8. Tas F, Bilgin E, Erturk K, Duranyildiz D. Clinical significance of circulating serum cellular heat shock protein 90 (HSP90) level in patients with cutaneous malignant melanoma. *Asian Pac J Cancer Prev.* (2017) 18:599–601. doi: 10.22034/APJCP.2017.18.3.599
9. Olejek A, Damasiewicz-Bodzek A, Bodzek P, Wielkoszyński T, Zamłyński J, Stołtny P, et al. Concentrations of antibodies against heat shock protein 27 in the sera of women with ovarian carcinoma. *Int J Gynecol Cancer.* (2009) 19:1516–20. doi: 10.1111/IGC.0b013e3181bf425b
10. Oka M, Sato S, Soda H, Fukuda M, Kawabata S, Nakatomi K, et al. Autoantibody to heat shock protein Hsp40 in sera of lung cancer patients. *Jpn J Cancer Res.* (2001) 92:316–20. doi: 10.1111/j.1349-7006.2001.tb01097.x
11. Bodzek P, Partyka R, Damasiewicz-Bodzek A. Antibodies against Hsp60 and Hsp65 in the sera of women with ovarian cancer. *J Ovarian Res.* (2014) 7:30. doi: 10.1186/1757-2215-7-30
12. Hamelin C, Cornut E, Poirier F, Pons S, Beaulieu C, Charrier P, et al. Identification and verification of heat shock protein 60 as a potential serum marker for colorectal cancer. *FEBS J.* (2011) 278:4845–59. doi: 10.1111/j.1742-4658.2011.08385.x

13. Campanella C, Rappa F, Scium,è C., Marino Gammazza A, Barone R, Bucchieri F, et al. Heat shock protein 60 levels in tissue and circulating exosomes in human large bowel cancer before and after ablative surgery. *Cancer.* (2015) 121:3230–9. doi: 10.1002/cncr.29499
14. Wyciszkiewicz A, Kalinowska-Łyszczarz A, Nowakowski B, Kazmierczak K, Osztynowicz K, Michalak S. Expression of small heat shock proteins in exosomes from patients with gynecologic cancers. *Sci Rep.* (2019) 9:9817. doi: 10.1038/s41598-019-46221-9
15. Chanteloup G, Cordonnier M, Isambert N, Bertaut A, Hervieu A, Hennequin A, et al. Monitoring HSP70 exosomes in cancer patients' follow up: a clinical prospective pilot study. *J Extracell Vesicles.* (2020) 9:1766192. doi: 10.1080/20013078.2020.1766192
16. Campanella C, Pace A, Caruso Bavisotto C, Marzullo P, Marino Gammazza A, Buscemi S, et al. Heat shock proteins in Alzheimer's disease: role and targeting. *Int J Mol Sci.* (2018) 19:2603. doi: 10.3390/ijms19092603
17. Wojsiat J, Prandelli C, Laskowska-Kaszub K, Martín-Requero A, Wojda U. Oxidative stress and aberrant cell cycle in Alzheimer's disease lymphocytes: diagnostic prospects. *J Alzheimer's Dis.* (2015) 46:329–50. doi: 10.3233/JAD-141977
18. Kilic A, Mandal K. Heat shock proteins: pathogenic role in atherosclerosis and potential therapeutic implications. *Autoimmune Dis.* (2012) 2012:502813. doi: 10.1155/2012/502813
19. Li Z, Song Y, Xing R, Yu H, Zhang Y, Li Z, et al. Heat shock protein 70 acts as a potential biomarker for early diagnosis of heart failure. *PLoS ONE.* (2013) 8:e67964. doi: 10.1371/journal.pone.0067964
20. Zhang C, Leng W, Sun C, Lu T, Chen Z, Men X, et al. Urine proteome profiling predicts lung cancer from control cases and other tumors. *EBioMedicine.* (2018) 30:120–8. doi: 10.1016/j.ebiom.2018.03.009
21. Dolgun A, Demirhan H. Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. *Commun Stat Simul Comput.* (2017) 46:5166–83. doi: 10.1080/03610918.2016.1146761
22. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds, et al. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology.* (2009) 23:255–64. doi: 10.1037/a0012850
23. Calderwood SK, Gong J. Heat shock proteins promote cancer: it's a protection racket. *Trends Biochem Sci.* (2016) 41:311–23. doi: 10.1016/j.tibs.2016.01.003
24. Albakova Z, Armeev GA, Kanevskiy LM, Kovalenko EI, Sapozhnikov AM. HSP70 multi-functionality in cancer. *Cells.* (2020) 9:587. doi: 10.3390/cells9030587
25. Murphy ME. The HSP70 family and cancer. *Carcinogenesis.* (2013) 34:1181–8. doi: 10.1093/carcin/bgt111
26. Albakova Z, Mangasarova Y, Sapozhnikov A. Heat shock proteins in lymphoma immunotherapy. *Front Immunol.* (2021) 12:660085. doi: 10.3389/fimmu.2021.660085
27. Kluger HM, Chelouche Lev D, Kluger Y, McCarthy MM, Kiriakova G, Camp RL, et al. Using a xenograft model of human breast cancer metastasis to find genes associated with clinically aggressive disease. *Cancer Res.* (2005) 65:5578. doi: 10.1158/0008-5472.CAN-05-0108
28. Balogi Z, Multhoff G, Jensen TK, Lloyd-Evans E, Yamashima T, Jäättelä M, et al. Hsp70 interactions with membrane lipids regulate cellular functions in health and disease. *Prog Lipid Res.* (2019) 74:18–30. doi: 10.1016/j.plipres.2019.01.004
29. Juhasz K, Lipp, A.-M., Nimmervoll B, Sonnleitner A, Hesse J, et al. The complex function of hsp70 in metastatic cancer. *Cancers.* (2013) 6:42–66. doi: 10.3390/cancers6010042
30. Klimczak M, Biecek P, Zylicz A, Zylicz M. Heat shock proteins create a signature to predict the clinical outcome in breast cancer. *Sci Rep.* (2019) 9:7507. doi: 10.1038/s41598-019-43556-1
31. Li Z, Hartl FU, Bracher A. Structure and function of Hip, an attenuator of the Hsp70 chaperone cycle. *Nat Struct Mol Biol.* (2013) 20:929–35. doi: 10.1038/nsmb.2608
32. Rousaki A, Miyata Y, Jinwal UK, Dickey CA, Gestwicki JE, Zuiderweg ER. Allosteric drugs: the interaction of antitumor compound MKT-077 with human Hsp70 chaperones. *J Mol Biol.* (2011) 411:614–32. doi: 10.1016/j.jmb.2011.06.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

SYSTEMATIC REVIEW
published: 02 November 2021
doi: 10.3389/frai.2021.736697

What Makes Artificial Intelligence Exceptional in Health Technology Assessment?

Jean-Christophe Bélisle-Pipon[1], Vincent Couture[2], Marie-Christine Roy[3], Isabelle Ganache[4], Mireille Goetghebeur[4] and I. Glenn Cohen[5]*

[1]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada, [2]Faculty of Nursing, Laval University, Quebec, QC, Canada, [3]École de Santé Publique, Université de Montréal, Québec, QC, Canada, [4]Institut National D'Excellence en Santé et en Services Sociaux (INESSS), Montréal, Québec, QC, Canada, [5]Harvard Law School, Cambridge, MA, United States

The application of artificial intelligence (AI) may revolutionize the healthcare system, leading to enhance efficiency by automatizing routine tasks and decreasing health-related costs, broadening access to healthcare delivery, targeting more precisely patient needs, and assisting clinicians in their decision-making. For these benefits to materialize, governments and health authorities must regulate AI, and conduct appropriate health technology assessment (HTA). Many authors have highlighted that AI health technologies (AIHT) challenge traditional evaluation and regulatory processes. To inform and support HTA organizations and regulators in adapting their processes to AIHTs, we conducted a systematic review of the literature on the challenges posed by AIHTs in HTA and health regulation. Our research question was: What makes artificial intelligence exceptional in HTA? The current body of literature appears to portray AIHTs as being exceptional to HTA. This exceptionalism is expressed along 5 dimensions: 1) AIHT's distinctive features; 2) their systemic impacts on health care and the health sector; 3) the increased expectations towards AI in health; 4) the new ethical, social and legal challenges that arise from deploying AI in the health sector; and 5) the new evaluative constraints that AI poses to HTA. Thus, AIHTs are perceived as exceptional because of their technological characteristics *and* potential impacts on society at large. As AI implementation by governments and health organizations carries risks of generating new, and amplifying existing, challenges, there are strong arguments for taking into consideration the exceptional aspects of AIHTs, especially as their impacts on the healthcare system will be far greater than that of drugs and medical devices. As AIHTs begin to be increasingly introduced into the health care sector, there is a window of opportunity for HTA agencies and scholars to consider AIHTs' exceptionalism and to work towards only deploying clinically, economically, socially acceptable AIHTs in the health care system.

Keywords: artificial intelligence, exceptionalism, ethical, social and legal implications, health technology assessment, health regulation

OPEN ACCESS

Edited by:
Pietro Lio, University of Cambridge, United Kingdom

Reviewed by:
Ibrahem Kandel, Universidade NOVA de Lisboa, Portugal
Rubul Kumar Bania, North-Eastern Hill Uiveristy, India

****Correspondence:***
Jean-Christophe Bélisle-Pipon jean-christophe_belisle-pipon@sfu.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *05 July 2021*
Accepted: *23 September 2021*
Published: *02 November 2021*

Citation:
Bélisle-Pipon J-C, Couture V, Roy M-C, Ganache I, Goetghebeur M and Cohen IG (2021) What Makes Artificial Intelligence Exceptional in Health Technology Assessment? Front. Artif. Intell. 4:736697. doi: 10.3389/frai.2021.736697

INTRODUCTION

Health technology assessment (HTA) is key to the introduction of artificial intelligence (AI) applications in health. HTA generally requires a systematic examination of health technologies' features, effects, and/or impacts allows for the appraisal of clinical, economic, social, organizational and ethical implications (Banta and Jonsson 2009; Kristensen et al., 2017; O'Rourke et al., 2020). While regulatory assessment often is conducted by supranational (e.g., European Medicines Agency, EMA) and national (US FDA, Health Canada) regulators, HTA is mostly conducted at regional, provincial or state-based level and represents the main gateway for a health technology (e.g., drugs, vaccines, medical devices) to be widely administered to patients (Vreman et al., 2020; Wang et al., 2018). A health technology that is positively evaluated by a health regulator or an HTA agency signals significant support for its use, causing clinicians, patients, hospital administrators and third-party payers (such as public or private health insurers) to consider deploying and reimbursing this technology in their health care system or setting (Allen et al., 2017; Wild, Stricka, and Patera 2017). However, AI is not just another health technology, and many commentators view its assessment as complex and particularly challenging (Harwich and Laycock 2018; Mason et al., 2018; Shaw et al., 2019). For instance, AI health technologies (AIHT) implementation within the healthcare system is often done in a fairly short timeframe after their development (months rather than years as for drugs and vaccines), with the result that there is not yet as much evidence of their effectiveness and impacts as would be required by traditional HTA for many other health technologies (Babic et al., 2019). Moreover, AI systems deployed within the healthcare system continue to learn and evolve over time based on the data they process (Reddy 2018); this is in contrast, for example, with drugs whose formulation, dosage and routes of administration are regulated, and to be modified for use in clinical context and service delivery, often require new approval by HTA. In addition, AI systems require to be trained on and use vast amounts of (potentially sensitive) data (about patients, research participants, clinicians, managers, health care systems, etc.) that raise issues of privacy, (cyber)security, informed consent, data stewardship and control over data usages (Wang and Preininger 2019; Dash et al., 2019; Sun and Medaglia 2019; Bartoletti 2019).

The application of AI in health is expected to transform the way we diagnose, prevent and treat as well as the way we interact with technologies (Patel et al., 2009; Hamet and Tremblay 2017; The Lancet 2017). This may advance healthcare by enhancing efficiency by automatizing routine tasks and decreasing health-related costs (Shafqat et al., 2020), broadening access to healthcare delivery (Harwich and Laycock 2018), targeting more precisely patient needs (Jameson and Longo 2015), and assisting clinicians in their decision-making (Lysaght et al., 2019; Smith 2020). For these benefits to materialize, governments and health authorities must efficiently regulate AI, and conduct appropriate health technology assessment (HTA). However, the very definition of AI in health is still the subject of discussion, debate and negotiation among both researchers and government authorities. AI in the health sector can be broadly defined as a field concerned with the development of algorithms and systems seeking to reproduce human cognitive functions, such as learning and problem-solving (Tang et al., 2018) with (current and anticipated) uses that include (without being limited to) supporting medical decision-making (Ahmed et al., 2020), pharmacovigilance (Leyens et al., 2017), and prediction and diagnosis (Noorbakhsh-Sabet et al., 2019). In fact, some AIHTs have already been approved by the FDA, such as AI-powered devices to diagnose eye diseases (Samuel and Gemma Derrick 2020). Risks and harms of AI in healthcare are described at all levels, from the clinical encounter (e.g., adverse effects of an AIHT that can spread to entire patient populations, inexplicability of an AI-based medical decision, issues with assigning responsibility for adverse events, and patients' loss of trust in their provider) to society as a whole (e.g., furthering inequalities due to algorithm training on biased data) (Sparrow and Hatherley 2019). Interestingly, one indication that current HTA processes are not yet well adapted is the fact that a significant number of AIHTs are benefiting from regulatory fast-track and do not undergo HTA review, a situation that is particularly noticeable in the United States (Benjamens et al., 2020; Gerke et al., 2020; Tadavarthi et al., 2020).

Even though AI solutions offer great potential for improving efficiency, health organizations are confronted with a vast array of AI solutions that have not yet been subject to extensive HTA (Love-Koh et al., 2018). Moreover, many authors have highlighted that these new technologies challenge traditional evaluation processes as well as the assessment of the ethical, legal and social implications (ELSI) that AIHTs may entail (He et al., 2019; Racine et al., 2019; Shaw et al., 2019; Ahmad et al., 2020; Benjamens et al., 2020), thus further impeding the already insufficient evaluative processes of AI health technologies (AIHTs). To inform and support HTA organizations in adapting their evaluation processes to AIHTs, we conducted a systematic review of the literature on the ethical, legal and social challenges posed by AIHTs in HTA. The present article was guided by this question: what makes artificial intelligence exceptional in health technology assessment? To our knowledge, this is the first review on this topic. After describing the methodology of the review, we will provide a comprehensive overview of AI-specific challenges that need to be considered to properly address AIHTs' intrinsic and contextual peculiarities in the context of HTA. This will lead to point possible explanations of this exceptionalism and solutions for HTA. Overall, this review is intended to build insights and awareness and allow to inform HTA practices.

METHODOLOGY

To map the exceptional challenges posed by AIHTs in HTA, we conducted a literature search for articles indexed in PubMed, Embase, Journals@Ovid, Web of Science and the International HTA database. Our review is part of a larger literature review addressing the full range of ethical, legal, social and policy implications that impact HTA processes for AIHTs. Therefore, the search strategy focused on three concepts: AI, HTA and ELSI. **Table 1** presents the search equations by theme for each reviewed database. In terms of definition of AI, we sought to remain agnostic and did not use specific definitions of AI. Instead, we used an inductive approach using a series of keywords (see

TABLE 1 | Search strategy.

Concepts	Terms
AI	**PB** = [(Artificial Intelligence) OR (Machine Learning) OR (Deep Learning) OR (Natural Language Processing) OR (Chatbot*) OR (Carebot*) OR (Big Data)] OR [Artificial intelligence OR Big Data (MeSH Terms)] **EM; OJ; WoS** = (Artificial Intelligence) OR (Machine Learning) OR (Deep Learning) OR (Natural Language Processing) OR (Chatbot*) OR (Carebot*) OR (Big Data) **iHTAd** = (Artificial Intelligence) AND
HTA	**PB** = (Health Technology Assessment) OR (HTA) OR (Technology Assessment) OR [Technology Assessment, Biomedical (MeSH Terms)] **EM; OJ; WoS** = (Health Technology Assessment) OR (HTA) OR (Technology Assessment) **iHTAd** = [*Empty*] AND
ELSI	**PB** = (ESLI) OR (Ethic*) OR (Bioethic*) OR (Moral*) OR (Legal*) OR (Law) OR (Societ*) OR (Polic*) OR (Governance) OR (Trust) OR (Mistrust) OR (Jurisprudence) OR (Public Policy) OR (Bioethics OR Ethics OR Jurisprudence OR "Public Policy" [MeSH Terms]) **EM; OJ; WoS** = (ESLI) OR (Ethic*) OR (Bioethic*) OR (Moral*) OR (Legal*) OR (Law) OR (Societ*) OR (Polic*) OR (Governance) OR (Trust) OR (Mistrust) OR (Jurisprudence) OR (Public Policy) **iHTAd** = [*Empty*]

Legend. *PB = PubMed; EM = Embase; OJ = Journals@Ovid Full Text; Databasel; WoS = Web of Science; iHTAd = International HTA.*

TABLE 2 | Selection criteria.

	Specifics
Date	2016–2020 (5 years)
Language	English; French
Study design	Descriptive; Experimental; Opinion/Perspective; Empirical Research; Literature Review
Type of publication	Original research; Commentary; Editorial

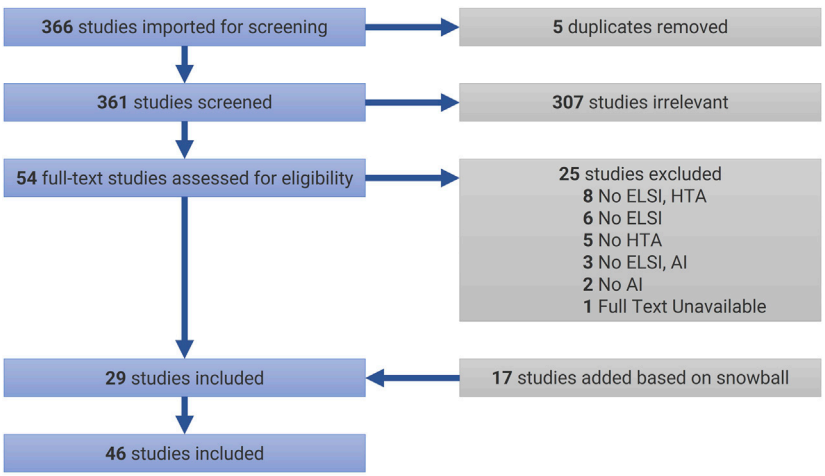

FIGURE 1 | PRISMA Flowchart. AI = artificial intelligence; ELSI = ethical, legal, and social implications; HTA = health technology assessment.

Table 1) to identify and collect articles that mention using or discussing AIHTs. The construction of the research strategy and the choice of equations was supported by librarians.

The initial search (as of December 27, 2020) returned a total of 366 articles, which were uploaded in Covidence. JCBP and VC conducted a careful analysis of the titles and abstracts that lead to

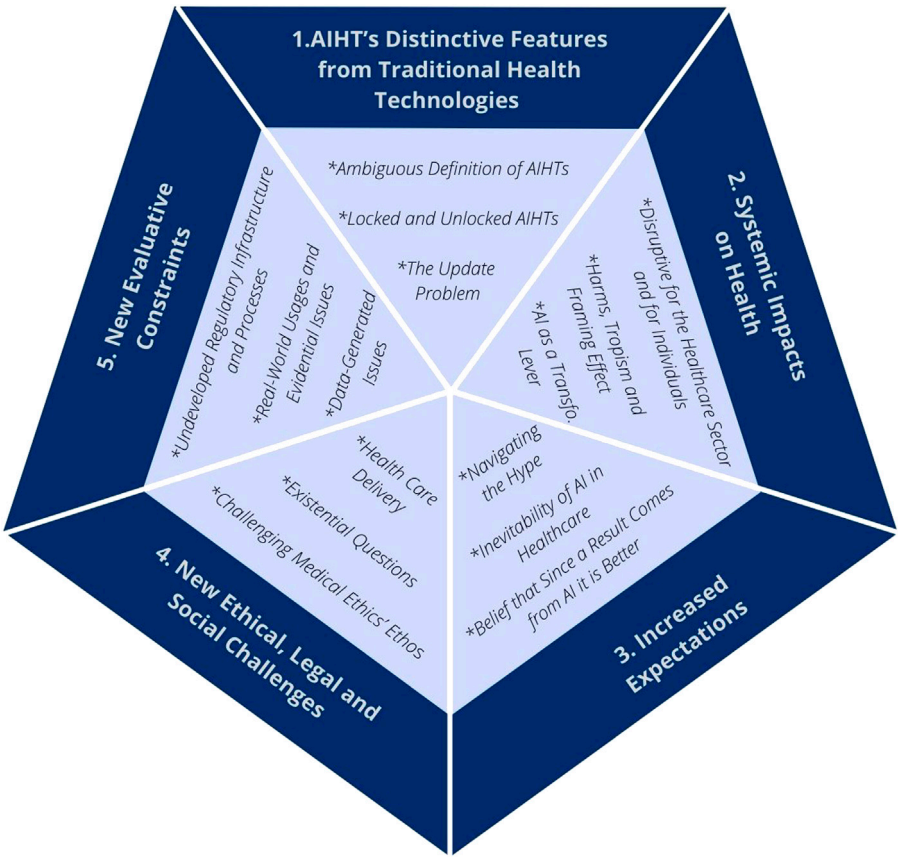

FIGURE 2 | The five main aspects of artificial intelligence health technologies' exceptionalism.

excluding 307 articles, and JCBP conducted the subsequent analysis of main texts allowed to select 29 articles for review (see **Table 2** for selection criteria). In case of doubt or ambiguity, articles were discussed with MCR to decide on inclusion or exclusion. In addition to this sample, in January 2021, a snowball process helped identify 17 additional papers that fitted the selection criteria. **Figure 1** presents our review flowchart following PRISMA's guidelines (Moher et al., 2009). Documents were thematically analyzed (Braun and Clarke 2006) with the help of NVivo 12. In the present article, we focus on the theme "exceptionalism of AI in HTA". Additional themes will be published in subsequent papers.

RESULTS

What follows is a presentation of the key considerations that have been raised in the reviewed literature regarding AI's peculiarities, and the challenges they raise, in the context of HTA. Twenty eight articles from the total sample discussed these peculiarities and challenges, which are presented as exceptional features of AI by authors. The "exceptionalism" of AIHTs can be broken down into five main aspects (see **Figure 2**): 1) AIHT's distinctive features; 2) their systemic impacts on health care and the health sector; 3) the increased expectations towards AI in health; 4) the new ethical, social and legal challenges that arise from deploying AI in the health sector; and 5) the new evaluative constraints that AI poses to HTA. **Table 3** presents a summary of the key considerations for each aspect.

Artificial Intelligence Health Technologies' Distinctive Features From Traditional Health Technologies

AIHT's exceptionalism is associated with the technology's definitional and foundational nature. Distinctive features include AIHTs' ambiguous definition; the fact that AIHTs may or may not continue to evolve; and the need to keep AIHTs up to date to reap the benefits and avoid the risks of harms.

Ambiguous Definition of Artificial Intelligence Health Technologies

According to Gerke et al. (2020), AIHTs are different from traditional health technologies for three reasons: their capacity to continuously learn, their potential for ubiquity throughout the health care system, and the opaqueness of their recommendations. However, AIHTs suffer from ambiguities

TABLE 3 | A summary of the five main aspects of AIHT's exceptionalism.

	Key considerations (In italic key sub-considerations)	Examples from the reviewed sample
1. AIHT's Distinctive Features from Traditional Health Technologies	AIHTs are different from traditional health technologies because of their capacity to continuously learn, their potential for ubiquity throughout the health care system, the opaqueness of their recommendations and the ambiguity of their definition (*Ambiguous Definition of AIHTs*) Locked algorithms will always yield the same result when it is fed by the same data. They are not *per se* safer and may require new regulatory approvals, though they are easier to assess than unlocked algorithms. Unlocked or adaptive algorithms improve over time, which demands that their safety and security must be continually re-evaluated. 'Lifecycle' regulation seems to be key in addressing these concerns, but for the most part burden lies on the regulators to adjusted their assessment of an AIHT in light of the evolving evidence, which is very resource intensive and for which HTA agencies are not yet equipped to conduct. (*Locked and Unlocked AIHTs*) Algorithms will need to be regularly updated (at high or even prohibitive prices) due to advances in medical knowledge and access to new datasets or at the risk of their usage becoming malpractice. Updating or replacing an AIHT will involve additional post-acquisition costs to the clinics and hospitals that purchased them. The difficulty of managing the consequences of an outdated algorithm outweighs those of a drug or other health product that must be withdrawn from the market (*The Update Problem*)	Locked AIHTs could become outdated potentially from the moment they are prevented from evolving. Thus, locking AIHT may cause it to become outdated and increase chance of contextual bias in real-life contexts
2. Systemic Impacts on Health	AI may have systemic effects that can be felt across an entire health care system, or across health care systems in several jurisdictions, initiating extensive and lasting transformations that are likely to affect all actors working in, using or financing the health system. In addition, AIHTs can have systemic real-world consequences for patients and non-ill or non-frequent users of the health care system. However, AI will not address everything that has to do with the overall well-being of people (*Disruptive for Both the Healthcare Sector and for Individuals*) Mistakes due to AIHTs used in clinical care and within the health care system have the potential to widely affect the patient population, suggesting that it is all the more necessary that all algorithms should submitted to extensive scrutiny. In addition, "tropic effects" (i.e., code embedded propensity towards certain behaviors or effects) may increase the risk of inappropriate treatment and care, and may result in importing AIHT-fueled standards and practices that are exogenous and non-idiosyncratic to local organizations. Furthermore, the large-scale systematization of certain behaviors may end up resulting in significant costs and harms (*Harms, Tropism and Framing Effect*) Some authors suggest AIHTs should be regarded as a "health system transformation lever" for improving health care and a key enabler of learning healthcare systems (LHS) (*AI as a Transformation Lever for the Health Sector*)	AI's role in health surveillance, care optimization, prevention, public health, and telemedicine will cause AIHTs to affect non-ill or non-frequent users of the health care system An AIHT trained on medico-administrative data in a context where physicians have often modified their billing to enter the highest paying codes for clinical procedures would cause the algorithm to infer that these codes represent the usual, standard, or common practice to be recommended, thus introducing a bias in the algorithm and leading to a cascade of non-cost effective recommendations
3. Increased Expectations	The "automation bias" describes the belief that an AI-generated outcome is inherently better than a human one. This is reinforced by the technological imperative, i.e., the pressure to use a new technology just because it exists (*Belief that Since a Result Comes from AI it is Better*) These high expectations toward AIHTs form the basis of the inevitability of AI in health. However, the concept of *AI chasm* refers to the phenomenon that while AIHTs are very promising,	The adoption and impact of AIHTs are unlikely to be uniform or to improve performance in all health care contexts because of the technology's distinctive features, its systemic effects on health care organizations and the human biases associated with the use of these technologies. AIHTs can significantly affect and highlight particularities of workflow and design of individual hospital systems, causing them not to respond in an intended way. Therefore, AIHTs represent great challenges for deciding whether marketing authorization is justified

(Continued on following page)

TABLE 3 | (*Continued*) A summary of the five main aspects of AIHT's exceptionalism.

	Key considerations (In italic key sub-considerations)	Examples from the reviewed sample
	very few will actually be successful once implemented in clinical settings and can help rebalance the expectations. HTA agencies have an important role to play here to contain this phenomenon (*Inevitability of AI in Healthcare*)	
	AI is currently in an era of promises rather than of fulfillment of what is expected from it. Possible consequences of this hype can be very significant but HTA agencies and regulators have an important role to play (*Navigating the Hype*)	
4. New Ethical, Legal and Social Challenges	AIHTs present new ethical, legal and social challenges in the context of health care delivery; by calling into question the roles of patients, HCPs and decision-makers; and by conflicting with medicine's ethos of transparency	Patients who compare very well with historic patient data will be the ones benefiting the most from AIHTs, calling for caution with regards to patient and disease heterogeneity
	Key AIHT-stemmed ethical challenges in care delivery are: AI-fostered potential bias; patient privacy protection; trust of clinicians and the general public towards machine-led medicine; new health inequalities (*Health Care Delivery*)	Practical and procedural ethical guidance for supporting HTA for AIHTs has not yet been thoroughly defined. For instance, distributive justice role in HTA for AIHT is not well specified
	AI being unlike most other health technologies, it forces the questioning of the very essence of humans. It also raises new existential questions regarding the role of regulators and public decision-makers AIHTs unparalleled autonomy intensifies ethical and regulatory challenges (*Existential Questions*)	AI-stemmed existential questionning includes the reflection that more and more clinicians are having about the proper role of healthcare professionals and what it means to be a doctor, a nurse, etc. And from the patients' perspective, what it means to be cared for by machines and to feel more and more like a number in a vast system run by algorithms
	AIHTs are often opaque, which poses serious problems for their acceptance, regulation and implementation in the health care system. AI's benefits for health care will come at the price of raising ethical issues specific to the technology (*Challenging Medical Ethics' Ethos*)	
5. New Evaluative Constraints	AIHTs raise new evaluative constrains at the technological level due to the data and infrastructure required (*Data-Generated Issues*)	The adoption and impact of AIHTs are unlikely to be uniform or to improve performance in all health care contexts because of the technology's distinctive features, its systemic effects on health care organizations and the human biases associated with the use of these technologies. Therefore, AIHTs represent great challenges for deciding whether marketing authorization is justified, and it forces to question whether marketing authorization at the 10,000 foot level for the product is appropriate and efficient as opposed to for more specific uses closer to the impacted communities and the point of delivery
	New constraints also appear at the clinical level because of the greater variation in AIHTs performance between the test environment and the real-word context than those of drugs and medical devices (*Real-World Usages and Evidential Issues*)	
	This high level of complexity requires a special regulation of AIHT, specifically adapted to its complexity (*Undeveloped Regulatory Infrastructure and Processes*)	

with respect to their definition and purpose as there is no agreed-upon definition that may help build an adapted and efficient policy and regulatory infrastructure (Pesapane et al., 2018). The drawback of AI exceptional features and of the high variability that exists among AI systems is that it poses definitional problems that affect AIHTs' regulation and slow down their deployment in the healthcare sector (Love-Koh et al., 2018; Haverinen et al., 2019). Compared with traditional health technologies (such as drugs, vaccines and medical devices), AIHTs are not static products and have the capability to learn and improve over time (Parikh, Obermeyer, and Navathe 2019; Dzobo et al., 2020). AIHTs are therefore in stark contrast with most technologies in medicine, which are fairly well defined and usually implemented when they are fairly well understood.

Locked and Unlocked Artificial Intelligence Health Technologies

Contributing to the distinctiveness of AIHTs in the health sector, self-learning and self-adaptation propensities clash with current regulatory frameworks and clinical practices (Alami et al., 2020; Fenech et al., 2020). It is easier to evaluate "locked" AIHTs, which are much more comparable to current health technologies (which cannot by themselves evolve). Currently, the majority of FDA-approved AIHTs have their capability to evolve locked (Dzobo

et al., 2020; Miller, 2020). A locked algorithm will always yield the same result when it is fed by the same data, therefore it does not change overtime with uses. Locked algorithms are not *per se* safer. They could be more harmful than "unlocked" or "adaptive" algorithms if they end up yielding erroneous results (based on legacy training data that are outdated), misleading patient care or systematizing biases (Prabhu 2019). Thus, a locked AIHT may require new regulatory approvals if during real-world usage significant and unexpected patterns of results are observed (i.e., stable and expected process produces outcomes unexpected because of incorrect priors about the data fed into an AIHT) or if it is deemed necessary to update the algorithms to match advances in medical knowledge (Gerke et al., 2020; Miller, 2020). There also is the issue of when AIHT is used (or not used) on new populations that differ from the training data, which raises questions about how the training data upon which an AIHT was developed and whether certain populations may be unduly excluded from benefiting from its development and implementation. Therefore the concept of locked may be misleading and should not be conveyed as safer (Babic et al., 2019). Unlocked or adaptive algorithms that improve over time is the future according to some (Prabhu 2019) as they will outperform humans (Dzobo et al., 2020). But some issues are to be expected. Unlocked AIHT may change as they process new data and yield new outcomes without the knowledge or oversight of its users, which demands that their safety and security must be continually re-evaluated (Abràmoff et al., 2020). Also, unlike traditional healthcare technologies and locked algorithms, unlocked AIHTs are more vulnerable to cyber-attacks and misuse that can cause the algorithm to generate problematic and highly damaging outputs (Babic et al., 2019; Miller, 2020).

The Update Problem

Another consideration that helps AIHT qualify for being an anomalous technology in the health sector is that the algorithms will need to be regularly updated (at high or even prohibitive prices) due to advances in medical knowledge and access to new datasets or at the risk of their usage becoming malpractice. To allow a rigorous analysis of the safety, efficiency, and equity of a given AIHT, it is necessary that the locked or unlocked state of the algorithm is always known to regulators and end-users (Char et al., 2020). Such transparency is necessary since due to the very distinct ethical and clinical implications that locked or unlocked AIHTs may generate. The update problem implies that a locked AIHT could quickly become outdated—potentially from the moment it is prevented from evolving (with or without supervision)—and that this could generate important risks as a result of the deployment and use of AIHTs in real-life contexts (Abràmoff et al., 2020). Although not all algorithms may need to evolve or be updated in the short term, at some point in time, updating or replacing an AIHT will involve additional post-acquisition costs. Post-acquisition updates and costs may seem counter-intuitive considering the distinctive characteristics attributed to AI, such as self-learning and continuous improvement. This may lead for certain organizations (in particular, in less affluent contexts or in periods of economic turmoil) not to deploy updates which will result in the uses of outdated algorithms and therefore sub-optimal benefits (if not harms) for some patients or services (Prabhu 2019). Since AIHT are considered as being more pervasive than physical technologies (such as drugs and other health products), some are arguying that managing the consequences of an outdated algorithm outweigh those of traditional health technologies (Babic et al., 2019; Prabhu 2019); even if it is very difficult to withdraw effectively a drug from the market, it is still possible to do so, while it may be much more challenging for AIHTs that are less visible, interpretable and tangible and more likely to be embedded in a hospital's or health system's IT systems.

Systemic Impacts on Health

Characteristic of disruptive technologies, AIHTs are said to have significant and systemic impacts on the healthcare sector. From the outset, what emerges is that AI has a capacity for information analysis that surpasses what is currently available from health professionals, healthcare managers or even from learning health systems (LHS) (Cowie 2017; Pesapane et al., 2018; Char and Burgart 2020). AI is geared towards changing healthcare practices by facilitating a better integration of innovations and of best practices that will yield optimal care delivery (Grant et al., 2020). These systemwide impacts may lead to both risks of harms and opportunities to optimize the health care system that must be taken into consideration in HTA.

Disruptive for Both the Healthcare Sector and for Individuals

Contrary to many health technologies, AI may have systemic effects that can be felt across an entire health care system, or even more so across health care systems in several jurisdictions (Dzobo et al., 2020). Gerhards et al. (2020) go as far as stating that AIHTs (especially those using machine learning) can yield significant changes to an entire healthcare system. These changes might not necessarily come from expected technological disruptions, but might come from the adaptation of the healthcare setting to certain methods and processes relying on AIHTs. This adaptation may initiate extensive and lasting transformations that are likely to affect all actors working in, using or financing the health system (Gerhards et al., 2020). The clinical use of some AIHTs may have the effect of transforming local health care administration practices by incorporating exogenous priors embedded within the technology. For instance, if a payer (public or private insurer) decides that a given AIHT recommendation become a precondition for reimbursement (i.e., making other care no longer reimbursable), this may have significant impacts on the way care is delivered, and will reduce patients', clinicians' and administrators' autonomy in making shared and appropriate decisions when the human-recommended care is different than a new gold standard based on AI on data and priors (Vayena, Blasimme, and Cohen 2018). There is therefore a process of importing practices, potentially very different, which can strongly contrast with local habits and

norms, requiring both adaptation and an impact assessment of these exogenous practices on the host environment.

AIHTs can have systemic real-world life-and-death consequences for patients (Miller, 2020), especially as AI will span across the life continuum from birth to death (Dzobo et al., 2020). AIHTs, unlike most drugs or medical devices, will also affect non-ill or non-frequent users of the health care system, be they due to AI increasing role in health surveillance, care optimization, prevention and public health, telemedicine (Love-Koh et al., 2018; Pesapane et al., 2018; Char and Burgart 2020). AI can help "democratizing health care" (i.e., in the sense of facilitating access) by extending care into patients' homes (Reddy et al., 2020), places where more individualized and personalized care can be facilitated. While being increasingly present in patient care, AI will not address everything that has to do with the overall well-being of people. Some aspects less related to illness, such as spirituality and sociality, will most likely not be resolved and supported by AI systems (Dzobo et al., 2020). Therefore, a systematic response to using AI in health care may systemically neglect important aspects of care.

Harms and Tropism

Mistakes due to AIHTs used in clinical care and within the health care system have the potential to widely harm the patient population. Some AI systems, especially in primary care settings, can have impacts on the entire population of a hospital or clinic (such as an AI-powered patient triage). This makes some people say that it is all the more necessary that all algorithms should be submitted to extensive scrutiny, with an increased attention on validation in clinical settings before they can be deployed in medical practice (Dzobo et al., 2020). In addition, a key challenge in implementing AI is that, without a comprehensive understanding of health needs (especially those not covered by AI), there is a risk of fragmenting healthcare delivery by silo use of AI systems. Such silo use may lead to weakening health systems capacity and efficiency in addressing patients needs.

AIHTs can have tropism effects on the healthcare system that may shape and normalize certain practices and expectations that are not necessarily accepted, widespread, cost-effective or standard in new contexts. An example of this would be an AIHT trained on medico-administrative data in a context where physicians have often modified their billing to enter the highest paying codes for clinical procedures, causing the algorithm to infer that these codes represent the usual, standard, or common practice to be recommended (Alami et al., 2020). Thus, the algorithmic inference would be biased because the procedure billed maximizes the clinician's remuneration, but potentially was not the one performed; this can lead to a cascade of non-cost effective recommendations. Such tropism effects may increase the risk of inappropriate treatment and care, and may result in importing AIHT-fueled standards and practices that are exogenous and non-idiosyncratic to local organizations and that may perpetuate latent biases in training data that are not present in certain health systems or contexts of care (Abràmoff et al., 2020; Alami et al., 2020; Miller, 2020). Therefore, the large-scale systematization of certain behaviors or inclinations may end up resulting in significant costs and harms for organizations and health systems as well for patients and HCPs (Alami et al., 2020; Hu et al., 2019). For example, higher sensitivities to clinical thresholds could lead to overdiagnosis or overprescription, while lower sensitivities could result in undiagnosed and untreated segments of the population; it is in the potential scope of the impacts that exceptionalism lies and must be carefully assessed in HTA (Alami et al., 2020; Topol 2020).

AI as a Transformation Lever for the Health Sector

According to Alami et al. (2020), instead of seeing AIHTs as a collection of distinct technologies, they should be regarded as a "health system transformation lever." AI can serve as a strategic lever for improving health care and services access, quality and efficiency. Used in such way, AI could have significant society-wide impacts, including technological, clinical, organizational, professional, economic, legal, and ethical.

AI can become a key enabler of learning healthcare systems (LHS) to achieve their full potential (Babic et al., 2019), especially since AIHTs are themselves learning systems (Ho, 2020). AIHTs and LHS can complement each other as both strive when there are porous boundaries between research and development and with clinical and organizational practices. Using data from the health care system, AI can learn and recalibrate both its performance and behaviors, and over time inform and refine the practices of the health care system (Babic et al., 2019). AI can allow for ongoing assessment of accuracy and usage and continuous risk monitoring (Ho, 2020).

AI, as a lever, can also have a systemic impact of putting forward the response to needs for which there are ready-to-use technologies, causing to pay little attention to serious unmet needs (Alami et al., 2020). According to Grant et al. (2020), AI may represent the "next major technologic breakthrough" in health care delivery, offering endless possibilities for improving both patient care and yielding health care system-wide optimizations. However, this blurring of boundaries poses significant problems for adequate regulatory design and should not be taken lightly (Babic et al., 2019).

Increased Expectations Towards Artificial Intelligence

Another key feature of AI exceptionalism is the increased expectations placed on AIHTs compared to other health technologies. According to Vollmer et al. (2020), AI systems often bear a misleading aura of obvious cutting-edge technology, which falsely limits the perceived need for careful validation and verification of their performance, clinical use, and general use once they begin to be used in routine clinical practice. The implications for HTA are three-pronged.

Belief that Since a Result Comes From an Artificial Intelligence It Is Better

A large part of the explanation for AI exceptionalism comes in particular from the belief that an AI-generated outcome is

inherently better than a human one (Char and Burgart 2020). This phenomenon, known as the automation bias, describes the fact that slowly but surely, AI is establishing itself as an authority over current practices and error-prone healthcare professionals. Part of this is reflected in the fact that it is now recognized as a problem that a person who disagrees with a result or recommendation generated by an AI must justify their opposition with much more data than those used by the AI to achieve that result (Char and Burgart 2020). The technological imperative—i.e., the mere fact that a sophisticated technological intervention exists creates pressure to use it because it is perceived to be superior to conventional practices, despite the risks—reinforces this belief and AI in medicine is currently having its technological imperative moment (Carter et al., 2020).

Inevitability of Artificial Intelligence in Healthcare

These high expectations toward AIHTs form the basis of the inevitability of AI in health. To the point where AI is seen as inevitably the future of medicine (Dzobo et al., 2020). Self-learning and the ability to perform arduous and repetitive tasks explains the growing interest for a greater place of AI in standard medical care. There are high hopes and, according to many commentators, good reasons to think that in a near future, virtually all physicians will be assisted by AI applications to expedite certain tasks and, in the median term, due to continuous learning, AIHTs might outperform humans in a wide range of areas (Dzobo et al., 2020; Abràmoff et al.2020; Gerke et al., 2020). It is not only for clinical or therapeutic reasons that AI seems to be inexorable; there is also competition within the AI ecosystem. The growing importance of AI and its inevitability also stems from the competition between political decision-makers from different jurisdictions to widely deploy AI in order not to lag behind others (Gerhards et al., 2020). Considering all the interests at stake, the massive investments and accelerated development of AI, the question is no longer whether AI will be part of routine clinical care, but when (Reddy et al., 2020).

Although the technological imperative is strong and that AI in health is very attractive and seems inevitable, caution is called for. In this regard, AI chasm is a powerful concept to rebalance and help manage expectations of overly rapid deployment and ubiquity of AI in health care (McCradden et al., 2020). AI chasm refers to the phenomenon that while AIHTs are very promising, very few will actually be successful once implemented in clinical settings. HTA agencies have an important role to play here to contain this phenomenon and reduce its frequency and spread (McCradden et al., 2020). One of the roles of evaluation and regulation is precisely to finely consider the implications of these technologies to overcome this phenomenon. This requires not only an analysis of technical efficiency and performance, but also an oversight of empirical and ethical validation to ensure the rights and interests of patients. This requires the development of regulatory tools that are well adapted to AIHTs so that there are clear procedures and processes to properly evaluate and screen AIHTs (Alami, Lehoux, Auclair, Guise, et al., 2020; Abràmoff et al., 2020). This is necessary to avoid ethical drifts and unacceptable (economic, health, social) costs that may be caused by technologies that are not adapted to the needs and specificities of a clinical or organizational context, or by milieus feeling pressure to deploy a technology and adapt its practices in a way that ultimately does not benefit patient care or sound health care management (Michie et al., 2017; Abràmoff et al.2020).

Navigating the Hype

AI is currently in an era of promises rather than of fulfillment of what is expected from it (Michie et al., 2017). This new science has yet to move beyond average outcomes on individuals to actual personalized benefits based on their situation, characteristics, and desired outcomes. It is important to remain critical and vigilant with respect to the rush to adopt these new technologies, possibly more so than politicians are at present (Cowie 2017). Especially when thought leaders' perspectives echo public wonderment and aspirations that AI transforms human life (Miller, 2020). With their development and implementation being largely driven by a highly speculative market and by proprietary interests, AIHTs are largely embedded into biocapital (Carter et al., 2020). That is to say, a vision of medical innovation that is based not on the actual creation of value, but on selling a certain vision of the future. It is through the sale of imaginary evoking unparalleled performance and disproportionate benefits to encourage all AI players to engage in the massive implementation of AI despite its uncertainties and shortcomings (Carter et al., 2020). That being said, in a study reported by Vayena et al. (2018), half of the surveyed American healthcare decision-makers expect that AIHTs will both improve medicine and fail meeting hyped expectations. Miller (2020) sums up the present phenomenon as follows: "No matter how sanguine the gurus touting game-changing AI technologies are, and no matter how much caregivers and patients hope that their benefits to medical practice and outcomes are not hype, all parties must remain vigilant." AIHTs are in their phase of promises and hype, which is creating inconsequent expectations (Reardon 2019).

The consequences of these unreasonable expectations can be very significant. For instance, patients' unsound expectations regarding the clinical outcomes of AIHTs can negatively affect their care experience (Alami et al., 2020). Certain areas of care, such as breast cancer, are particularly fertile ground to AI companies' hype and promises (Carter et al., 2020), because it resonates with patient unfulfilled demands. The counterweight to these expectations is not yet in place as, despite all the hype, the science of AI is still young and possibly not yet mature, including gaps in clinical validation and perhaps imprecise health recommendations (Dzobo et al., 2020). HTA agencies and regulators have an important role to play, particularly in developing a regulatory infrastructure that is as exceptional as technology can be for health and as powerful as the "unfounded hype" can be, to use Mazurowski (2019) expression.

New Ethical Challenges

There seems to be broad agreement that AIHTs present new ethical challenges (Vollmer et al., 2020). According to Michie et al. (2017), AIHTs presents "new challenges and new versions of old challenges" which require new evaluative methods and

legislative motivation to address health data and AIHT-specific ethical and regulatory issues.

Health Care Delivery

Reddy et al. (2020) identified three key AIHT-stemmed ethical challenges in care delivery: AI-fostered potential bias, patient privacy protection and trust of clinicians and the general public towards machine-led medicine. AI is also prone to generating new health inequalities; perhaps more potent than its ability to reduce existing ones (Fenech et al., 2020). An important caveat in terms of health care equity comes from the fact that those who compare very well with historic patient data will be the one benefiting the most from AIHTs. A cautious attitude is therefore called for with regards to patient and disease heterogeneity, taking into account that patterns detected by AI are largely deduced from smaller historical data sets (Dzobo et al., 2020). In addition to the current disparities, digital literacy and access to technologies are adding up, so that if nothing is done, large segments of the population may be excluded from enjoying the benefits of AIHTs, resulting in significant issues of justice (Fenech et al., 2020).

Existential Questions

According to Fenech et al., 2020, AI is unlike any other health technology, due to its capability of being a general-purpose technology forcing to question the very essence of humans. This technology is particularly sensitive for the healthcare sector as it raises new existential questions that regulators and public decision-makers must now face. One of such key existential challenge for HTA, according to Haverinen et al. (2019), is that AI is becoming a new decision-maker. This adds an actor with a decision-making role on the fate of patients and the health care system in addition to the role of HCPs and increases the complexity of performing comprehensive HTA. For Ho (2020), a distinctive ethical concern that stems from AIHT is the technology's unparalleled autonomy, which intensifies ethical and regulatory challenges, especially in terms of patient safety. While this obviously raises questions about liability (who is at fault for harm, and who is responsible for explaining it and being accountable to patients), it also requires thorough thinking about appropriate ways to ensure that care is humane and respects the dignity of persons (Pesapane et al., 2018; Vayena et al., 2018; Fenech et al., 2020).

Challenging Medical Ethics' Ethos

Exceptionalism also stems from the fact that the field of medicine is structured around the transparency and explainability of clinical decisions, which poses serious problems for the acceptance, regulation and implementation of (too often inexplicable) AI in the health care system (Reddy et al., 2020). As Miller (2020) points out that technology insertion is never neutral, both the success of AI in health care and the integrity and reputation of health care professionals depend on the alignment between the ethos of medical ethics and the ability of AIHTs (notwithstanding its benefits and performance) to respond to the challenges that its very characteristics pose to the health care system (Reddy et al., 2020). It is therefore widely acknowledged that AI will have considerable benefits on health care (optimized process, increased quality, reduced cost, and expanded access) that will come at the price of raising ethical issues specific to the technology (Abràmoff et al., 2020). This moral cost and related ethical considerations partly explain that the field of AI ethics has recently "exploded" as academics, organizations and other stakeholders have been rushing to examine the ethical dimensions of AI development and implementation (Fiske et al., 2020). However, some are skeptical, such as Fenech et al. (2020) who was warning that data ethics is fashionable. While Bærøe et al. (2020) go as far as arguing that "exceptional technologies require exceptional ethics" and that "an intentional search for exceptionalism is required for an ethical framework tasked with assessing this new technology".

New Evaluative Constraints

According to Dzobo et al. (2020), by being very distinct from more traditional technologies, AI must be regulated differently. Zawati and Lang (2020) argue that the uncertainty regarding AI decisional processes and outcomes make AIHTs particularly challenging to regulate. Regulators, policy-makers and HTA agencies are faced with unprecedented complexity for evaluating and approving AI (Alami et al., 2020). AIHTs raise new evaluative constrains, be they technological, clinical, organizational, that affect how ethical, legal and social dimensions may be tackled (Gerke et al., 2020). Evaluation constraints are related to data, real-world uses and the embryonic nature of the regulatory infrastructure and processes.

Data-Generated Issues

AI uses larger than ever volumes of data generated by individuals, governments, and companies, and according to Fenech et al. (2020), the greater complexity of health data raise new questions about the governance of data use and storage, especially as AI technologies are only effective and relevant with up-to-date, labelled, and cleaned big data. In addition to data, the hardware infrastructure will need to be updated over time, requiring major financial investments to maintain the use of AI in the healthcare system (Dzobo et al., 2020). However, too few technical studies are helping to appreciate and help managing AIHTs' complexity. In most studies, contextual, clinical and organizational considerations, implementation and uses of the technology are neglected, which complicates regulators' assessment as they are mostly informed about the significance of AI applications' technical performance (Alami et al., 2020). Caution should therefore be exercised, particularly since the complex ethical and regulatory issues involved deserve careful consideration before deploying these technologies in routine clinical care (Prabhu 2019).

Real-World Usages and Evidential Issues

AIHTs raise new regulatory challenges in part due to the greater variation in their performance between the test environment and the real-word context than those of drugs and medical devices. According to Gerke et al. (2020), AIHTs have potentially more risks and less certainty associated with their use in real-world contexts, which is central to regulatory concerns. However, most

AIHTs have not been objectively validated in and for real-world usages (Alami et al., 2020). In that sense, one important caveat is that the adoption and impact of AIHTs are unlikely to be uniform or to improve performance in all health care contexts (Gerke et al., 2020). This is attributable concurrently to the technology itself (and its distinctive features that renders it disruptive), to the contexts of implementation (the systemic impact of the technology across the health care system, and clash with local practices) and to the human biases associated with the use of these technologies (inability to reason with AI-provided probabilities, small samples and noise induced extrapolation and false patterns identification, and undue risk aversion) (Gerke et al., 2020). For regulatory authorities, these represent great challenges for deciding whether marketing authorization is justified. But it is also puzzling for hospital, clinic and health care system purchasers to determine whether an AIHT will actually add value and increase performance of care and service delivery. There is a lot to be studied and understood on the broad systemic policy implications of AIHTs in real-world context of care and services (Alami, Lehoux, Auclair, de Guise, et al., 2020).

Undeveloped Regulatory Infrastructure and Processes

AIHTs' exceptional characteristics have significant regulatory implications as regulation is emerging, but at a far slower pace than technological changes, which are virtually infinite (Char and Burgart 2020). Regulative complexity is furthered by the fact that existing standards (e.g., those of the Food and Drug Administration, European Medicines Agency, Health Canada) do not translate well for self-evolving technologies (Dzobo et al., 2020; Shaffer 2020; Topol 2020). This definitional deficit complicates the regulation of this technology and the implementation of appropriate policy infrastructure (Pesapane et al., 2018). Recent approvals of algorithms highlighted some limitations of existing regulatory standards and processes (Haverinen et al., 2019; Parikh et al., 2019). These considerations are threefold and concern the levels of requirements, the speed of AIHT developments and the equilibrium posture that regulators must adopt.

A challenge for existing regulatory regimes lies in the extensive information requirements on both the nature and effects of health technology, as well as clinical data on efficacy and patient safety, and population and societal impacts. However, regulators have yet to develop an infrastructure and processes that are appropriate and optimal for AI, and this requires more knowledge about how algorithms work (Dzobo et al., 2020). This complicates the problem because AI is a less transparent and explainable technology than drugs or medical devices can be (Abràmoff et al., 2020; Reddy et al., 2020). Privacy concerns are also important and there is yet no public agreement regarding data collection and sharing for commercial purposes, nor regarding for-profit data ownership (Michie et al., 2017). This calls for finding collective responses to these considerations, which must accompany the work of structuring HTA practices and infrastructures by regulators (Fenech et al., 2020).

Another dimension putting pressure on regulation is the speed of development. For regulating a fast-changing and unpredictably sector such as AI, time is of the essence to ensure that regulatory standards and practices are relevant (Pesapane et al., 2018). Currently, regulation has to constantly catch up with the private sector which leads to important gaps in terms of ethical examination of AIHTs (Shaffer 2020). Since, most developments are done by the private sector and HTA processes are not yet well designed and optimized, there is a problem of scrutiny (Abràmoff et al., 2020). So to keep up, regulation must be as fast as technological developments in AI, therefore it requires to conduct assessment and oversight at an unprecedented pace (Haverinen et al., 2019). However, this need to proceed quickly, in particular to match the private sector's pace, must be put into perspective with the very acceptability of a significant presence of commercial interests in the big data and AIHTs sector.

Achieving the right balance is delicate for HTA agencies between accelerating the development of HTA policies and procedures and not falling prey to the sirens of AIHT's hype (Cowie, 2017; Miller, 2020). Regulators want to see the health care system reap AI's benefits quickly, but if their assessment is too hasty and the implementation of the first generations of AIHTs encounters difficulties or, worse, generates adverse effects, social and professional acceptability may be shattered and further delay the deployment of AI in healthcare. According to Reddy et al. (2020), it could need a single serious adverse incident caused by an AIHT to undermine the public's and HCP communities' confidence. Considering that AI's acceptance is still fragile, and that AI is expected to have an expanded presence in all aspects of the health care system, HTA agencies will have to be extra careful in considering the ethical and regulatory implications of IA. If not well managed, these considerations will become major barriers that will play against the deployment of AI in healthcare (Pilotto et al., 2018; Vayena et al., 2018; Bærøe et al., 2020).

DISCUSSION

The current body of literature appears to portray AI health technologies as being exceptional to HTA. This exceptionalism is expressed along five dimensions. Firstly, the very nature of the technology seems to be the primary cause of the difficulty in fitting AIHTs into current HTA processes. Thus, the still ambiguous definition and the consequences of its changing and evolving nature pose new challenges for its assessment. Secondly, the scope of its impact far surpasses those of current health technologies. AIHTs will have impacts that extend significantly beyond the targeted patients and professionals. It is therefore in the interest of HTA agencies to consider the disruptive effects on individuals as well as on the entire health care system. Hence, the importance for HTA to consider the potential harms, the systematization of biases and to anticipate the clashes between current practices that are working well and the framing effect that will come with the deployment of AIHTs. But also, AI can act as a transformational lever that, beyond the risks of AI in healthcare, appears to be capable of redressing and reorienting the healthcare system to better respond to the full

range of healthcare needs, to create synergies so that the of learning healthcare systems are operational and to take this opportunity to adjust the regulatory design. Third, the advent of AI in healthcare comes with a lot of high expectations. The quality of outcomes generated by AIHTs is expected to be higher than that of current human-driven processes. This positive perception of the added value of AIHTs in the healthcare system makes AI in healthcare appear inevitable. However, while the technology is currently casted as exceptional and highly promising, some caution should be kept towards the current hype, which should prompt regulators to be prudent towards unreasonable expectations. Fourth, AIHTs are challenging HTA from an ethical perspective as AI has a strong potential to generate greater inequity whether arising from algorithmic decisions or in access to AIHTs. The fact that AIHTs are becoming new decision-makers, due to their autonomy, raises important issues of patient safety as well as liability. *In fine*, medical ethics' ethos is even shattered since, with AI, ethical dilemmas seem to be amplified, calling potentially for ethical standards revamping that would be as exceptional as the technology. Finally, AI technologies in health are increasing regulatory complexity and are pressuring current HTA structure and processes. The new evaluation constraints relate to data, real-world uses and the rather embryonic nature of AI-ready regulatory infrastructure and processes. Therefore, be they the extensive information requirements on both AIHTs' features and effects for regulatory review, the speed of AIHTs' developments, and the need to regulate quickly, but in a way that benefits the entire population.

A key point emerging from the views of the authors reviewed is that *exceptionalising* views of AIHTs, in the context of HTA, appear to come as much from the technology itself as they do from the broader social, cultural, and political contexts surrounding AI in the health sector. In other words, AIHTs are exceptional because of their technological characteristics *and* potential impacts on society at large. This is quite consistent with HTA, which seeks to assess the diversity of impacts of a technology. It is therefore quite reasonable that a technology with multi-dimensional impacts on society severely affects a process that is based on these same dimensions. The key takeaway may be that, to adapt and remain relevant, HTA must continue to focus on and strengthen these evaluative processes, which must be capable of a comprehensive assessment of the technical, social, cultural, ethical and health dimensions.

Interestingly, no author in the reviewed sample clearly promoted the idea that AI is an unexceptional technology for HTA. Many reasons could explain this phenomenon. First, the hype is still very strong when it comes to AIHT (Mazurowski 2019; Carter et al., 2020). Thus, it is possible that discussions about AI (un)exceptionalism are not yet ripe to mark the literature. This can be seen in the literature reviewed, which, without focusing on the limits of exceptionalism, currently has its strongest criticisms on AI hype. This leads us to think that hype and exceptionalism may be linked: hype feeds on exceptionalism while the latter needs hype to surface and to strike a chord within the literature and the rhetoric about AI in health. Second, trivially, this may be because there is less incentive to write (and publish) on the advent of a new technology in health by stating that nothing is new under the Sun (Caulfield 2018). Third, AIHTs may be so recent in the HTA pipeline that HTA as not yet addressed all the dimensions of AIHT.

Even if AI's exceptionalism appears significant in the current body of literature, there is some caveats to promoting AI exceptionalism in HTA. First, two authors noted some limitations to AI exceptionalism. Michie et al., (2017) pointed out that AIHTs are not only raising new challenges, they also bear issues that are common to existing health technologies. This is echoed by Char et al. (2020) who acknowledge the phenomenon, but puts a limit to the enthusiasm for AI exceptionalism in health when it comes to AIHTs sporting some features similar to standard health technologies. Second, currently, the literature discussing AI exceptionalism is still piecemeal, and it would be relevant for future research to address the issue by looking more holistically at the full range of issues posed by AI (i.e., outside the sole realm of HTA). There is still some space to apprehend and analyze the exceptionality of AIHTs' in HTA and the implications this has for both the evolution of HTA and the development and use of AIHTs. The literature is still quite young, and this can be observed from the fact that some highly discussed considerations in the broader AI ethics and AI in medicine literatures have not been discussed in the body of literature at review. For example, the more structural implications related to data-generated issues—privacy, data stewardship and intellectual property (Bartoletti 2019; Cohen and Mello 2019)—or to issues pertaining to informed consent, patient autonomy and human rights (Sparrow and Hatherley 2019; Cohen 2019; Racine et al., 2019; Ahmed et al., 2020), or to human-machine comparison in medical decision making and diagnosis (London 2019) have not been explored in the studied subset. A comprehensive exploration of the themes generally associated with health technologies will provide a better understanding of and clarity on whether AIHTs are exceptional or not. Third, exceptionalism in the context of health innovation is not a new topic. New health interventions or discoveries often generate a lot of hype, and the sector is hungry for predictions about the next revolution in healthcare (Emanuel and Wachter 2019). Twenty years ago, the health sector was living the "genomic revolution" and was assessing the exceptional implications of genetics in healthcare (Suter 2001). As AIHT's literature mature, it may continue to be centered on its exceptionality; but it is also plausible to consider that, as with the genetic revolution in the early 2000s and the nanotechnologies in the 2010s, AI exceptionalism will pass what Murray (2019) calls its "sell-by date". Thus, AI exceptionalism may end up following a rather similar pattern where the hype will slowly wear off as the health sector will become more accustomed to the technology; better understand its actual strengths, limitations, and capabilities.

Therefore, possibly it is a matter of time before a coherent body of literature addresses the limits of an *exceptionalist* view of AI in health and HTA. At the same time, if the AI revolution really takes off, exceptionalism will no longer be an important consideration. Indeed, as regulatory systems, the healthcare system and human agents (clinicians, patients, regulators,

managers, etc.) adapt, exceptionalism will probably pass and habituation will make AI in health the new normal, as after any major disruption that lasts over time. However, it may be of interest to the AI, HTA and health regulation communities and scholars to remain vigilant about AIHTs' exceptionalism (by means of the manifestation of its five dimensions) in health and HTA.

Another avenue that the literature could explore is whether the exceptionality arises from the technology (i.e., AI) or the sector of application (i.e., health)? In other words, is it AI that is exceptional in health or rather health that is a sector of exception for AI? Healthcare is possibly the most regulated sector that AI has come across so far. Health's exceptionality may explain the significant regulation in the healthcare sector (i.e., attention given to this sector in terms of regulation, ethics, law, and society) (Daniels 2001; Bird and Lynch 2019), while no other sector has an assessment process that has the breadth and systematism of HTA. Therefore, it would be interesting to reverse the question at the very basis of this review and consider how, for the AI ecosystem, health is *per se* exceptional and calls for additional and distinct norms, practices and contingencies that need to be considered to develop and implement an AIHT. Thereby, in addition to offering insights and guidance to communities strongly engaged in HTA, our results can also help the AI research and development sector better understand the unique evaluative considerations that exist in the health sector. The five dimensions raised by our paper can help guide those developing AIHTs to better understand and respond to the specific expectations and priors that underlie the use, administration and acceptability of health technologies. This can potentially help better align AIHT developers' desire to create value with HTA agencies' value appreciation and thus facilitate the congruence between technology development and healthcare priorities (Chalkidou 2021).

More broadly, the literature review raises key institutional questions, in terms of the exceptional issues posed by AI, such as which body is best placed to incorporate the new and added concerns that AIHTs raise? Is it the (supra)national regulators (e.g., EMA, FDA, Health Canada and the like which are mostly responsible for evaluating safety, efficacy, and quality concerns) or the HTA bodies (who are more concerned with appropriate use, implementation, coverage and reimbursement)? Can certain issues (e.g., the ethical and social ones) be better addressed by one body versus another? A possible limitation of the literature review is that overall the authors do not generally make a clear distinction between the regulatory processes (those of the FDA, EMA, and Health Canada that aim to allow the marketing of AIHTs) and the HTA (which focus on assessing implementation, optimal use, and whether or not to recommend reimbursement by third-party payers), so it was not possible to specify the unique considerations that arise specifically for either or both. One thing is for certain, the exceptional challenges of AIHT further raise the importance, for regulators and health technology assessors, to consider the impacts of AI uses in healthcare in a holistic way. This points to pivoting current rather linear regulatory and HTA process towards a "lifecycle" approach, which would allow for a better consideration of the five exceptional dimensions of AIHT. This may sound demanding, but AIHTs already represent additional evaluative burdens, especially when it comes to long-term real-world usages (e.g., when AIHT are used on new populations or for new purposes that differ from the data on which it was trained, or simply behave differently from what was expected at the time of the regulatory or HTA assessment) and the difficulties of withdrawing AIHT from the market. This calls for more cooperation between regulators and HTA agencies, but also hint towards a global health technology governance reform to allow increased scrutiny capability, and also to help AIHT regulatory and HTA assessment adjust overtime (i.e., by using a lifecycle approach) based on the evolving (clinical, economic, social, ethical) evidence.

In any case, there is a strong argument for taking into consideration the exceptional aspects of AIHTs, especially as their impacts on the healthcare system will be far greater than that of drugs and medical devices (Vayena et al., 2018; Babic et al., 2019). As AI applications begin to be much more readily introduced into the health care sector, there is a window of opportunity for HTA agencies and scholars to consider the broad spectrum of impacts that AIHTs may generate (Bostrom and Yudkowsky 2011; Helbing 2015; Burton et al., 2017; Herschel and Miori 2017; Knoppers and Mark Thorogood 2017). AI implementation by governments and health organizations carries risks of generating new and amplifying existing, challenges due to a shift from the current mostly human-driven systems to new algorithm-driven systems (Vayena et al., 2018; Zafar and Villeneuve 2018; Reddy et al., 2020). Hence the need to address the distinct (without the need for them to be exceptional) characteristics of AIHTs to inform HTA developments in a way to ensure that only clinically, economically, socially acceptable AIHTs are deployed in the health care system.

CONCLUSION

Therefore, is it possible to assert that there is such thing as an AI exceptionalism in HTA? It may be too early to be decisive on this issue, although the literature reviewed seems to point in this direction. Our review of the literature has allowed to identify five dimensions through which AIHTs are exceptional, from an HTA standpoint: nature, scope, increased expectations, new ethical challenges and new evaluative constrains. Most importantly, what underlies the promises of AI, the hype, and the exceptionalism is that we are mostly in an era of speculation; while some applications have begun to work their way into the healthcare system, the much-anticipated revolution is still a ways off. It is the test of time that will determine the veracity and breadth of the exceptionalist perspective. But whether or not exceptionalism proves to be valid, HTA must certainly adapt to the massive arrival of AI in health. This must be done by considering and responding to the five dimensions of exceptionalism and their many implications that can undermine the appropriateness, efficiency, and

relevancy of current and future HTA infrastructure and processes. Our results should help inform where HTA stakeholders need to pay special attention and adapt their policy architecture and processes so that they become agile to adopt a regulatory posture capable of appreciating the distinct characteristics and impacts that AIHTs pose in the health sector.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

J-CBP conceived and designed the study, collected the data, performed the analysis, wrote the first draft of the paper, and managed the revision process with the co-authors. VC and M-CR conceived and designed the study, contributed to data collect and analysis, and reviewed extensively various draft of the article. IG, MG, and IGC supported the analysis, and reviewed and contributed to the article.

FUNDING

IGC was supported by the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program, which is supported by grant NNF17SA0027784 from the Novo Nordisk Foundation. The SFU Open Access Fund has generously covered the publishing fee to make this article widely accessible.

ACKNOWLEDGMENTS

The authors would like to thank Renaud Lussier (INESSS) for supporting and validating the research strategy, Erica Monteferrante (McGill University) for helping conduct a preliminary search for planning the current study, and Matthew Herder (Dalhousie University) and Corélia Kostovic (INESSS) for providing insightful comments on the final version of the paper.

REFERENCES

Abràmoff, M. D., Tobey, D., and Char, D. S. (2020). *Lessons Learned about Autonomous AI: Finding a Safe, Efficacious, and Ethical Path through the Development Process American Journal of Ophthalmology*, 214. Iowa City, IA, United StatesDallas, TX, United: Department of Ophthalmology and Visual Sciences, University of IowaAbramoff) IDx Technologies, Coralville, IA, United States(Tobey) DLA PiperStates(Char) Division of Pediatric Cardiac Anesthes), 134–142. doi:10.1016/j.ajo.2020.02.022

Ahmad, O. F., Stoyanov, D., and Lovat, L. B. (2020). Barriers and Pitfalls for Artificial Intelligence in Gastroenterology: Ethical and Regulatory Issues. *Tech. Innov. Gastrointest. Endosc.* 22 (2), 80–84. doi:10.1016/j.tgie.2019.150636

Ahmed, Z., Mohamed, K., Zeeshan, S., and Dong, X. (2020). Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine. *Database : J. Biol. Databases Curation* 2020, 101517697. doi:10.1093/database/baaa010

Alami, H., Lehoux, P., Auclair, Y., de Guise, M., Gagnon, M.-P., Shaw, J., et al. (2020). Artificial Intelligence and Health Technology Assessment: Anticipating a New Level of Complexity. *J. Med. Internet Res.* 22 (7), e17707. doi:10.2196/17707

Alami, H., Rivard, L., Lehoux, P., et al. (2020). Artificial Intelligence in Health Care: Laying the Foundation for Responsible, Sustainable, and Inclusive Innovation in Low- and Middle-Income Countries. *Glob. Health* 16 (1), 52. doi:10.1186/s12992-020-00584-1

Allen, N., Walker, S. R., Liberti, L., and Salek, S. (2017). Health Technology Assessment (HTA) Case Studies: Factors Influencing Divergent HTA Reimbursement Recommendations in Australia, Canada, England, and Scotland. *Value in Health* 20 (3), 320–328. doi:10.1016/j.jval.2016.10.014

Babic, B., Gerke, S., Evgeniou, T., and Cohen, I. G. (2019). Algorithms on Regulatory Lockdown in Medicine. *Science* 366 (6470), 1202–1204. doi:10.1126/science.aay9547

Banta, D., and Jonsson, E. (2009). History of HTA: Introduction. *Int. J. Technol. Assess. Health Care* 25 (Suppl. 1July), 1–6. doi:10.1017/S0266462309090321

Bærøe, K., Jansen, M., and Kerasidou, A. (2020). Machine Learning in Healthcare: Exceptional Technologies Require Exceptional Ethics. *Am. J. Bioeth.* 20 (11), 48–51. doi:10.1080/15265161.2020.1820103

Bartoletti, I. (2019). *AI in Healthcare: Ethical and Privacy Challenges.* Cham: Lecture Notes in Computer ScienceSpringer International Publishing, 7–10. doi:10.1007/978-3-030-21642-9_2

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database. *Npj Digit. Med.* 3 (1), 1–8. doi:10.1038/s41746-020-00324-0

Bird, G., and Lynch, H. (2019). Introduction to the Politics of Life: A Biopolitical Mess. *Eur. J. Soc. Theor.* 22 (3), 301–316. doi:10.1177/1368431019838455

Bostrom, Nick., and Yudkowsky, Eliezer. (2011). "The Ethics of Artifical Intelligence.," in *The Cambridge Handbook of Artificial Intelligence*. Editors K. Frankish and W. M. Ramsey (Cambridge University Press), 316–35.

Braun, V., and Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* 3 (2), 77–101. doi:10.1191/1478088706qp063oa

Burton, E., Goldsmith, J., Koenig, Sn., Kuipers, B., Mattei, N., and Walsh, y. (2017). *Ethical Considerations in Artificial Intelligence Courses*. ArXiv:1701.07769 [Cs], January http://arxiv.org/abs/1701.07769.

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B., and Houssami, N. (2020). "The Ethical, Legal and Social Implications of Using Artificial Intelligence Systems in Breast Cancer Care, the Breast," in *((Carter) Australian Centre for Health Engagement, Evidence and Values (ACHEEV)* (Northfields Avenue, New South Wales 2522, Australia: School of Health and Society, Faculty of Social Science, University of WollongongRogers) Department of Philosophy and Dep)), 49, 25–32. doi:10.1016/j.breast.2019.10.001

Caulfield, T. (2018). Spinning the Genome: Why Science Hype Matters. *Perspect. Biol. Med.* 61 (4), 560–571. doi:10.1353/pbm.2018.0065

Chalkidou, A. (2021). A Different Animal but the Same Beast? Using Development-Focused Health Technology Assessment to Define the Value Proposition of Medical Technologies. *Int. J. Technol. Assess. Health Care* 37 (1), 1–2. doi:10.1017/S0266462321000209

Char, D. S., Abràmoff, M. D., and Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *Am. J. Bioeth.* 20 (11), 7–17. doi:10.1080/15265161.2020.1819469

Char, D. S., and Burgart, A. (2020). "Machine-Learning Implementation in Clinical Anesthesia, Anesthesia & Analgesia," in *(Char, Burgart) Division of Pediatric Anesthesia* (Stanford, CA, United States: Department of Anesthesia, Center for Biomedical Ethics, Stanford University School of Medicine), 130, 1709–1712. doi:10.1213/ANE.0000000000004656

Cohen, I. G. (2019). "Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? Symposium: Law and the Nation's Health. *Georgetown L. J.* 108 (6), 1425–1470.

Cohen, I. G., and Mello, M. M. (2019). Big Data, Big Tech, and Protecting Patient Privacy. *JAMA* 322 (12), 1141. doi:10.1001/jama.2019.11365

Cowie, M. R. (2017). Digital Health: Hype or Hope? *Dialogues Cardiovasc. Med.* 23 (1), 39–42.

Daniels, N. (2001). Justice, Health, and Healthcare. *Am. J. Bioeth.* 1 (2), 2–16. doi:10.1162/152651601300168834

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big Data in Healthcare: Management, Analysis and Future Prospects. *J. Big Data* 6 (1), 54. doi:10.1186/s40537-019-0217-0

Dzobo, K., Adotey, S., Thomford, N. E., and Dzobo, W. (2020). Integrating Artificial and Human Intelligence: A Partnership for Responsible Innovation in Biomedical Engineering and Medicine. *OMICS: A J. Integr. Biol.* 24 (5), 247–263. doi:10.1089/omi.2019.0038

Emanuel, E. J., and Wachter, R. M. (2019). Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA* 321, 2281–2282. doi:10.1001/jama.2019.4914

Fenech, M. E., Buston, O., and Buston, Olly. (2020). AI in Cardiac Imaging: A UK-Based Perspective on Addressing the Ethical, Social, and Political Challenges. *Front. Cardiovasc. Med.* 7 (101653388), 54. doi:10.3389/fcvm.2020.00054

Fiske, A., Tigard, D., Müller, R., Haddadin, S., Buyx, A., and McLennan, S. (2020). Embedded Ethics Could Help Implement the Pipeline Model Framework for Machine Learning Healthcare Applications. *Am. J. Bioeth.* 20 (11), 32–35. doi:10.1080/15265161.2020.1820101

Gerhards, H., Weber, K., Bittner, U., and Fangerau, H. (2020). Machine Learning Healthcare Applications (ML-HCAs) Are No Stand-Alone Systems but Part of an Ecosystem - A Broader Ethical and Health Technology Assessment Approach Is Needed. *Am. J. Bioeth.* 20 (11), 46–48. doi:10.1080/15265161.2020.1820104

Gerke, S., Babic, B., Evgeniou, T., and Cohen, I. G. (2020). The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device. *Npj Digit. Med.* 3, 101731738. doi:10.1038/s41746-020-0262-2

Grant, K., McParland, A., Mehta, S., and Ackery, A. D. (2020). Artificial Intelligence in Emergency Medicine: Surmountable Barriers with Revolutionary Potential. *Ann. Emerg. Med.* 75 (6), 721–726. doi:10.1016/j.annemergmed.2019.12.024

Hamet, P., and Tremblay, J. (2017). Artificial Intelligence in Medicine. *Metabolism* 69 (Suppl. ment), S36–S40. doi:10.1016/j.metabol.2017.01.011

Harwich, E., and Laycock, K. (2018). *Thinking on its Own: AI in the NHS.*" London: Reform.

Haverinen, J., Keränen, N., Falkenbach, P., Maijala, A., Kolehmainen, T., and Reponen, J. (2019). Digi-HTA: Health Technology Assessment Framework for Digital Healthcare Services. *FinJeHeW* 11 (4), 326–341. doi:10.23996/fjhw.82538

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nat. Med.* 25 (1), 30–36. doi:10.1038/s41591-018-0307-0

Helbing, Dirk. (2015). *"Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies." SSRN Scholarly Paper*. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2594352.

Herschel, R., and Miori, V. M. (2017). Ethics & Big Data. *Tech. Soc.* 49 (May), 31–36. doi:10.1016/j.techsoc.2017.03.003

Ho, C. W.-L. (2020). Deepening the Normative Evaluation of Machine Learning Healthcare Application by Complementing Ethical Considerations with Regulatory Governance. *Am. J. Bioeth.* 20 (11), 43–45. doi:10.1080/15265161.2020.1820106

Hu, M., Babiskin, A., Wittayanukorn, S., Schick, A., Rosenberg, M., Gong, X., et al. (2019). Predictive Analysis of First Abbreviated New Drug Application Submission for New Chemical Entities Based on Machine Learning Methodology. *Clin. Pharmacol. Ther.* 106 (1), 174–181. doi:10.1002/cpt.1479

Jameson, J. L., and Longo, D. L. (2015). Precision Medicine - Personalized, Problematic, and Promising. *N. Engl. J. Med.* 372 (23), 2229–2234. doi:10.1056/NEJMsb1503104

Knoppers, B. M., and Thorogood, A. M. (2017). Ethics and Big Data in Health. *Curr. Opin. Syst. Biol.* 4 (August), 53–57. doi:10.1016/j.coisb.2017.07.001

Kristensen, F. B., Lampe, K., Wild, C., Cerbo, M., Goettsch, W., and Becla, L. (2017). The HTA Core Model -10 Years of Developing an International Framework to Share Multidimensional Value Assessment. *Value in Health* 20 (2), 244–250. doi:10.1016/j.jval.2016.12.010

Leyens, L., Reumann, M., Malats, N., and Brand, A. (2017). Use of Big Data for Drug Development and for Public and Personal Health and Care. *Genet. Epidemiol.* 41 (1), 51–60. doi:10.1002/gepi.22012

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent. Rep.* 49 (1), 15–21. doi:10.1002/hast.973

Love-Koh, J., Peel, A., Rejon-Parrilla, J. C., Ennis, K., Lovett, R., Manca, A., et al. (2018). The Future of Precision Medicine: Potential Impacts for Health Technology Assessment. *PharmacoEconomics* 36 (12), 1439–1451. doi:10.1007/s40273-018-0686-6

Lysaght, T., Lim, H. Y., Xafis, V., and Ngiam, K. Y. (2019). AI-assisted Decision-Making in Healthcare. *Abr* 11 (3), 299–314. doi:10.1007/s41649-019-00096-0

Mason, J., Morrison, A., and Visintini, S. (2018). *"An Overview of Clinical Applications of Artificial Intelligence." 174. CADTH Issues in Emerging Health Technologies*. Ottawa: Canadian Agency for Drugs and Technologies in Health.

Mazurowski, M. A. (2019). Artificial Intelligence May Cause a Significant Disruption to the Radiology Workforce. *J. Am. Coll. Radiol.* 16 (8), 1077–1082. doi:10.1016/j.jacr.2019.01.026

McCradden, M. D., Joshi, S., Mazwi, M., and Anderson, J. A. (2020). Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning. *The Lancet Digital Health* 2 (5), e221–e223. doi:10.1016/S2589-7500(20)30065-0

Michie, S., Yardley, L. R., West, R., Patrick, K., and Greaves, F. (2017). Developing and Evaluating Digital Interventions to Promote Behavior Change in Health and Health Care: Recommendations Resulting from an International Workshop. *J. Med. Internet Res.* 19 (6), e232. doi:10.2196/jmir.7126

Miller, D. (2020). *Machine Intelligence in Cardiovascular Medicine Cardiology in Review*, 1120. Augusta, GA 30912: Douglas Miller) Department ofMedicine, Radiology and Population Health Sciences, Medical College of Georgia (Gb 3330)15th Street. United States: 53–64. doi:10.1097/CRD.0000000000000294

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G.The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Plos Med.* 6 (7), e1000097. doi:10.1371/journal.pmed.1000097

Murray, T. H. (2019). Is Genetic Exceptionalism Past its Sell-By Date? on Genomic Diaries, Context, and Content. *Am. J. Bioeth.* 19 (1), 13–15. doi:10.1080/15265161.2018.1552038

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., and Abedi, V. (2019). Artificial Intelligence Transforms the Future of Health Care. *Am. J. Med.* 132 (7), 795–801. doi:10.1016/j.amjmed.2019.01.017

O'Rourke, B., Oortwijn, W., and Schuller, T.the International Joint Task Group (2020). The New Definition of Health Technology Assessment: A Milestone in International Collaboration. *Int. J. Technol. Assess. Health Care* 36 (3), 187–190. doi:10.1017/S0266462320000215

Parikh, R. B., Obermeyer, Z., and Navathe, A. S. (2019). Regulation of Predictive Analytics in Medicine. *Science* 363 (6429), 810–812. doi:10.1126/science.aaw0029

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., et al. (2009). The Coming of Age of Artificial Intelligence in Medicine. *Artif. Intelligence Med.* 46 (1), 5–17. doi:10.1016/j.artmed.2008.07.017

Pesapane, F., Volonté, C., Codari, M., and Sardanelli, F. (2018). Artificial Intelligence as a Medical Device in Radiology: Ethical and Regulatory Issues in Europe and the United States. *Insights Imaging* 9 (5), 745–753. doi:10.1007/s13244-018-0645-y

Pilotto, A., Boi, R., and Petermans, J. (2018). Technology in Geriatrics. *Age and Ageing* 47 (6), 771–774. doi:10.1093/ageing/afy026

Prabhu, S. P. (2019). Ethical Challenges of Machine Learning and Deep Learning Algorithms. *Lancet Oncol.* 20 (5), 621–622. doi:10.1016/S1470-2045%2819%2930230-X

Racine, E., Boehlen, W., and Sample, M. (2019). Healthcare Uses of Artificial Intelligence: Challenges and Opportunities for Growth. *Healthc. Manage. Forum* 32 (5), 272–275. doi:10.1177/0840470419843831

Reardon, S. (2019). Rise of Robot Radiologists. *Nature* 576 (7787), S54–S58. doi:10.1038/d41586-019-03847-z

Reddy, S., Allan, S., Coghlan, S., and Cooper, P. (2020). A Governance Model for the Application of AI in Health Care. *J. Am. Med. Inform. Assoc.* 27 (3), 491–497. doi:10.1093/jamia/ocz192

Reddy, S. (2018). Use of Artificial Intelligence in Healthcare Delivery. *EHealth - Making Health Care Smarter, November*. doi:10.5772/intechopen.74714

Samuel, G., and Derrick, G. (2020). Defining Ethical Standards for the Application of Digital Tools to Population Health Research. *Bull. World Health Organ.* 98 (4), 239–244. doi:10.2471/BLT.19.237370

Shaffer, L. (2020). WHO Wants to Bring Order to Health Data. *Nat. Med.* 26 (1), 2–3. doi:10.1038/s41591-019-0717-7

Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., and Ahmad, H. F. (2020). Big Data Analytics Enhanced Healthcare Systems: A Review. *J. Supercomput* 76 (3), 1754–1799. doi:10.1007/s11227-017-2222-4

Shaw, J., Rudzicz, F., Jamieson, T., and Goldfarb, A. (2019). Artificial Intelligence and the Implementation Challenge. *J. Med. Internet Res.* 21 (7), e13659. doi:10.2196/13659

Smith, H. (2020). Clinical AI: Opacity, Accountability, Responsibility and Liability. *AI Soc.* 36, 535–545. doi:10.1007/s00146-020-01019-6

Sparrow, R., and Hatherley, J. (2019). The Promise and Perils of Al in Medicine. *Int. J. Chin. Comp. Phil. Med.* 17 (2), 79–109. doi:10.24112/ijccpm.171678

Sun, T. Q., and Medaglia, R. (2019). Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare. *Government Inf. Q.* 36 (2), 368–383. doi:10.1016/j.giq.2018.09.008

Suter, S. M. (2001). *The Allure and Peril of Genetics Exceptionalism: Do We Need Special Genetics Legislation*, 79. " Washington University Law Quarterly, 669–748.

Tadavarthi, Y., Vey, B., Krupinski, E., Prater, A., Gichoya, J., Safdar, N., et al. (2020). The State of Radiology AI: Considerations for Purchase Decisions and Current Market Offerings. *Radiol. Artif. Intelligence* 2 (6), e200004. doi:10.1148/ryai.2020200004

Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., et al. (2018). Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J.* 69 (2), 120–135. doi:10.1016/j.carj.2018.02.002

The Lancet (2017). Artificial Intelligence in Health Care: Within Touching Distance. *The Lancet* 390 (10114), 2739. doi:10.1016/S0140-6736(17)31540-4

Topol, E. J. (2020). Welcoming New Guidelines for AI Clinical Research. *Nat. Med.* 26 (9), 1318–1320. doi:10.1038/s41591-020-1042-x

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine Learning in Medicine: Addressing Ethical Challenges. *Plos Med.* 15 (11), e1002689. doi:10.1371/journal.pmed.1002689

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., et al. (2020). Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness. *BMJ* 368 (March), l6927. doi:10.1136/bmj.l6927

Vreman, R. A., NaciMantel-Teeuwisse, H., Goettsch, W. G. H. G. M., Mantel-Teeuwisse, A. K., Schneeweiss, S. G., Leufkens, H. G. M., et al. (2020). Decision Making under Uncertainty: Comparing Regulatory and Health Technology Assessment Reviews of Medicines in the United States and Europe. *Clin. Pharmacol. Ther.* 108 (2), 350–357. doi:10.1002/cpt.1835

Wang, F., and Preininger, A. (2019). AI in Health: State of the Art, Challenges, and Future Directions. *Yearb. Med. Inform.* 28 (1), 016–026. doi:10.1055/s-0039-1677908

Wang, T., McAuslane, N., Liberti, L., Leufkens, H., and Hövels, A. (2018). Building Synergy between Regulatory and HTA Agencies beyond Processes and Procedures-Can We Effectively Align the Evidentiary Requirements? A Survey of Stakeholder Perceptions. *Value in Health* 21 (6), 707–714. doi:10.1016/j.jval.2017.11.003

Wild, C., Stricka, M., and Patera, N. (2017). Guidance for the Development of a National HTA-Strategy. *Health Pol. Tech.* 6 (3), 339–347. doi:10.1016/j.hlpt.2017.06.006

Zafar, A., and Villeneuve, S. (2018). "Adopting AI in the Public Sector: Turning Risks into Opportunities through Thoughtful Design." Brookfield Institute for Innovation + Entrepreneurship (Blog). Available at: https://brookfieldinstitute.ca/commentary/adopting-ai-in-the-public-sector April 25, 2018).

Zawati, M. n., and Lang, M. (2020). What's in the Box?: Uncertain Accountability of Machine Learning Applications in Healthcare. *Am. J. Bioeth.* 20 (11), 37–40. doi:10.1080/15265161.2020.1820105

Conflict of Interest: IGC serves as a bioethics consultant for Otsuka on their Abilify MyCite product, is a member of the Illumina ethics advisory Board, and serves as an ethics consultant for Dawnlight.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Medicine

ORIGINAL RESEARCH
published: 29 November 2021
doi: 10.3389/fmed.2021.758690

Development and Validation of a Deep Learning Algorithm to Automatic Detection of Pituitary Microadenoma From MRI

Qingling Li[1,2†], Yanhua Zhu[1†], Minglin Chen[3†], Ruomi Guo[4†], Qingyong Hu[5], Yaxin Lu[6], Zhenghui Deng[3], Songqing Deng[3], Tiecheng Zhang[7], Huiquan Wen[4], Rong Gao[1], Yuanpeng Nie[1], Haicheng Li[1], Jianning Chen[8], Guojun Shi[1], Jun Shen[9], Wai Wilson Cheung[10], Zifeng Liu[6], Yulan Guo[3*] and Yanming Chen[1*]*

[1] Department of Endocrinology and Metabolism, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [2] Department of VIP Medical Service Center, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [3] School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China, [4] Department of Radiology, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [5] Department of Computer Science, University of Oxford, Oxfordshire, United Kingdom, [6] Department of Medical Artificial Intelligence Center, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [7] Department of Magnetic Resonance, The Second Affiliated Hospital, Harbin Medical University, Harbin, China, [8] Department of Pathology, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [9] Department of Radiology, The Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China, [10] Department of Pediatrics, University of California, San Diego, San Diego, CA, United States

Background: It is often difficult to diagnose pituitary microadenoma (PM) by MRI alone, due to its relatively small size, variable anatomical structure, complex clinical symptoms, and signs among individuals. We develop and validate a deep learning -based system to diagnose PM from MRI.

Methods: A total of 11,935 infertility participants were initially recruited for this project. After applying the exclusion criteria, 1,520 participants (556 PM patients and 964 controls subjects) were included for further stratified into 3 non-overlapping cohorts. The data used for the training set were derived from a retrospective study, and in the validation dataset, prospective temporal and geographical validation set were adopted. A total of 780 participants were used for training, 195 participants for testing, and 545 participants were used to validate the diagnosis performance. The PM-computer-aided diagnosis (PM-CAD) system consists of two parts: pituitary region detection and PM diagnosis. The diagnosis performance of the PM-CAD system was measured using the receiver operating characteristics (ROC) curve and area under the ROC curve (AUC), calibration curve, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score.

Results: Pituitary microadenoma-computer-aided diagnosis system showed 94.36% diagnostic accuracy and 98.13% AUC score in the testing dataset. We confirm the robustness and generalization of our PM-CAD system, the diagnostic accuracy in the internal dataset was 96.50% and in the external dataset was 92.26 and 92.36%, the AUC was 95.5, 94.7, and 93.7%, respectively. In human-computer competition, the diagnosis performance of our PM-CAD system was comparable to radiologists with >10

OPEN ACCESS

Edited by:
Jingjing You,
The University of Sydney, Australia

Reviewed by:
Yuanyuan Qin,
Huazhong University of Science and Technology, China
Sarat Sunthornyothin,
Chulalongkorn University, Thailand

****Correspondence:***
Yanming Chen
chyanm@mail.sysu.edu.cn
Yulan Guo
guoyulan@sysu.edu.cn
Zifeng Liu
liuzf@mail.sysu.edu.cn

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *14 August 2021*
Accepted: *28 October 2021*
Published: *29 November 2021*

Citation:
Li Q, Zhu Y, Chen M, Guo R, Hu Q, Lu Y, Deng Z, Deng S, Zhang T, Wen H, Gao R, Nie Y, Li H, Chen J, Shi G, Shen J, Cheung WW, Liu Z, Guo Y and Chen Y (2021) Development and Validation of a Deep Learning Algorithm to Automatic Detection of Pituitary Microadenoma From MRI. Front. Med. 8:758690. doi: 10.3389/fmed.2021.758690

years of professional expertise (diagnosis accuracy of 94.0% vs. 95.0%, AUC of 95.6% vs. 95.0%). For the misdiagnosis cases from radiologists, our system showed a 100% accurate diagnosis. A browser-based software was designed to assist the PM diagnosis.

Conclusions: This is the first report showing that the PM-CAD system is a viable tool for detecting PM. Our results suggest that the PM-CAD system is applicable to radiology departments, especially in primary health care institutions.

Keywords: pituitary microadenoma, magnetic resonance imaging, deep learning, algorithm, computer-aided diagnosis

INTRODUCTION

A pituitary microadenoma (PM) is a tumor <10 mm in diameter. PMs can occur in either sex. As many as 10% of the population may have a microadenoma, but most do not cause symptoms (1, 2). However, some PMs cause symptoms by secreting hormones that exert harmful consequences, for example, in Cushing's disease, acromegaly, infertility, and hyperprolactinemia (1). Due to its small size and variable anatomical structure among individuals, the diagnosis of PM is not easy by applying the technique of MRI alone (3). Manual analysis of MRI data is usually biased and time-consuming, and the diagnostic accuracy is closely related to the experience of radiologists. A shortage of experienced radiologists may cause a delay in diagnosis and compromise the overall quality of service to patients with PM (4, 5). Deep learning has the potential to revolutionize disease diagnosis and management by improving the diagnostic accuracy of PM while reducing the workload of radiologists. The development of a convolutional neural network (CNN) has significantly improved the performance of image classification and object detection (6). Recent reports showed that a computer-aided diagnosis (CAD) system can accurately diagnose patients with pituitary adenoma from MR images (7–9). In this work, we have developed and validated an image-based deep learning model to aid the detection of PM.

MATERIALS AND METHODS

Ethical Approval

This study is approved by the research ethics committee of the Institute of Basic Research in Clinical Medicine, The Third Affiliated Hospital of Sun Yat-sen University ([2020]02-089-01). This research is registered at the Chinese Clinical Trials Registry (http://www.chictr.org.cn/index.aspx) with the number ChiCTR2000032762.

Data Collection and Pre-procession of MRI Data

The original intention to develop and validate the technique of deep learning algorithms assisting PM diagnosis was prompted by several misdiagnosed PM cases in our hospital (**Supplementary Figure 1**). We developed and validated an automatic diagnosis model for the detection of PM. The training set was a retrospective study, the data were extracted from January 2012 to September 2019 at The Third Affiliated Hospital of Sun Yat-sen University (TianHe and LuoGang hospital). The validation set 1 was a prospective temporal validation using data from October 2019 to April 2021 at The Third Affiliated Hospital of Sun Yat-sen University. Validation sets 2 and 3 are geographic prospective external validation with data from two additional hospitals (Sun Yat-sen Memorial Hospital of Sun Yat-sen University, and The Second Affiliated Hospital of Harbin Medical University) from March 2020 to April 2021. All data were recruited using the same inclusion and exclusion criteria.

The workflow diagram for the overall experimental design is in **Figure 1** and **Supplementary Figure 2**. Inclusion criteria were participants suffered from infertility (defined as the inability of a sexually active couple to achieve pregnancy within a year or more with regular unprotected intercourse) and at least exhibited one or more of the following clinical symptoms/signs (menstrual irregularity, amenorrhea, galactorrhea, premature ejaculation, erectile dysfunction, or hypogonadism). Exclusion criteria were as follows: lactation, pregnancy, with primary thyroid, adrenal and/or gonadal diseases, malignant tumors, pituitary macroadenoma, sellar/pituitary masses or cyst, congenital disease of the pituitary gland, pituitaries, and MR images without complete pituitary scan or with too many MRI artifacts. Further examination was performed on the participants. We measured serum hormone levels of the participants (such as prolactin, adrenocorticotrophic hormone, follicle-stimulating hormone, luteinizing hormone, serum thyroid-stimulating hormone, and growth hormone) and performed a pituitary MR examination on those participants. Patients with functional and non-functional PM and patients with normal pituitary function were included for further deep learning analysis. The coronal dynamic enhancement T1-weighted imaging (T1WI) sequences of MRI (DICOM) from those participants were downloaded with a standard image format according to the software and instructions of the manufacturer. All pituitary images were read by two junior neuroradiologists (with <10 years of professional experience) and one senior neuroradiologist (with >10 years of professional experience), and the final diagnosis was mutually agreed upon by all three neuroradiologists have then proceeded for further investigation. In the training set, all images present with PM or normal pituitary images were selected by four general radiologists (>5 years of professional experience) and reviewed by two neuroradiologists (with >10 years of professional experience). All images of coronal dynamic enhancement T1WI sequence were used for the validation set without additional human intervention. MRI was performed with a 1.5 or 3.0 T MRI unit

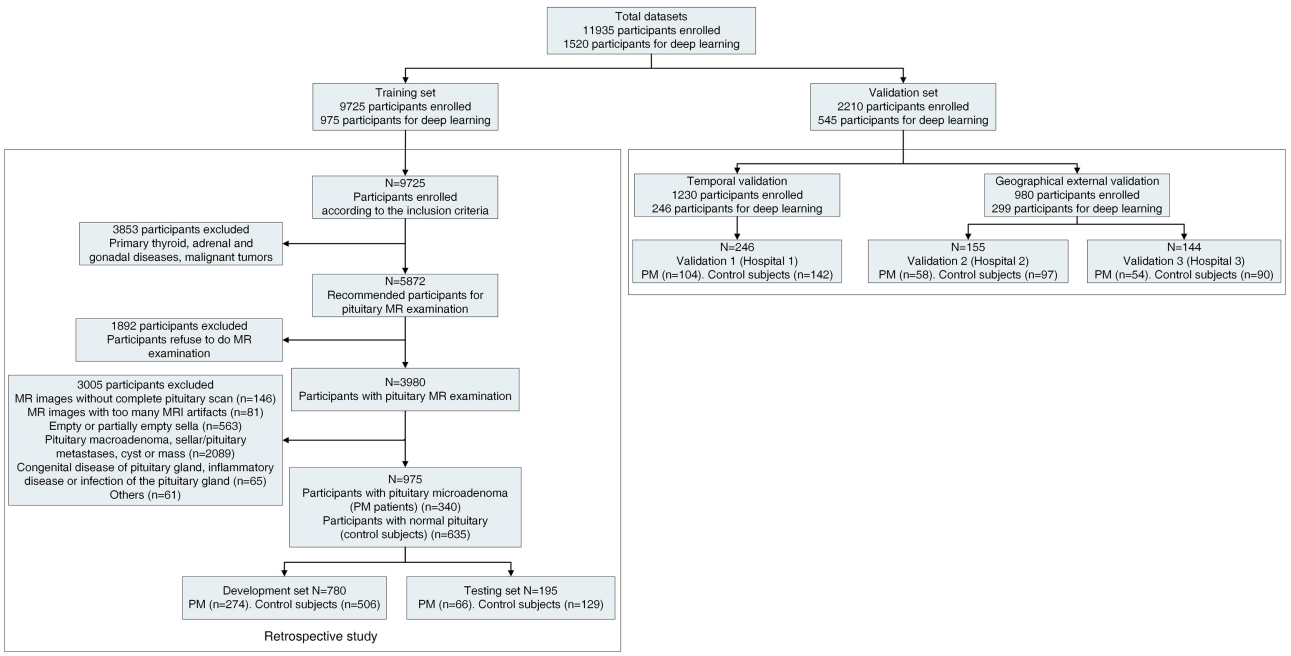

FIGURE 1 | Workflow diagram for the overall experimental design. The detailed workflow diagram of the validation datasets are in **Supplementary Figure 2**. PM, pituitary microadenoma; MRI, magnetic resonance imaging. The Third Affiliated Hospital of Sun Yat-sen University as hospital 1. Sun Yat-sen Memorial Hospital of Sun Yat-sen University as hospital 2, and The Second Affiliated Hospital of Harbin Medical University as hospital 3.

(GE, Philips company, Amsterdam, the Netherlands) in the head-first supine position, 380 ms/12.5 ms (repetition time /echo time), and 1 or 3 mm thick sections. Six medical fellows in the division of clinical endocrinology were involved in collecting patient clinical information, and the dataset was reviewed and verified by two endocrinologists.

Model Structure (Overview of Our PM-CAD System)

The pipeline of our PM-CAD system is shown in **Supplementary Figure 3**, and it consists of two parts: (1) pituitary region detection and (2) PM diagnosis. All programs are implemented with Python (https://www.python.org/) language on PyTorch (https://pytorch.org/) platform. In pituitary region detection, we develop a pituitary detection model based on Faster R-CNN (10) [with ResNet-50 FPN (11) as its backbone]. The input MR image is processed by this model to generate classification and regression maps, which have been further used to extract the pituitary bounding box in MR images. The pituitary bounding box is used to crop the pituitary region patch from the MR image (**Supplementary Method A**). In PM diagnosis, we proposed a novel CNN (namely, PM-CAD) to diagnose the PM from the cropped MR images. All the cropped pituitary region images are resized to 256 × 256, normalized into (0,1), and processed with histogram matching normalization (HM) for the enhancement of microadenoma features. In the PM-CAD system, we modify the ResNet architecture to preserve fine-grained features during forward propagation. An attention module is used to further improve the discriminativeness of feature representation. To handle the overfitting problem, HM normalization, intensity shift data augmentation, and label-smoothing loss are used (**Supplementary Method B**). The training procedure is stopped after 500 epochs (iterations through the entire dataset) due to the absence of further improvement in terms of both the area under receiver operating curve (AUC) and label-smoothing loss (**Supplementary Figure 4**).

Model Discrimination and Calibration

A total of 1,520 participants were included for the further study. We partitioned the data into three non-overlapping sets, with 780 participants for model development, 195 participants for model testing (developing and testing dataset as 8:2), and 545 participants for model validation. To reduce the time bias, the training set was a retrospective study from January 2012 to September 2019. The validation set was a prospective validation from October 2019 to April 2021. The detailed statistics for each set are summarized in **Figure 1** and **Supplementary Figure 2**.

Evaluation of the Diagnosis Performance of Our PM-CAD System and Statistical Analysis

In the testing set, we used accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score to evaluate our PM-CAD system. The validation set A had been used to evaluate the generalization ability and stability of our PM-CAD system. The receiver operating characteristics (ROC; showing both true-positive rate and false-positive rate for

diagnosis performance) curves and AUC were used in testing, internal and external validation sets (12, 13). We also used binary logistic regression methods to re-fit the prediction probability data rooted in PM-CAD, and calibration curves were used to test the fitting ability of the model (14). Validation set B consists of 100 participants and has been used to compare the performance of the PM-CAD system to general radiologists. A wide range of performance metrics has been adopted, such as diagnosis accuracy, sensitivity, specificity, PPV, NPV, F1-score, weighted error, positive likelihood ratio (PLR), negative likelihood ratio (NLR), and AUC (12). A weighted error was used for further analysis, specifically, a penalty weight of 2 was assigned to false-negative cases and a penalty weight of 1 was assigned to false-positive cases (12). Six radiologists were recruited for this study. Radiologists 1 and 2 have professional experience of <5 years, Radiologists 3 and 4 have professional experience between 5 and 10 years, and Radiologist 5 and 6 have professional experience over 10 years. Each radiologist read MR images of 100 participants independently. The Bland-Altman plot was used to evaluate the interobserver consistency of pituitary MRI finding independently measured by the six radiologists. The diagnostic accuracy of those radiologists was evaluated, and the experience of each radiologist in reading images of the cranial and pituitary MR or CT is shown in **Supplementary Table 1**. In validation set C, we tested the diagnosis accuracy of our PM-CAD system on three cases misdiagnosed by radiologists. Descriptive statistics included mean (SD) for continuous variables and proportions for categorical variables. All the metrics were calculated using Python-3.9.5 (https://www.python.org/), and R-4.0.3 (15) was used to provide visual analyses.

Browser-Based Software Application

A browser-based software was designed to assist the diagnosis from pituitary MR images. Once pituitary MR images (DICOM files) are uploaded to the software, PM diagnosis outputs can be presented.

RESULTS

Study Participants

A total of 11,935 infertility participants were initially recruited for this project. After applying the exclusion criteria, 1,520 participants (556 PM patients and 964 controls subjects) were included for further study whereby we have partitioned data from 975 participants (340 PM patients and 635 control subjects) for the training set, such as 780 participants (19,573 images) for development set and 195 participants (4,927 images) for the testing set. In the validation set, 545 participants (13,239 images) were recruited for the study. The validation set A consisted of 163 PM patients and 279 control subjects came from three hospitals. The validation set B consisted of 100 participants (50 PM patients, and 50 control subjects). In validation set C, we tested the diagnosis accuracy of our PM-CAD system on three misdiagnosed PM cases. The detailed statistics for each set are summarized in **Figure 1** and **Supplementary Figure 2**. Among patients with PM, there were 397 cases of non-functional PMs and 159 cases of functional PMs. The clinical and baseline characteristics of these participants are shown in **Table 1**.

Performance of PM-CAD System

The PM-CAD system consists of two parts: pituitary region detection and PM diagnosis. In pituitary region detection, we use the well-known average precision (AP) as the evaluation metric. We achieved an AP of 0.9783 at an intersection-of-union (IOU) threshold of 0.5 (**Supplementary Method A**). For testing the accuracy of PM diagnosis, 975 participants have been used for the development and testing set (**Supplementary Method B**). We showed that our PM-CAD system achieved an AUC of 98.13% (**Figure 2A**), an F1-score of 92.09%, an accuracy of 94.36%, a sensitivity of 96.97%, a PPV of 87.67%, a specificity of 93.02%, and an NPV of 98.36% on the testing set. The calibration curve of the testing set is listed in **Figure 3A**, the intercept on the testing is −6.098, and the probability weight W is 10.069. We employed PM-CAD for further investigation.

PM-CAD System Application in the Validation Set (Internal and External Datasets)

We used the internal and external datasets to validate the robust generalization performance of our PM-CAD system. The system was further tested in 442 participants from three different hospitals (Validation set A). The PM-CAD system achieved the diagnosis performance of AUC (95.46%) (**Figure 2B**), F1-score (97.30%), accuracy (96.50%), sensitivity (97.83%), PPV (96.77%), specificity (94.12%), and NPV (96.00%) in hospital 1. In hospital 2, the AUC is 94.72% (**Figure 2C**), F1-score is 93.62%, accuracy is 92.26%, sensitivity is 90.72%, PPV is 96.70%, specificity is 94.83%, and NPV is 85.94%, respectively. The diagnosis performance is AUC (93.70%) (**Figure 2D**), F1-score (93.71%), accuracy (92.36%), sensitivity (91.11%), PPV (96.47%), specificity (94.44%), and NPV is (86.44%) in hospital 3 (**Table 2**). The ROC curve is described in **Figures 2B–D**. The calibration curve of the validation set A is in **Figures 3B–D**, the intercept is −4.26, −3.465, and −2.963, respectively. And the probability weight W was 9.928, 11.06, and 9.909, respectively. The classification confusion matrices report the number of true positive, false positive, true negative, and false negative, which are resulted in **Supplementary Table 2**. We showed that our PM-CAD system achieves excellent diagnostic performance in internal and external datasets.

Performance of the PM-CAD System vs. Radiologists

An independent validation set B (100 participants: 50 PM patients and 50 controls from hospital 1) was used to compare the performance of the PM-CAD system vs. radiologists. For this comparison, six radiologists were recruited. The diagnosis performance of PM-CAD system is F1-score (93.88%), accuracy (94.00%), sensitivity (92.00%), PPV (95.83%), specificity (96.00%), and NPV is (92.31%) (**Supplementary Table 3**). In contrast, the performance of our best radiologist #6 is F1-score (94.95%), accuracy (95.00%), sensitivity (94.00%),

TABLE 1 | Description and characteristics of the training and validation datasets.

Characteristics	Training set				Validation set					
	Development set		Testing set		Temporal validation (hospital 1)		Geographical validation (hospital 2)		Geographical validation (hospital 3)	
	Patients	Controls	Patients	Controls	Patients	Controls	Patients	Controls	Patients	Controls
Full cohort	274	506	66	129	104	142	58	97	54	90
Sex [No. (%)]										
Male	56 (20.4)	98 (19.4)	13 (19.7)	30 (23.3)	19 (18.3)	25 (17.6)	12 (20.7)	22 (22.7)	10 (18.5)	17 (18.9)
Female	218 (79.6)	408 (80.6)	53 (80.3)	99 (76.7)	85 (81.7)	117 (82.4)	46 (79.3)	75 (77.3)	44 (81.5)	73 (81.1)
Age (Mean ± SD)	30.92 ± 6.56	31.26 ± 7.36	30.82 ± 6.02	30.58 ± 6.04	30.79 ± 6.86	30.66 ± 5.50	31.43 ± 7.61	30.86 ± 5.46	29.81 ± 4.78	30.14 ± 5.35
BMI (Mean ± SD)	23.07 ± 2.50	23.09 ± 2.52	22.85 ± 2.38	23.91 ± 2.48	23.20 ± 2.40	22.67 ± 2.31	23.42 ± 2.79	23.21 ± 2.57	23.27 ± 2.50	23.48 ± 2.66
Blood biochemical indices (Mean ± SD)										
PRL, ulU/mL	1,184.92 ± 1,353.99	321.14 ± 144.32	1,142.82 ± 1,332.77	302.21 ± 150.47	1,121.06 ± 1,362.23	301.31 ± 152.69	1,053.70 ± 1,346.33	329.89 ± 149.50	1,150.89 ± 1,280.17	304.69 ± 162.74
ACTH, pmol/L	5.69 ± 2.46	5.61 ± 1.80	5.48 ± 3.57	5.34 ± 1.82	5.91 ± 4.06	5.29 ± 2.03	5.40 ± 1.71	5.17 ± 1.69	5.35 ± 1.54	5.12 ± 1.70
FSH, mIU/mL	4.73 ± 2.32	4.72 ± 2.04	5.02 ± 2.27	4.53 ± 2.01	5.17 ± 1.94	4.47 ± 2.26	5.53 ± 2.10	4.58 ± 2.07	5.14 ± 2.26	4.49 ± 2.00
LH, mIU/mL	4.24 ± 2.02	4.39 ± 1.93	4.34 ± 2.32	4.32 ± 1.82	4.97 ± 2.22	4.12 ± 1.98	5.80 ± 2.13	4.34 ± 1.75	4.79 ± 1.84	4.38 ± 1.79
TSH, ulU/mL	2.07 ± 0.93	2.48 ± 1.21	2.10 ± 0.89	2.19 ± 1.08	2.02 ± 0.81	1.92 ± 0.84	1.99 ± 1.47	1.96 ± 0.85	1.90 ± 0.79	2.08 ± 0.87
MRI examination and PM Functional diagnosis [No. (%)]										
Normal pituitary of MRI scan	–	506	–	129	–	142	–	97	–	90
PM of MRI scan	274	–	66	–	104	–	58	–	54	–
Non-functional PM	194 (70.8)	–	47 (71.2)	–	75 (72.1)	–	42 (72.4)	–	39 (72.2)	–
Functional PM	80 (29.2)	–	19 (28.8)	–	29 (27.9)	–	16 (27.6)	–	15 (27.8)	–
PRL-PM	76	–	17	–	24	–	15	–	15	–
ACTH-PM	3	–	2	–	3	–	0	–	0	–
GH-PM	1	–	0	–	2	–	0	–	0	–
TSH-PM	0	–	0	–	0	–	1	–	0	–

Data are mean (S.D.) or a number of individuals (%). BMI, Body Mass Index; PRL, Prolactin; ACTH, adrenocorticotrophic hormone; FSH, Follicle-Stimulating Hormone; LH, Luteinizing Hormone; TSH, Serum Thyroid-stimulating Hormone; GH, Growth hormone; MRI, Magnetic Resonance Imaging; PM, pituitary microadenoma.—means the participants did not calculate.

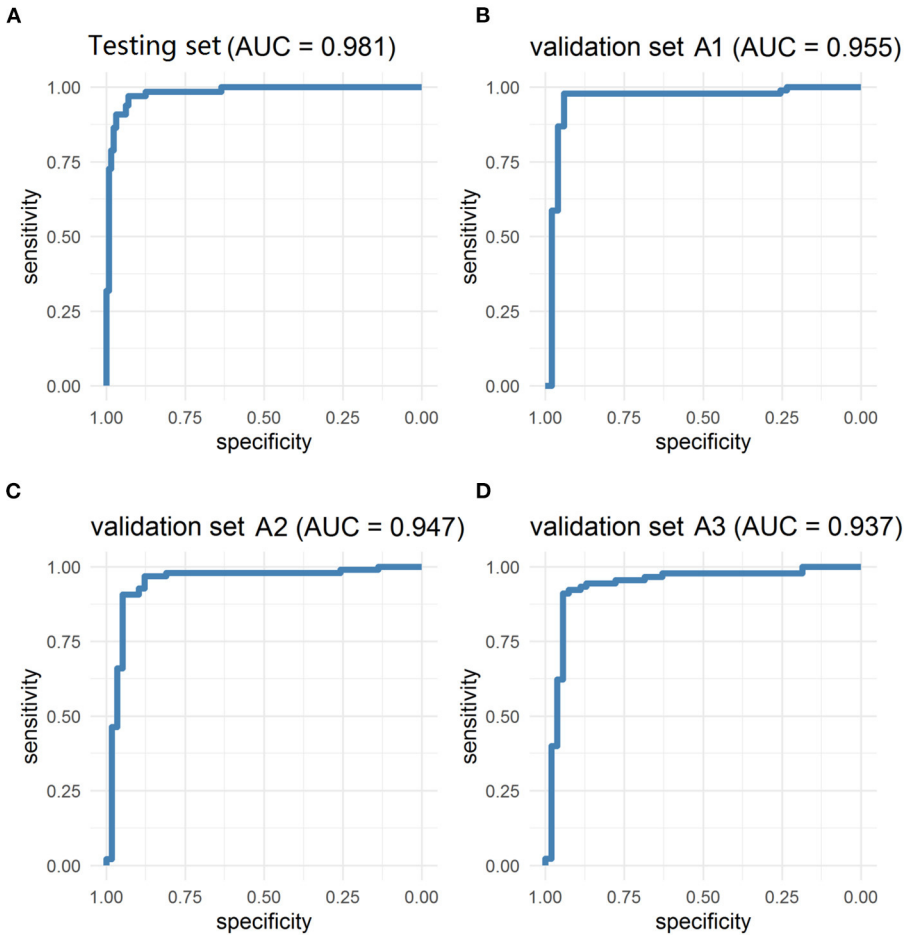

FIGURE 2 | The ROC curves of testing and validation set A1 (Internal dataset), validation set A2 and A3 (external dataset). The model has achieved excellent diagnosis performance in internal and external data sets. **(A)** The AUC of the testing set was 98.13%. **(B)** The validation set A1 is a temporal internal dataset, the AUC was 95.46%. **(C,D)** In the geographical external dataset, the AUC of the validation set A2 and A3 was 94.72 and 93.70%, respectively. AUC, area under the ROC curve; ROC, the receiver operator curve.

PPV (95.92%), specificity (96.00%), and NPV is (94.12%) (**Supplementary Table 3**). The ROC curves are shown in **Supplementary Figure 5A**, the AUC of the PM-CAD system was 95.56% and outperformed our six radiologists (best radiologist #6 as 95.00%), at the same false-positive rate, the true positive rate of the PM-CAD system was higher than six radiologists (**Supplementary Figure 5A**). Weighted error scoring (10) was incorporated during modeling and evaluation, the PM-CAD system produces a weighted error of 10.00%, which is far below the average weighted error of 21.67% achieved by six radiologists (**Supplementary Figure 5B**). The difference of NLRs or PLRs (10) between our PM-CAD system and radiologists is shown in **Supplementary Figures 5C,D**, our model demonstrates excellent diagnostic performance. The classification confusion matrices report the number of true positive, false positive, true negative, and false negative resulted for the PM-CAD system and radiologists in **Supplementary Table 4**. Thus, we showed that the diagnosis performance of our PM-CAD system is comparable to general radiologists with more than 10 years of professional experience. A Bland-Altman plot was used to analyze the interobserver consistency of the six radiologists' independent measurements of the pituitary MRI finding. The 95% limits of agreement were −0.4500 to 0.4300, −0.2958 to 0.2558, −0.1860 to 0.2060, −0.1860 to 0.2060, −0.1860 to 0.2060, and −0.2060 to 0.1860, respectively, indicating high interobserver consistency.

Further Assessment for the Diagnosis Performance of the PM-CAD System

We sampled three double positive cases of PM (both diagnosed by radiologists and PM-CAD system), which underwent surgical treatment, the double positive cases were confirmed by a subsequent pathological examination (one case of Cushing's disease, one case of Acromegalia, and one case of prolactinoma; **Supplementary Figure 6A**).

A false-negative diagnosis leads to delay in treatment of PM, PM-CAD system showed 100% diagnosis accuracy of detecting three clinically misdiagnosed PM cases which subsequently underwent surgical treatment (two cases of Cushing's disease and

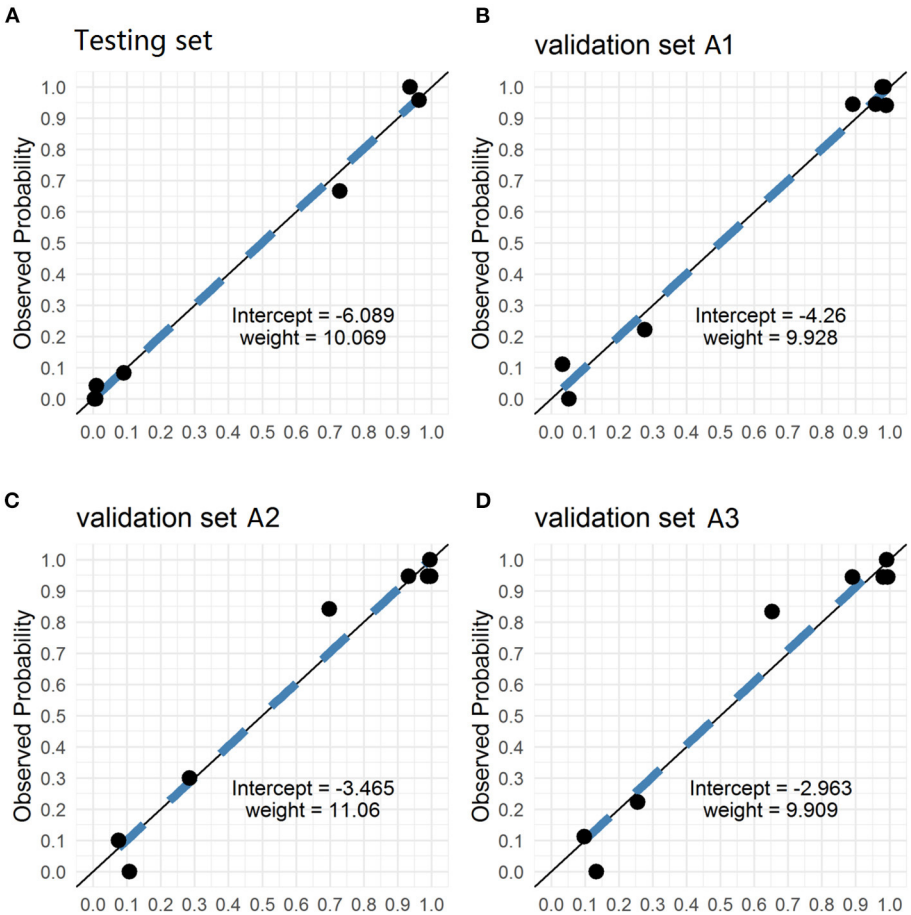

FIGURE 3 | The Calibration curves of testing and validation set A1 (Internal dataset), validation set A2 and A3 (external dataset). The calibration curves of the predicted probability from our PM-CAD vs. the observed probability for PM in **(A)** the testing set, **(B)** the validation set A1, **(C)** the validation set A2, and **(D)** the validation set A3. We used logistic regression to rebuild the prediction probability from our CNN model. The intercepts on the testing and verification set A are −6.098, −4.26, −3.465, and −2.963, respectively. And the probability weight W is 10.069, 9.928, 11.06, and 9.909, respectively. CNN, convolutional neural network; PM-CAD, Pituitary microadenoma-computer-aided diagnosis.

one case of thyroid-stimulating hormone, TSH, secreting PM; **Supplementary Figure 6B**). The diagnosis of thE misdiagnosed PM was confirmed by histopathology examination and relevant clinical information (**Supplementary Figure 6** and **Supplementary Table 5**).

Browser-Based Software Application

The browser-based software was designed to assist the PM diagnosis of pituitary MR images from different hospitals, which is hosted at http://www.pituitarymicroadenoma.com. Even without graphics processing unit (GPU) acceleration, the application takes only 1–2 s to analyze all MR images from a patient. Once DICOM files (the coronal dynamic enhancement T1-weighted imaging (T1W) sequence) are uploaded to the software, PM diagnosis outputs can be presented. The software interface is presented in **Supplementary Figure 7**. In a prospective study, we have tested the efficacies of our PM-CAD in the division of endocrinology in our hospital. Our results indicate that the PM-CAD system is an excellent screening test for the presence of PM. Over a period of 1 month, our PM-CAD system was able to detect the presence of 11 PM patients with a 97% accuracy rate (of 48 infertile patients and 25 patients with pituitary MR examination).

DISCUSSION

In this work, we developed a deep learning system (namely, PM-CAD) to diagnose PM from MRI. As we know, it is the first attempt to focus on PM diagnosis by using deep learning, although similar works have been proposed for pituitary adenoma (7–9, 16). Diagnosis of PM is challenging due to its tiny size and various anatomical structure (1–3). We found that our PM-CAD system can accurately diagnose PM from MRI without additional information, the system achieves a 96.5% diagnostic accuracy, which is comparable to radiologists with over 10 years of professional expertise.

TABLE 2 | The diagnosis performance of the PM-CAD system in the validation set A (internal and external datasets).

Evaluation metrics	Validation set A (set A1: internal dataset, hospital 1)	Validation set A (set A2: external dataset, hospital 2)	Validation set A (set A3: external dataset, hospital 3)
AUC (95% CI)	0.9546 (0.9028–0.9923)	0.9472 (0.8978–0.9858)	0.9370 (0.8821–0.9802)
Sensitivity	0.9783 (0.9237–0.9974)	0.9072 (0.8312–0.9567)	0.9111 (0.8324–0.9608)
Specificity	0.9412 (0.8376–0.9877)	0.9483 (0.8562–0.9892)	0.9444 (0.8461–0.9884)
Accuracy	0.9650 (0.9203–0.9885)	0.9226 (0.8687–0.9594)	0.9236 (0.8674–0.9613)
PPV	0.9677 (0.9086–0.9933)	0.9670 (0.9067–0.9931)	0.9647 (0.9003–0.9927)
NPV	0.9600 (0.8629–0.9951)	0.8594 (0.7498–0.9336)	0.8644 (0.7502–0.9396)
F1 score	0.9730 (0.9381–0.9912)	0.9362 (0.8912–0.9666)	0.9371 (0.8903–0.9682)

AUC, the area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value.

Several previous works have attempted to analyze pituitary adenoma using MRI. Ugga et al. (9) used a machine learning method to extract MRI-based radiomics to predict the proliferative index of pituitary macroadenomas. Qian et al. (7) employ a CNN network to diagnose pituitary adenoma from MRI, they evaluated a 149 participants dataset, which includes pituitary macroadenoma and microadenoma. Wang et al. (16) created an automated segmentation method for the sellar region, several tools to extract invasiveness-related features of pituitary adenoma and evaluate their clinical usefulness by predicting the tumor consistency. In this study, we focus on the diagnose of PM from the PM-CAD system with a large dataset. We show that our PM-CAD system outperforms the model developed by Qian et al. (7). Because of our PM-CAD system can specifically extract PM features from pituitary MR images and trained with more data. In addition, our model was validated in three hospitals and showed excellent generalization ability.

Strengths and Limitations

Our work has the following strengths. First, we showed that this PM-CAD system is a rapid, reliable tool to diagnose PM with a high accuracy in both internal and external datasets. Second, PM diagnosis requires experienced radiologists, but the exhausting workload raises the misdiagnose rate. Our PM-CAD system can be used as an assistant tool to reduce the workload of radiologists. Our PM-CAD system achieves comparable diagnostic accuracy to experienced radiologists and can make a decision in 1–2 s. Third, medical resources are not evenly distributed, that is, experienced radiologists mostly worked in economically developed areas hospitals while economically underdeveloped areas are lack experienced radiologists (4, 5). Our online accessible PM-CAD system can provide PM diagnosis to these areas and improve their PM diagnostic capabilities. Last, training a radiologist is costly and time consuming. It usually takes more than 10 years to train a qualified radiologist (4, 5). Our PM-CAD system is trained from annotated data and takes few time (about 30 s per patient) to improve its performance when more data are provided.

Our PM-CAD system remains several problems to be solved. First, although our PM-CAD system achieves a 96.5% diagnostic accuracy, this implies that 3.5% of cases may potentially be misdiagnosed in practice. To further improve the diagnosis performance of the PM-CAD system, more data should be collected and used to train our models. Second, when more new data are available, it would be better than our PM-CAD system can perform model self-update, a continual learning approach can be introduced to keep our system learning. Third, MRI scan data are unique to patients, with privacy concerns, these data are not allowed to distribute out of the hospitals. Therefore, our PM-CAD system cannot be fine-tuned in a specific hospital. In future work, we will use a federated learning framework to fine-tune our models in a privacy-preserving manner.

CONCLUSIONS

In summary, we have developed a deep learning-based system (namely, PM-CAD) to detect PM from MRI. A Total of 1,520 participants datasets have been used to train, validate, and test our system. Our PM-CAD system achieves a diagnostic accuracy comparable to radiologists with over 10 years of professional expertise. In the study, our PM-CAD system shows excellent generalization ability. Results from this work highlight the potential applications of deep learning on the diagnosis of patients with PM. With the rapid development of computing power, deep learning algorithms can surpass the gold diagnosis standard for the detection of PM. Machine learning for the diagnosis of PM will serve as an important component in improving patient care and outcomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

CODE AVAILABILITY STATEMENT

The software and code of the proposed method have been separated into two files and are available as Supplementary Software files. https://github.com/MinglinChen94/PituitaryMicroadenomaDiagnosis.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee of the Institute of Basic Research in Clinical Medicine, The Third Affiliated Hospital of Sun Yat-sen University ([2020]02-089-01). This research is registered at the Chinese Clinical Trials Registry (http://www.chictr.org.cn/index.aspx) with the number ChiCTR2000032762. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YC and YG: have full access to all the data in the study, take responsibility for the integrity of the data and the accuracy of the data analysis, administrative, technical, material support, and supervision. QL, YZ, MC, ZL, and GS: concept and design. QL and MC: drafting of the manuscript. WC and GS: critical revision of the manuscript for important intellectual content. ZL, MC, and QL: statistical analysis. YC, YG, and RGu: obtained funding. All authors acquisition, analysis, or interpretation of data.

FUNDING

This study was funded by the National Key R&D Program of China (2017YFA0105803), the National Natural Science Foundation of China (U20A20185, 81770826, 61972435, and 81801757), the Key Area R&D Program of Guangdong Province (2019B020227003), the Natural Science Foundation of Guangdong Province (2019A1515011271, 2019A1515012051, and 2018A030310322), the Science and Technology Plan Projects of Guangzhou (202007040003), and the Science and Technology Innovation Committee of Shenzhen Municipality (JCYJ20190807152209394).

ACKNOWLEDGMENTS

We thank all the patients and investigators for their participation in this study. We thank Chunkui Shao and Jing Wang (Professor, Department of Pathology and radiology, The Third Affiliated Hospital of Sun Yat-sen University) for their assistance in the research. We thank six fellows and six radiologists involved in data collecting and human-computer competition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.758690/full#supplementary-material

Supplementary Method | A detailed description of PM-CAD Model. PM-CAD, Pituitary microadenoma-computer-aided diagnosis.

Supplementary Figure 1 | Cases of 4 misdiagnosed pituitary microadenomas. **(A)** 4 consecutive pituitary MRI scans over a period of 20 months in a misdiagnosed patient with pituitary microadenoma. The radiologists have not detected the pituitary microadenoma during the first 3 MRI examinations. A functional microadenoma has been localized by the subsequent ACTH examination of the inferior petrosal sinus in the region of right pituitary gland. On the 4th MRI scanning, two microadenoma are detected by radiologist. **(B)** Additional 3 cases of misdiagnosed microadenoma. Patient 1 has a very small microadenoma with a diameter < 3 mm. Patient 2 has an irregularly shaped microadenoma. Patient 3 has two microadenoma (with diameters of 2.8 mm and 6.1 mm, respectively) and the smaller one was misdiagnosed. The comprehensive clinical data for patients were listed in **Supplementary Table 1**. ACTH, Adrenocorticotropic Hormone; MRI, magnetic resonance imaging; T1WI-COR, T1 weighted imaging-coronal. MRI bar = 5 mm. The yellow arrow and the area inside the red circle represent adenomas.

Supplementary Figure 2 | Workflow diagram for the validation datasets. PM, pituitary microadenoma; MRI, magnetic resonance imaging.

Supplementary Figure 3 | Overview of our PM-CAD system. **(A)** First the MR images are fed into our PM-CAD system for automatic diagnosis. The proposed PM-CAD system consists of two models: **(B)** the pituitary detection model localizes the pituitary region in cerebral MRI. The MR images are processed with multiple convolutional layers and two maps (classification map is used to predict the center and the regression map is used to refine the height and width of the rectangle box) are produced to predict a rectangle box enclosing the pituitary region. The pituitary rectangle region is cropped, stacked, and then fed into the PM diagnosis model. **(C)** It employs the proposed PM-CAD model to extract features. A softmax layer is employed to transform the feature into the presence probability of PM. CAD, computer-aided diagnosis; MRI, magnetic resonance imaging; MR, magnetic resonance; PM, pituitary microadenoma.

Supplementary Figure 4 | Performance of the PM-CAD system on the training datasets. **(A)** Accuracy curves achieved by the PM-CAD system on the development and testing datasets. **(B)** Cross entropy loss curves achieved by the PM-CAD system on the development and testing datasets. We train the PM-CAD system for 500 epochs.

Supplementary Figure 5 | The PM-CAD system outperforms 6 radiologists in AUC of PM diagnosis. **(A)** ROC and AUC: ROC curve shows the true positive rates (sensitivity) with respect to different false-positive rates (1-specificity). The ROC curve shows that the PM-CAD system outperforms 6 radiologists. The AUC of PM-CAD system is 95.6% better than our best radiologist#6 (AUC 95.0%). **(B)** Weighted error. A penalty weight of 2 is applied to false-negatives and a penalty weight of 1 is assigned to false-positives. The PM-CAD system produces a weighted error of 10%, whereas the radiologists produce a weighted error of 21.67%. **(C,D)** The negative likelihood ratio and the positive likelihood ratio: The negative likelihood ratio is defined as the false-negative rate over the true negative rate, so that a decreasing likelihood ratio < 1 indicated increasing probability the absence of PM. The positive likelihood ratio is defined as the true positive rate over the false-positive rate, so that an increasing likelihood ratio > 1 indicated increasing probability the diagnosis of PM. The confidence intervals show that the PM-CAD system demonstrates statistically better screening performance in terms of both negative likelihood ratio and positive likelihood ratio than radiologists. Radiologist 1 & 2: with < 5 years professional experience, Radiologist 3 & 4: with 5 - 10 years professional experience, Radiologist 5 & 6: with > 10 years professional experience. PM, pituitary microadenoma; receiver operating characteristics (ROC); the area under ROC curve (AUC).

Supplementary Figure 6 | The MRI and histological validation of double positive and false-negative cases. **(A,B)** 3 double positive and 3 false-negative cases, which were functional PM, as confirmed by subsequent pathological examination. The comprehensive clinical data for these patients are listed in **Supplementary Table 5**. PM, pituitary microadenoma; MRI, magnetic resonance imaging; AI, Artificial intelligence; HE, hematoxylin and eosin; ACTH, adrenocorticotropic hormone; GH, growth hormone; TSH, thyroid stimulating hormone; PRL, prolactin. MR bar = 5mm. Pathology bar =100 μm. The yellow arrow indicates a pituitary microadenoma.

Supplementary Figure 7 | The browser-based software to aid the diagnosis of PM. As long as we upload the pituitary MR images (DICOM), the software will tell you whether the patient suffering from PM disease. This browser based tool can be accessed at http://82.157.181.77/.

Supplementary Table 1 | The workload of radiologists with different professional experience in human-computer competition. All participating radiologists are

general radiologists (no specialization). Workload analysis was performed on the participating radiologists for 1 year.

Supplementary Table 2 | Confusion Matrices for testing and validation of dataset A (internal and external datasets). Data are numbers of images. a, true-positive; b, false-positive; c, false-negative; d, true-negative.

Supplementary Table 3 | The diagnostic performance for Human-computer competition according to temporal validation set B ($n = 100$). Unless otherwise specified, data are percentages, with numbers of images in parentheses and 95% confidence intervals in brackets. F1 score, the harmonic mean of PPV and sensitivity. NPV, negative predictive value; PPV, positive predictive value. Radiologist 1 & 2, < 5 years professional experience; Radiologist 3 & 4, 5 - 10 years professional experience; Radiologist 5 & 6, > 10 years professional experience.

Supplementary Table 4 | Confusion Matrices for Human-computer competition according to temporal validation set B ($n = 100$). Data are numbers of images. a, true-positive; b, false-positive; c, false-negative; d, true-negative.

Supplementary Table 5 | The patient clinical data in **Supplementary Figures 1**, **6**. PM, pituitary microadenoma; BMI, Body Mass Index; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; HR, Heart Rate; TSH, Serum Thyroid-stimulating Hormone; FT4, Free T4; FT3, Free T3; TSTO, Testosterone; PRL, Prolactin; PRGE, Progesterone; LH, Luteinizing Hormone; E2, Estradiol; GH, Growth hormone; IGF-1, Insulin-like Growth factor-1; COR, cortisol; ACTH, adrenocorticotrophic hormone; PZC24, 24-hour urine free cortisol; IPSS, inferior petrosal sinus sampling; MRI, Magnetic Resonance Imaging. FT3 (range 3.5–6.5 pmol/L). FT4 (range 11.5–22.7 pmol/L). TSH (range 0.55–4.78 uIU/mL). TSTO (range female 0.5–2.6, male <50 years 4.94–32.01 nmol/L). FSH (range female 2.5–10.2, male 0.95–11.95 mIU/mL). PRL, (range female 59–619, male 72.66–407.4 uIU/mL). PRGE (range female 0.5–4.5, male 0.2–1.040 nmol/L). LH (range female, 1.9–12.5, male 0.57–12.07 mIU/mL). E2 (range female, 71.6–529.2, male 40.4–161.5 pmol/L). GH (range <8 ng/mL). IGF-1 (range 116–358 ng/mL). COR (8 Am range 118.6–618 nmol/L 0.4 Pm range 85.3–459.6 nmol/L). ACTH (8 Am range <10 pmol/L). PZC24 (range 153.2–789.4 nmol/ 24-h)—means the patient did not measured.

REFERENCES

1. Molitch ME. Diagnosis and treatment of pituitary adenomas: a review. *J Am Med Assoc.* (2017) 317:516–24. doi: 10.1001/jama.2016.19699
2. Molitch ME. Nonfunctioning pituitary tumors. *Handb Clin Neurol.* (2014) 124:167–84. doi: 10.1016/B978-0-444-59602-4.00012-5
3. Wang H, Hou B, Lu L, Feng M, Zang J, Yao S, et al. PET/MRI in the diagnosis of hormone-producing pituitary microadenoma: a prospective pilot study. *J Nucl Med.* (2018) 59:523–8. doi: 10.2967/jnumed.117.191916
4. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ.* (2017) 359:j4683. doi: 10.1136/bmj.j4683
5. Rimmer A. BMA urges more career flexibility and better occupational support to fight workforce crisis. *BMJ.* (2017) 358:j4381. doi: 10.1136/bmj.j4381
6. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst.* (2018) 42:226. doi: 10.1007/s10916-018-1088-1
7. Qian Y, Qiu Y, Li CC, Wang ZY, Cao BW, Huang HX, et al. A novel diagnostic method for pituitary adenoma based on magnetic resonance imaging using a convolutional neural network. *Pituitary.* (2020) 23:246–52. doi: 10.1007/s11102-020-01032-4
8. Niu J, Zhang S, Ma S, Diao J, Zhou W, Tian J, et al. Preoperative prediction of cavernous sinus invasion by pituitary adenomas using a radiomics method based on magnetic resonance images. *Eur Radiol.* (2019) 29:1625–34. doi: 10.1007/s00330-018-5725-3
9. Ugga L, Cuocolo R, Solari D, Guadagno E, D'Amico A, Somma T, et al. Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning. *Neuroradiology.* (2019) 61:1365–73. doi: 10.1007/s00234-019-02266-1
10. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39:1137–49. doi: 10.1109/TPAMI.2016.2577031
11. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Honolulu, HI. (2017). p. 2117–25. doi: 10.1109/CVPR.2017.106
12. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell.* (2018) 172:1122–31.e9. doi: 10.1016/j.cell.2018.02.010
13. Zhou LQ, Wu XL, Huang SY, Wu GG, Ye HR, Wei Q, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology.* (2020) 294:19–28. doi: 10.1148/radiol.2019190372
14. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *J Am Med Assoc.* (2017) 318:1377–84. doi: 10.1001/jama.2017.12126
15. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing (2020). Available online at: https://www.R-project.org/ (acessed March 19, 2021).
16. Wang H, Zhang W, Li S, Fan Y, Feng M, Wang R. Development and evaluation of deep learning-based automated segmentation of pituitary adenoma in clinical task. *J Clin Endocrinol Metab.* (2021) 106:2535–46. doi: 10.1210/clinem/dgab371

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Medicine

REVIEW
published: 25 January 2022
doi: 10.3389/fmed.2021.784455

Data Integration Challenges for Machine Learning in Precision Medicine

Mireya Martínez-García[1] *and Enrique Hernández-Lemus*[2,3]*

[1] Clinical Research Division, National Institute of Cardiology 'Ignacio Chávez', Mexico City, Mexico, [2] Computational Genomics Division, National Institute of Genomic Medicine (INMEGEN), Mexico City, Mexico, [3] Center for Complexity Sciences, Universidad Nacional Autnoma de Mexico, Mexico City, Mexico

A main goal of Precision Medicine is that of incorporating and integrating the vast corpora on different databases about the molecular and environmental origins of disease, into analytic frameworks, allowing the development of individualized, context-dependent diagnostics, and therapeutic approaches. In this regard, artificial intelligence and machine learning approaches can be used to build analytical models of complex disease aimed at prediction of personalized health conditions and outcomes. Such models must handle the wide heterogeneity of individuals in both their genetic predisposition and their social and environmental determinants. Computational approaches to medicine need to be able to efficiently manage, visualize and integrate, large datasets combining structure, and unstructured formats. This needs to be done while constrained by different levels of confidentiality, ideally doing so within a unified analytical architecture. Efficient data integration and management is key to the successful application of computational intelligence approaches to medicine. A number of challenges arise in the design of successful designs to medical data analytics under currently demanding conditions of performance in personalized medicine, while also subject to time, computational power, and bioethical constraints. Here, we will review some of these constraints and discuss possible avenues to overcome current challenges.

Keywords: precision medicine, machine learning, data integration, meta-data mining, computational intelligence

OPEN ACCESS

Edited by:
Yuguang Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Juan M. Banda,
Georgia State University, United States
Sridhar Goud,
National Institutes of Health (NIH), United States

***Correspondence:**
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *27 September 2021*
Accepted: *28 December 2021*
Published: *25 January 2022*

Citation:
Martínez-García M and Hernández-Lemus E (2022) Data Integration Challenges for Machine Learning in Precision Medicine. Front. Med. 8:784455. doi: 10.3389/fmed.2021.784455

1. INTRODUCTION

Contemporary biomedical research and medical practices are increasingly turning into data-intensive fields, for which computational intelligence approaches, such as those based on artificial intelligence and machine learning (AI/ML) methods are becoming the norm. Due to the specific nature of these fields, the integration and management of the ever-growing volumes of heterogeneous data involved, often presents a number of challenges. These challenges become even more relevant in the light of the importance that AI/ML are gaining, establishing themselves at the core of the state-of-the-art in biomedical research and clinical medicine (1–3), as well as public health and healthcare policy (4–6).

From the standpoint of biomedical research, a number of large, data-intensive collaborative projects, such as the International Hap Map project (7, 8), The Cancer Genome Atlas (TCGA) (9–12), the 1000 Genomes (1000G) study (13–16), the GTEX consortium (17–19), and the Human Cell Atlas (HCA) (20, 21), and others are establishing novel frameworks for the molecular study

of health and disease. Such frameworks are firmly supported by robust database management and integration strategies that are allowing them to develop into central tools for basic and translational biomedical research.

Relevant as genomics and high throughput molecular studies are for biomedicine, there are other relevant players in the medical data arena. Among the more important in the present context are large scale clinical and phenotypic studies. Large clinical cohorts creating data-intensive outputs are of course not new, but the extent of their outreach and the complexity of the resulting data sets are growing exponentially fast. Starting from large scale clinical surveys, such as the Framingham Heart study (22, 23), the Wellcome Trust Case Control Consortium (24) and moving unto efforts like the UK Biobank that combines large scale clinic and phenotypic data with ultra-high-throughput genomic testing (25–28) that for the last 15 years has been generating massive data corpora used for their own means but also encouraging and feeding other data-intensive analytical efforts from genetic disease association (29) to brain imaging (30) to psychology (31) and social determinants of health (32), to name just a few instances. It goes without saying that the impact that these projects have reached on the basic and clinical settings, but also in the epidemiology and public health areas has been enormous.

In the context of AI/ML, however, the focus is shifting into *translating* the astronomical amounts of data generated ultimately into *products* and *policies* able to impact both the patients' and the general public health. This has been, for instance, one of the central goals of the U.S. initiative in Personalized Medicine (33, 34). That is, how to develop analytic strategies—many of them founded on automated learning, essential, given the size of complexities of current health-related data corpora—to pass from large scale, heterogeneous data to useful (even actionable) medical information (35).

Aside from large scale, even multi-national efforts—such as the ones in the consortia just discussed—, another area of intensive interest regarding data-mining in medicine has been the development of analytical strategies to effectively mine the ever growing body of Electronic Health Records (EHR), that has been perceived as a largely forgone and under-utilized data source (6, 36–39).

One main challenge in knowledge discovery from EHRs is that electronic medical records are highly heterogeneous data sources with a complex array of quantitative, qualitative, and transactional data. Disparate data types include ICD codes (mainly used for pricing and charging hospital procedures), biochemical and lab tests, clinical (text-based) notes, historical archives of medical interventions, therapies and even pharmaceutical deliveries. These data sources are often captured by dozens of individuals (sometimes with biased criteria) for each instance. Hence EHR data is quite difficult to analyze, in particular if one is looking (as is often the case if AI/ML techniques are being considered) multi-patient institutional and even multi-centric levels.

In brief, EHRs were not developed to be used as a resource for automated learning so they are not designed with *data structures* in mind. Since EHRs are first and foremost adapted for clinical and hospital logistics, data modeling and learning will often face challenges related to structural heterogeneity from their early stages, either by adapting existing EHR strategies or by re-designing them (40–44).

In the quest for more efficient healthcare interventions, based on information-optimized clinical practice and policy, AI/ML will certainly play a key role in going from a medicine approach—based mainly in the skills of the well-trained clinician—to one based also in detailed (often automated) analysis of the individualized interplay of molecular interactions and physiological traits with environmental and even social elements, thus, delivering the promise of personalized medicine (1, 2, 45, 46). The development of this analytic approach to personalized medicine (often termed *Precision Medicine*) involves a number of theoretical frameworks from systems biology to computational biology, biomedical informatics, and computational medicine. This is so, since health and healthcare are multi-dimensional in nature, hence, their study must consider information at the genetic, molecular, clinical and population levels. Health and healthcare analytics, however, must also evaluate and assess how to cope with the complexity and natural biases of the plethora of medical-related databases in which said molecular, clinical, and epidemiological data resides. This, again, points out to the need of customized, scalable computational and analytical tools for pattern discovery and hypothesis generation and testing. AI/ML is turning into a cornerstone of personalized medicine (6, 47–49).

In order to present a panoramic view on how these and other challenges may be overcome toward an optimized application of machine learning and artificial intelligence to analyze biomedical and health-related data in a Precision Medicine context, the rest of this work will proceed as follows: The next section (The role of data in training good AI/ML models) will establish the necessity to have proper data as input to machine learning and AI models useful in Precision Medicine. We will discuss how having very large data corpora (a.k.a *Big Data*) is great, but often carries with it the so-called *curse of dimensionality* and the need to perform *feature selection*, i.e., to select relevant pieces of information among very large and complex databases. We will also elaborate on the challenges created by diverse and heterogeneous data types and sources, bringing problems, such as *class imbalance* (study groups of sometimes extremely disparate sizes, that are problematic to analyze for many machine learning algorithms).

The following section (Precision medicine: transforming biomedical evidence with data analytics) will outline how the tenets of computational intelligence and machine learning may be used to advance medicine turning it (even more) into a full-evidence based science. We will see that in order to impact biomedical research, clinical practice and public policy, AI/ML approaches could be helpful to extend our capacities to generate biomedical knowledge, contribute to knowledge dissemination, translate personalized medicine into clinical practice and even empowering the patients. In order to develop, large scale data analytics in medicine should be able to become *translational*, i.e., moving faster from research environments to clinical settings to ultimately benefit the patients. Then, we will move on in the next section, to discuss the main challenges involved in the use of computational learning toward Precision Medicine.

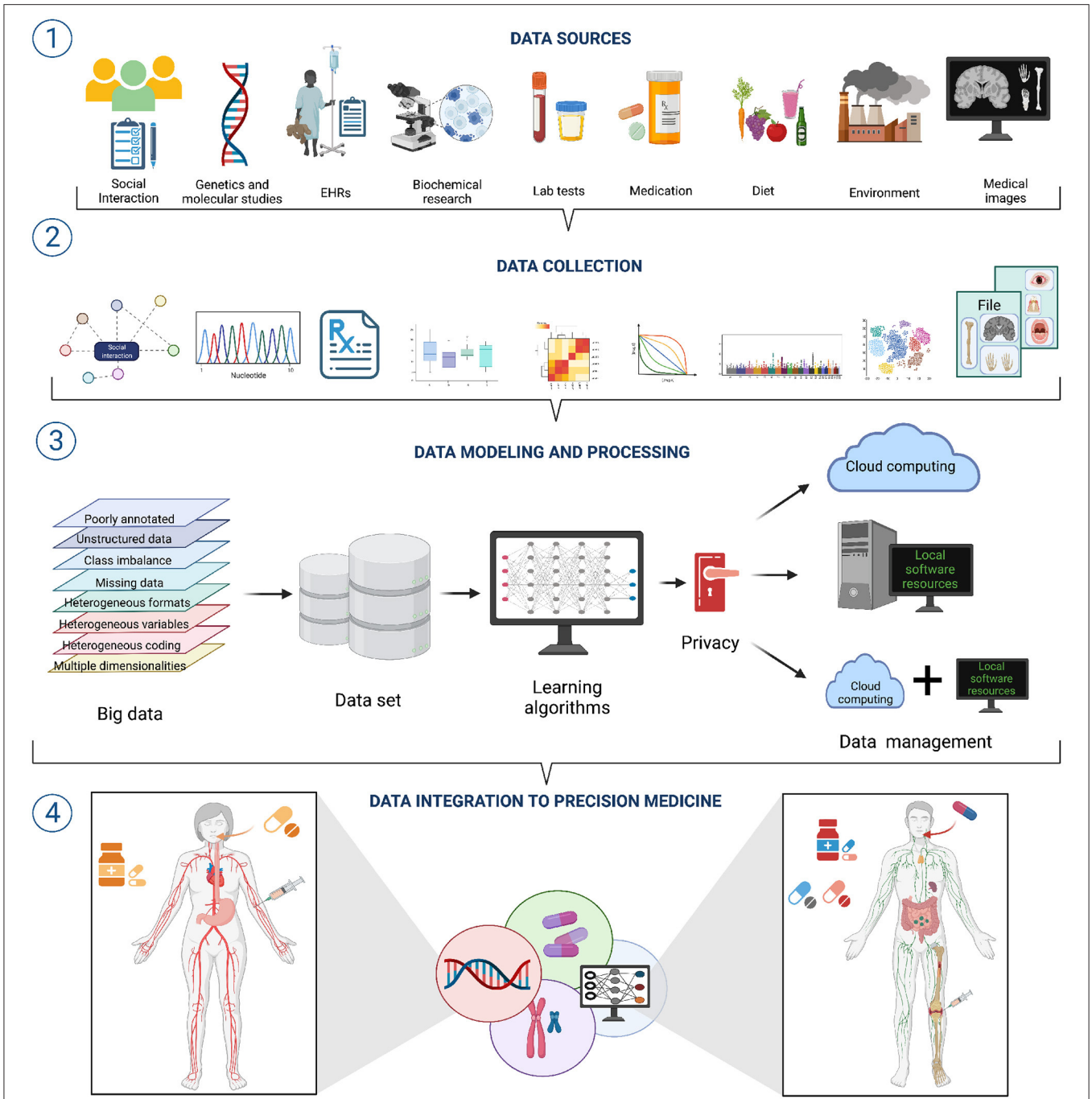

FIGURE 1 | A workflow for data integration for AI/ML modeling in precision medicine. (1) A wide variety of data sources with diverse features exists. Hence, different approaches to data collection and pre-processing are needed (2). (3) Integrating such diverse and heterogeneous data is one of the grand challenges to the successful application of AI/ML approaches to Precision Medicine. Overcoming such challenges will bring important improvements to Precision Medicine (4).

These include processing heterogeneous and unstructured data, working on collaborative and cloud-based resources, developing standards for data sharing and collaboration, implementing software solutions to support large scale data analytics under the biomedical and clinical diverse data ecosystems.

Section 5 will deal with one of the main challenges involved in the quest to effectively implement AI/ML in Precision Medicine: Data Integration. Biomedical and clinical knowledge deals with a plethora of phenomena, ranging from the molecular to the socio-political. Currently, we have technologies to massively

measure or infer data from most of these domains. How to *make sense* of these different dimensions to turn them into a coherent, intelligible body of knowledge useful for the researchers, but more importantly, for practising clinicians, the healthcare providers and the patients is an extremely challenging endeavor. Interestingly, a source of information that is becoming key for AI/ML approaches in Precision Medicine is metadata. Metadata, i.e., auxiliary data sources often used to define other data types. Having one's genome sequence is of little use if we do not have a proper *annotation* file; and knowledge of the zip code or educational level of a patient may provide actual clues for their *personalized* treatment. Since many data types are actually pre-processed prior to the analysis, it is also relevant to know how has the data been treated prior to its current form. Information of this kind is also considered metadata. Metadata is, hence, becoming more and more relevant. Managing such large amounts of *personal* data (what can be more personal for us than our healthcare data?), however, does not come without a price. Ethical and legal considerations pose no small problem if one is to provide fair and *minimally invasive* use of the data, especially if it is of a sensible or private nature. Some of these issues are discussed in section 6. Section 7 is devoted to present the Data Management Plan, a document that will be extremely useful to set the guidelines of any data-intensive project being a research protocol, a clinical trail or a healthcare management design. Finally, in section 8, we present some Conclusions and Perspectives.

2. THE ROLE OF DATA IN TRAINING GOOD AI/ML MODELS

The current development of highly sophisticated and often quite effective AI/ML and the accompanying proliferation of large scale data sources in the biomedical setting, has raised the expectations regarding the many potential benefits that can be derived from the marriage of *good methods + good data*. However, in order for these large amounts of data to be useful in producing good AI/ML models, size is not the only thing that matters, a question that is often overlooked (50, 51).

Clinical and biomedical data comes in a wide variety of sizes, forms, and formats; it is often complex, heterogeneous, poorly annotated, and often unstructured. Now, each of these issues: size, variety, formatting, complexity, heterogeneity, bad annotation, and lack of structure, pose a challenge to effective AI/ML modeling (see **Figure 1** section (1)) (52).

Regarding size, for instance, even when we often deal with *big data*—usually considered an advantage—, it is common that these data sources suffer from the so-called *curse of dimensionality* (CoD), a situation in which the number of variables or features is much larger than the number of experimental samples or realizations. CoD is particularly evident in the case of genomic and transcriptomic analyses for which the number of genes or transcripts is in the order of tens of thousands whereas the number of samples is rarely larger than a few hundreds or a few thousands at most. Even more complex is the scenario when one is measuring, for instance, chemical modifications, such as DNA methylation; the current experimental protocols allow for the simultaneous measures of five thousand hundred or more methylation probes (52).

CoD leads to the $p >> n$ problem in machine learning (53): increased data dimensionality may cause AI/ML methods to suffer from **overfitting**. Overfitting, in turn, implies that the methods are highly accurate on training data while showing low performance on generalization or handling unseen data. Potentially good methods will fail to deliver in real life applications. One approach to deal with the CoD is performing data dimensionality reduction prior to training the ML methods. The most common means of data dimensionality reduction are *feature extraction* in which data is projected from a high dimensional space to a lower dimensional space and *feature selection* that reduces dimensions by identifying a relevant or *informative* subset of the original set of features (54).

Feature extraction methods, such as principal component analysis (PCA) and other methods based on eigenvalue decompositions, non-negative matrix factorization (NNF), *t*-distributed stochastic neighbor embedding (t-SNE) and others, allow for easier data visualization, exploration, and compression, as well as latent factor profiling. On the other hand, feature selection methods consists in one or more of the following strategies: data filtering (DF), data wrapping (DW), and data embedding (DE). The purpose the former (DF) is to select a subset of relevant features in a model independent fashion an include methodological approaches, such as ANOVA, Pearson's correlation, information theoretical measures, such as entropy and mutual information, constrained regressions, and maximal relevance minimal redundancy (mRMR) methods. DW methods look for the best combination of features trained by a particular predictive model and include the recursive feature elimination (RFE), jackstraw and the Boruta-Random Forests (BRF). DE are a combination of DF and DW that works by performing feature selection while building a predictive model, perhaps the best known example of DE method is the least absolute shrinkage and selection operator (LASSO) and its extensions, such as the elastic net algorithm (52).

Data variety/diversity and data heterogeneity also result problematic for the implementation of AI/ML modeling in Precision Medicine. Heterogeneity emerges from many situations, such as substantially different types of variables (or different coding) in the various data sets (think of EHRs from different hospitals), mismatched distributions or scaling including disparate dynamic ranges (say we have combined expression data from microarrays and RNASeq technologies), diverse data modalities (continuous signals, counts, intervals, categories, pathways, etc., derived from molecular and imaging experiments) and formats (say European versus American reporting standards) (**Figure 1** section (2)). Integrating heterogeneous data types may be done naively, by just concatenating features from disparate data sources, but this reduce the number of working to the use of decision tree (DT)—like models that suffer from overfitting. An alternative would be to use penalized regression (e.g., elastic nets) with several

regularization strategies, though this may in turn bring challenges regarding interpretability of results (51, 52). Better results may be obtained by resorting to block-scaling (55) or multiple kernel learning methods (56).

Due to the complexity intrinsically associated to biomedical and clinical data, but also due to difficulties in subject/sample procuration and in data acquisition (data generating/sampling technologies may fail) it is common to have problematic circumstances, such as missing data (from instances not measured or measured defectively), class imbalance (widely different sample sizes in different feature groups) and even rarity (an extreme form of class imbalance) (57). There are several learning strategies to cope with missing data and class imbalance, ranging from the so-called *listwise deletion* (i.e., completely deleting the problematic sample from the study), imputation (i.e., inferring the missing value from expectation methods from the sample-wise profiles or even from feature-wise profiles) suing methods, such as *k*-nearest neighbor replacement, full conditional specification, stochastic gradient boosted trees, and other ensemble regression frameworks (52).

Class imbalance is another problematic-yet-pervasive situation in large scale data analytics (LSDA) of biomedical and clinical data. This fact becomes quite relevant since the most machine learning methods, such as support vector machines, random forests, and artificial neural networks assume balanced class distributions. Hence, these classifiers tend to overestimate patterns from the majority class, and underestimate those features characteristic of the minority class or classes. To overcome this limitation a class of ML approaches termed class imbalance learning (CIL) methods have been developed. CIL algorithms can be based on *data sampling* (e.g., random undersampling, bootstrap sampling, etc.); on *algorithm modifications* incorporating the inherent biases or skewness in the learning steps (e.g., weightedSVM, weigthedELM) or in *ensemble learning* in which several ML methods are applied and the results are consensed or averaged (52, 58).

Furthermore, even if most of these problematic issues may be solved, at least partially, with the analytic approaches just discussed, two relevant issues remain. First, real life datasets often have not one, but several (even all) of these challenging features. The ML methods useful to tackle some of these limitations may have poor performance due to others. Leveraging alternatives by evaluating the pros and cons may not be trivial. Second, every one of the methods for LSDA in imperfect/real-life datasets has its own set of assumptions and limitations. AI/ML researchers in biomedicine should be very aware of this and very cautious when combining methods and taking conclusions. However, as we will see in the next section, advancing biomedical and clinical research by using AI/ML approaches often worth all the efforts.

3. PRECISION MEDICINE: TRANSFORMING BIOMEDICAL EVIDENCE WITH DATA ANALYTICS

Since the later years of the 20th century, following the pioneering work by Cochrane, Eddy, and others (59–62) efforts have been directed toward building a systematic approach to medical and public health decisions, one founded not on anechdotic or individual expertise, but rather in the light of a full inspection of the existing clinical and biomedical research. This approach, called Evidence-Based Medicine (EBM) (63) aimed at the comprehensive use of all the accumulated scientific and clinical evidence to develop health related interventions and policy. At that time EBM was founded on anecdotal clinical experience, published case reports, meta-analyses and systematic reviews, and randomized controlled trials (64, 65). No *high-throughput* molecular or individual disaggregated information was considered at the time; even the already existing large-scale epidemiological data was not exploited fully due to data availability constraints (66, 67).

Even if the EBM paradigm has been superseded for various reasons, perhaps its main relevance resided in bringing to attention the fact that, as a rule, healthcare-related decisions should be supported by objective, stringent evidence rather than being left to the subjective opinion of some *individual* professional, expert as they may be. With the advent of larger, well-curated data corpora and more powerful ways to analyze the data and transforming it into useful information, EBM ideals have been embraced and incorporated into what has been called Precision Medicine (68–71).

Aside from the spectacular changes in information technologies in recent times, another main booster of this transformation was the *genomic revolution* driven by the human genome project (HGP) (72–74). The promises of the HGP,—many of them still undelivered (75)—pointed out to data-based biomedicine (particularly the identification of genetic variants behind the diseased phenotypes), as a key player to identify targets and customize pharmacological and other therapeutic interventions leading to a dramatic improvement of population and individual health (76, 77).

In view of this emerging paradigm, what is the role that AI/ML may play in its establishment as the standard approach in biomedical research, clinical practice and public policy? It has been argued (2, 6, 78) that there are at least four development avenues in which LSDA may impact healthcare: (i) LSDA may enlarge the capacity to generate new biomedical knowledge, (ii) LSDA may provide a support for healthcare-related knowledge dissemination, (iii) LSDA can become a tool for translating personalized medicine initiatives into clinical practice (for instance, by integrating molecular and EHR data on a single framework), and (iv) LSDA supplemented with simplified user interfaces can become a vehicle for empowering of the patients, helping them play a more active role in their own healthcare decision making.

In order to deliver such benefits, LSDA needs to be able to address questions, such as how to deal with highly unstructured heterogeneous data (say from EHRs) via high-performance computational techniques for quantitative analytics, but also for data mining, literature mining, and natural language processing algorithms over integrated pipelines. Particularly challenging are the scenarios related to clinical practice since they would be ideally processing such enormous amounts of unstructured data in cuasi-real time, if LSDA is intended to be beneficial for the

individual patient (79, 80). In the following sections, we will discuss some of the opportunities and limitations of applying AI/ML (often in the form of LSDA) in health-related settings.

3.1. Personalized Medicine: From Data Lakes to Patient Beds

LSDA and AI/ML may also play a role in supporting the clinical practitioners to keep up-to-date with the current scientific literature in their fields, an issue that has been struggling attending physicians for a while. In brief, if a medical doctor wants to treat their patients with the current best available therapeutic options, difficulties arise in trying to define what is *currently* considered better. As is known, the available scientific literature regarding a single medical speciality has been already overwhelming. The situation becomes much worse when one is dealing with multi-morbid patients since clinical guidelines and algorithms are often aimed at the single condition scenario (81–85).

Embracing the computational learning paradigm, the clinician may be armed with a new set of tools allowing for suggestions/surveys supported by real-time patient data analytics integrating, both the complexity of the patient's genetic background, environmental conditions, and the corresponding comorbidities with the current literature standards of care (**Figure 1** section (3)) (6, 33, 46, 86–88).

Aside from standard biomedical and clinical data, LSDA allows to further integrate occupational, social, physiological, and even behavioral information of the individual patient (available in social network, wearable devices, and other cloud-based resources) (89–92) to enhance the clinical profiles. To reach this point, however, there are important conundrums to be solved. In particular, novel computing and analytical frameworks should be designed to find patients' similarities and differences, but also to discover patterns highlighting their connections and discrepancies with the aim of calculating, for instance, personalized disease risk profiles, akin to polygenic risk scores, but under a much more general view—engulfing all the already discussed data types—allowing for individualized proactive medicine (93–95).

Hence, by integrating phenotype and disease-history based approaches, LSDA aims to advance personalized disease prediction, improve healthcare management and even contribute to an overall positive impact to individual wellness (**Figure 1** section (4)) (96–100). In doing so, AI/ML approaches are collaborating to a shift in the emphasis of clinical medicine from a disease-centered view to a patient-based practice (101, 102), a paradigm that has been long known since Hippocratic times and has been resumed a hundred years ago by the Spanish endocrinologist Gregorio Marañón who stated that *there are no diseases but patients*.

The panorama we have just discussed seem to be quite promising, indeed AI/ML and LSDA have already brought relevant advances toward Personalized Medicine (34, 70, 103). However, a consensus has not been reached as to how to integrate the large scale data of EHR, the many heterogeneous databases on molecular, phenotypical and environmental information derived from large scale experimental, clinical and epidemiologic studies and the individual-wise data gathered from disparate sources, such as social networks and wearable devices to develop a personalized approach to medicine? (46, 48, 104–106).

4. CHALLENGES TO COMPUTATIONAL LEARNING IN PRECISION MEDICINE

Of the many challenges posed to AI/ML by ever-growing health and biomedicine data sources, one of them is paradoxically related to what is often perceived as its main driving force. Having large amounts of data is obviously beneficial for computational learning algorithms, the more data you have, the more robust your classifiers, regressions, and mining strategies will be. However, as the tendencies move toward Precision Medicine, we can see how some major sources of primary biomedical information, such as genomics (in particular next generation sequencing) and imaging are becoming progressively cheaper (107–109), hence allowing their widespread use, nevertheless the computational costs of processing and analyzing the data are, for obvious reasons, growing fast (110–114).

Hence, aside from the already discussed challenges of database structural heterogeneity and data type integration, a number of major limitations for the development of AI/ML in biomedicine belong to the computer systems domain (115). Those challenges are, for instance, in the development of consolidation, characterization, validation, and processing standards for the data; creating ontologies and knowledge relationships for entities, such as genes, drugs, diseases, symptoms, patients, and treatments, as well as their corresponding entity-relationship schemes (116–119).

Along these lines, recent advances in AI, in particular those directed to Natural Language Processing (NLP) have been incorporating tools of semantic web analysis, such as conceptual relational networks (120, 121), semantic-syntactic classification (122), and similarity mapping (123). The problem, again, is a matter of throughput: effective implementation (training, in particular) of such NLP tools is only enabled if one has extremely large data corpora being accessed on a concurrent fashion (124). The vast majority of hospitals, research labs and even pharmaceutical development facilities do not currently have access to the storage and computational power resources needed to perform these analyses. The current alternative to local processing is, of course, cloud computing (125–127). However, as we will see in the next subsection, performing LSDA in medical and biomedical data in the cloud is not a problem-free solution.

4.1. Precision Medicine, Machine Learning and Cloud Computing

The use of cloud computing in the analysis of clinical, biomedical and healthcare data has many advantages: (i) it helps to solve the issue of processing large amounts of data in real time (128, 129), (ii) may provide scalable, cost-efficient data analytics solutions (130). Cloud computing, however, brings some technical difficulties, such as the ones related to high-throughput data transfer infrastructures, distributed computer

power over very large non-parallelizable tasks and perhaps the main challenge (that we will discuss more in depth in a forthcoming section) which lies in adapting the current distributed storage and processing paradigms in big data, while simultaneously allowing for full confidentiality of the data (since some of it may be highly sensible in nature) (131).

However, a number of cloud computing resources are becoming a standard for several omic studies, as it can be exemplified by `Basespace` a cloud-based sequencing analysis environment by Illumina, by the `EasyGenomics` platform of the Beijing Genomics Institute (BGI) and by European-based `Embassy` clouds as part of the Elixir Collaboration, by the NGScloud2 over Amazon Web Services (AWS) or by Galaxy-Kubernetes integrated workflows to name but a few instances (132–139).

It is worth noticing that standard cloud computing designs using distributed systems, grid computing, parallel programming, and virtualization on top of multi-layered environments (134, 140) are becoming adopted in LSDA in precision medicine due to their applications in the development of robust and secure distributed analysis (132). Indeed, as we already mentioned, cloud computing in LSDA may be implemented under several paradigms, such as: Platform as a Service (PAAS) (141–143), Infrastructure as a Service (IAAS) (144, 145), and Software as a Service (SAAS) (35, 146, 147).

These different standards for cloud computing have their particular pros and cons when applied to LSDA in Precision Medicine; for instance PAAS designs are suited for in-house software development or to integrate already designed libraries that can be implemented either by the user or by the cloud provider. Here we can mention healthcare, biomedicine, and bioinformatics services by providers, such as Google App Engine, Microsoft Azure MapReduce Hadoop, and others. In contrast, IAAS providers commonly offer high performance computing and massive storage facilities (sometimes called *HPC-farms* or *data centers*) including only the minimum operating system/computing environment requirements: this is often the case of general plans offered by companies, such as Amazon Web Services, HP Cloud, Rackspace, and Joyent (148–151).

Of these different paradigms, SAAS results as the more complete, as well as the more costly and less flexible. In SAAS the user is able to perform LSDA via pre-established (sometimes customized) applications sitting on a remote cloud infrastructure. This provides almost immediate access and usability with minimum installation and customization requirements from the user. However, due to these very reasons, the user has less control over the specifics of both, the computing environment and the actual algorithms used to perform analysis. The risk is that some of the more sophisticated methods will develop into *black boxes*. A somewhat intermediate solution is what can be called *Code-as-a-service* that is, SAAS with full access to the code (often only by specific requirement of the user). This is the case of the Cloud BioLinux service (152). The Cloud BioLinux suite has a set of pre-installed services, like a Galaxy server (153), access to the BioPerl programming language (154), BLAST (155), R/Bioconductor (156), Glimmer (157), ClustalW (158), and other general purpose (mostly bioinformatic-related) libraries/packages/environments (35, 159, 160).

Aside from molecular biology and genomics oriented applications, SAAS has also been developed in areas, such as medical diagnostics. In this regard, one can mention *DXplain*, one of the earliest developed decision support systems available. DXplain that was created by scientists, physicians, and software engineers at Massachusetts General Hospital http://www.mghlcs.org/projects/dxplain. DXplain may be used as a search engine (akin to a searchable eBook) providing the concise yet detailed description of more than 2,600 medical conditions, indexed by their main signs and symptoms, as well as their etiology, pathology, and prognosis. More relevant to this discussion is the use of DXplain as a case analytics tool, processing a set of clinical findings (signs, symptoms, laboratory data) as an input to a computational intelligence engine that computes a ranked list of diagnoses related to the given clinical manifestations. Furthermore, DXplain provides supports its suggestions with evidence sources, suggests what further clinical information would be useful to collect for the conditions under consideration, and displays a list of relevant clinical manifestations (161, 162). IBM's *Watson Health* constitutes another example of a (commercial) SAAS system aimed to support clinical decision making by the use of computational intelligence methods www.ibm.com/watson-health/ (163). However, many researchers and clinicians have become skeptical of the tool due to initial over-promises from the company (164). Many other diagnostic support applications have been developed, most of them aimed at commercial use such is the case of *ISABEL* https://www.isabelhealthcare.com/ (165, 166) and others. However, due to commercial restrictions, their AI/ML assessment and their use in LSDA has been rather restricted (167, 168).

In the end, each health/biomedical/clinical research team will have to make a choice between these different levels of cloud services depending on its availability of technical staff (computational biologists, data scientists, statisticians, bioinformaticians, software engineers, and so on), the computer literacy and involvement of the biomedical researchers and the clinicians, the scope and extension of the projects and other constraints, including financial issues, local infrastructure, and confidentiality matters (169–173).

It is also needed to take into account that some LSDA applications in health and biomedicine demand usually high computing resources. One alternative that is gaining relevance recently is the design of hybrid servers combining traditional CPUs with Graphical Processing Units (GPUs). The use of GPUs on cloud-based environments is indeed favored, given their massively parallel architechture (MPA). MPA results advantageous not only for actual computations, but also for input/output (I/O) operations (174). An important fraction of GPU-based applications in computational biology and biomedicine are implemented under (175–177). However, it remains a challenging endeavor to develop and implement parallelization algorithms, efficient enough to make sense of heterogeneous data sources, such as the ones coming from omic technologies, from EHRs, population surveys (127, 178).

Aside from the already mentioned cloud-based solutions, most research and clinical institutions will need to build some local infrastructure and algorithmics suited for their particular needs. In the search for semi-automation and reproducibility, some relevant general tasks are better managed by resorting to specialized software and algorithmic suites developed with building workflows and pipelines in mind. We will present some of the more widely used of such suites or packages for LSDA useful in Precision Medicine in the following subsection.

4.2. Software Resources for Computational Medicine

Whether implementing local, cloud-computing, or hybrid solutions, choices need to be made regarding appropriate algorithms and software for data pre-processing, processing, and analytics. A number of general purpose approaches have been developed, such is the case of the suite of R-based algorithms and programs in the `Bioconductor` repositories (156), the pipeline management tools, such as `Snakemake` (179, 180) and `Taverna` (181) or the cloud-based development suites `Helastic` (182) and `BioNimbus` (183).

For sequence analytics, a central player for quite some time has been the genome analysis toolkit (`GATK`) by the Broad Institute (184, 185). The `GATK` suite has been developed for LSDA of genome sequencing data mainly focused on high-accuracy variant discovery and genotyping useful in the clinical and biomedical research environments (186). Other computational omic analysis tools useful in the context of Precision Medicine include `dRanger` for the automatic identification of somatic rearrangements in Cancer (187), `Athlates` for the determination of HLA immuno-genotypes from exome sequencing data (188), the `Trinity` suite for De Novo RNA-Seq analysis (189), the `Hail` library for scalable (bio-bank scale) genomic data exploration (190), and the GWAS analysis suite `Plink` (191), to name but a handful instances.

More broadly applicable suites have been also developed, such as `GenePattern` (192, 193), the running/development platform `Galaxy` (153, 194, 195). Biological function databases like `Gene Ontology` (196, 197) and its generalizations (198, 199), the `MONA` (multi-level ontology analyses) programs (200), and other medium-to-high level analysis tools, such as the network analysis suite `Cytoscape` (201) or the structural biology libraries `BioDAS` (202) to mention but a handful of the many available options.

Aside from genomics and purely molecular/omic studies, other computational tools have been developed and widely used in the biomedical and clinical settings. Such is the case of `CellProfiler` for image analysis and processing (203) that has been proved to be quite useful for machine learning applications (204, 205). Automating data throughput in biomedical and clinical applications may also be useful even for relatively low demand tasks under certain circumstances; for example, automated RT-PCR data processing as implemented in `ARPA` (Automated RT-PCR analysis) turned out to be crucial for testing efforts during the COVID-19 pandemic (206). AI/ML modeling based on facilitated access data may indeed become a key tool to tackle with current and future pandemics (207).

Moving on to clinical applications, some of the most popular computational tools for managing clinical data (particularly with clinical trials in view) are `OpenClinica` (208), the Integrated Data Repository Toolkit `IDRT` (209) and the `VISTA` trials suite (210), and the comorbidity risk assessment tool `comoR` (211). Tools for the management of high-throughput day-to-day clinical records commercial and academic/open source have flourished in recent times. Some of the more widely adopted open source software solutions are `OpenEMR` (212), `OpenMRS` (213), `WorldVistA` (214). Some of these tools are actually enabling capacities to allow for the implementation of data mining and computational learning on their databases (54), however, as previously discussed, caution must be taken when using EHR data for automated discovery since a number of potential biases and confounders may arise (215, 216).

There are also some R-packages useful to manage EHR data. Such is the case of `EHR`: an Electronic Health Record and Data Processing and Analysis Tool https://cran.r-project.org/web/packages/EHR/index.html (217, 218), as well as `rEHR` https://github.com/rOpenHealth/rEHR (219).

Other software solutions from the R ecosystem useful in the LSDA applications in the clinical practice include `babsim.hospital`, a hospital resource planner and simulator https://cran.r-project.org/web/packages/babsim.hospital/index.html (220); `bp` a blood pressure analytics tool https://cran.r-project.org/web/packages/bp/index.html; and `card` a toolkit to evaluate the autonomic regulation of cardiovascular physiology via integrating electrocardiography, circadian rhythms, and the clinical risk of autonomic dysfunction on cardiovascular health data https://cran.r-project.org/web/packages/card/index.html (221).

Other software packages include `radtools` a set of utilities to extract and analyze medical image metadata https://cran.r-project.org/src/contrib/Archive/radtools/ (222); `psrwe` a library useful to incorporate real-world evidence (RWE) into regulatory and health care decision making https://cran.r-project.org/web/packages/psrwe/index.html (223, 224); `clinDataReview` https://cran.r-project.org/web/packages/clinDataReview/index.html an environment to support exploratory analysis of data in clinical trial settings, `patientProfilesVis` a tool to create patient profile visualizations for exploration, diagnostic or monitoring purposes during a clinical trial https://cran.r-project.org/web/packages/patientProfilesVis/index.html; and even `healthyR` a full suite to review common administrative hospital data. Although this latter application does not seem to be related to LSDA in Precision Medicine, it is not uncommon the application of AI/ML methods to administrative data to infer, for instance, social determinants of health.

5. DATA INTEGRATION: CURRENT CHALLENGES

Computational limitations in LSDA for Precision Medicine are gradually being overcome. Deeper challenges, however, arise

when we consider the question of how to develop coherent ways to *make sense of the data*, that is how to build models and analytical frameworks that allow biomedical scientists and clinicians to use all these currently available data types and resources in the best possible way as diagnostic and prognostic tools (225). In the context of genomics (and other omics) in biomedicine, important international efforts along these lines have been developed, such is the case of the ELIXIR-EXCELERATE collaboration (136), the STATegra project (226, 227), the SeqAhead consortium (228), and others (229, 230).

It must be stressed that most of the efforts of these—extremely relevant—endeavors are directed toward the integration of information on the *molecular* side of the spectrum of biomedical related data. Data integration at this level provides mathematical and relational models able to give a mechanistic description of the interplay between the molecular components of the cells (225, 231). This is of course fundamental to understand the rise of cellular and tissular phenotypes from its biochemical origins, but may result insufficient to account for the rise of disease in organs, individuals, and even populations. Recent advances have been done to extend these efforts to encompass LSDA on biological databases incorporating individual EHR data (232), as well as social and environmental information [the so-called social determinants of health (233)]; perhaps even incorporating constraints representing healthcare policy within a precision medicine framework (93, 234). Advances in AI will surely play a central role in the development of such integrated frameworks (235).

In this context, data integration allows the use of multiple data sources with several different (eve disparate) *pieces of evidence* to build (hopefully) interpretable models of the systems under study (236). Since these broad array of data sources may have quite different structures, levels of granularity and, in the case of quantitative measurements, different distributions and dynamic ranges, data integration is indeed a demanding endeavor, briefly subsumed in the question *how can we put together these data sources to improve knowledge discovery?* (237). Hence, being able to perform complex queries, build heterogeneous models and develop hierarchically nested data retrieval operations on multiple databases are core goals for data integration strategies useful for AI/ML models in Precision Medicine (235, 238–241).

LSDA in Precision Medicine is driven by two major sets of goals. On the one hand, we aim to develop *high level intuition* (HLE) from inductive analyses, via statistical learning and causal inference techniques. HLE may serve to sketch guidelines for current and future experimental and clinical research (242). On the other hand, AI/ML approaches may be useful for *automated reasoning* (AR), i.e., the non-supervised or semisupervised extraction of non-trivial patterns in *dynamic* databases (243–245).

5.1. The Need for Guidelines and Standardization to Support Precision Medicine

Machine learning and artificial intelligence approaches able to live up to these envisioned objectives will depend on the underlying data resources to a great extent. We will need, not only high throughput carefully curated databases, but also inter-operable data strategies. By creating integrated/integrable databases related to Precision Medicine we will enhance our *data discovery* and *data exploitation* capabilities, refine our algorithms for *statistical assessment of data-driven discovery* and improve our *data standardization*. Regarding data standards, there have been some advancements from the early days of the MIAME requirements (246, 247) for genomic data formats, now updated for next generation sequencing data (248) and even for single cell RNASeq experiments (249); to some more recent efforts for meta-data standardization (250, 251).

Focused efforts toward data standardization with AI/ML approaches in mind have been recently advanced. For instance, a multi-institutional group has recently compiled a document establishing guidelines on *Minimum information about clinical artificial intelligence modeling* by means of the MI-CLAIM checklist (252). MI-CLAIM has been developed as a tool to make reporting of AI/ML algorithms in medicine more transparent. This approach looks to solve issues related to interpretability, opaque documentation and scope of AI/ML methods in medicine. It consists of six parts: (i) Study design, (ii) Separation of data into partitions for model training and testing, (iii) Optimization and final model selection, (iv) Performance evaluation, (v) Model examination and (vi) Reproducible pipeline. Central to this standard is the MI-CLAIM checklist [Table 1 in (252)].

Aside from methods, standards need to be developed for all different aspects involved in biomedical data analytics and computational intelligence. From the patients/subjects to the clinical and analytical research, to academic and industrial approaches and back to the patients and clinicians. The National Patient-Centered Clinical Research Network (PCORNET) initiative https://pcornet.org/ of the US has been developed as *a national resource where health data, research expertise, and patient insights are available to deliver fast, trustworthy answers that advance health outcomes* (253). PCORNET was designed as a distributed data research network (DRN) built to facilitate multi-site observational and interventional research across the diverse (existent-at-the -time and future) clinical data research networks and other relevant players in the health data ecosystem.

By standardizing procedures, formats and approaches PCORNET looks up to deliver greater sample size and power of the studies, the ability to analyze the effects of the differences in practice and assessing heterogeneity in treatments and populations. It included the creation of a Data Standards Security and Network Infrastructure (DSSNI) task force aimed to identify the minimal data standards and technical specifications for data to be effectively shared and disseminated effectively. These actions will be directed to optimize the evaluation and improving quality assessment of the research projects and to maximize their concurrent impact (254). Other task forces within PCORNET are devoted to issues, such as Governance, Data privacy, Ethics, and regulation, Health system interactions, Patient and consumer engagement, Patient-generated outcomes, Clinical trials, Rare diseases, Biorepositories, and Obesity. These

task forces (and other that are being added as they develop) are supervised by PCORNET's Project Management Office operating under a network-like structure rather than as a traditional hierarchical organization. The development and functioning of the approach are subject to continuous assessment and evaluation via the Foundational Data Quality model founded on the premises of optimal data curation (255).

A related initiative put forward by the National Center for Biomedical Computing of the US is the I2B2 (Informatics for Integrating Biology and the Bedside) https://www.i2b2.org/index.html. I2B2 was developed with the aim of *enabling effective collaboration for precision medicine, through the sharing, integration, standardization, and analysis of heterogeneous data from healthcare and research; through engagement and mobilization of a life sciences-focused open-source, open-data community.*. I2B2 was created as part of the NIH roadmap to advance precision medicine to provide the community of clinical investigators with a toolbox to integrate medical records, clinical data, and genomic technologies all at once (256). One of the foundations of I2B2's approach to data interoperability is data-model harmonization based on ontological representations, particularly those facilitating the involvement of subjects/patients and clinicians aside from biomedical researchers (257). The extent of influence of these actions is designed to further improve the way subjects are enrolled and followed-up in research study protocols, clinical trials and observational cohorts (258).

Ontologies are useful to provide a conceptual framework. In the case of automated and semi-automated data mining methods in biomedicine it is desirable to have a *standardized language*, easily translated into machine-readable text. This is precisely the aim of the *Biological Expression Language* (BEL). BEL is presented as *a language for representing scientific findings in the life sciences in a computable form. BEL is designed to represent scientific findings by capturing causal and correlative relationships in context, where context can include information about the biological and experimental system in which the relationships were observed, the supporting publications cited and the process of curation* https://bel.bio/. The elementary elements of BEL are known as BEL-assertions that are built as intermediate steps connecting natural language (as presented in say, academic writing or medical records) into machine-readable expressions. Such expression will then be *computable* with applications in tasks, such as logical modeling in database learning, systems biology verification studies or next generation EBM to name a few (259–261). Implementing language standards, such as BEL may prove beneficial, since it has been shown, for instance, that different approaches to process clinical notes using natural language analytics substantially affects the performance of predictive models in intensive care settings (262).

The biomedical data ecosystem is turning so complex that new standards are needed even to define what we call *evidence*. The large amounts of seemingly anecdotal data that are being produced nowadays have brought to attention issues like the so-called *real world evidence* (RWE). RWE refers to *data regarding the use, or the potential benefits or risks, of a drug derived from sources other than randomized clinical trials* (263). Large sampling spaces are behind RWE move from anecdotal to referential. However, not all the real world information should be treated as RWE. In this regard, there is a growing need for methods to assess when are these data sources rigorous and trustworthy enough as to be useful as a guideline or to be considered actual *evidence*. These issues result particularly relevant toward the definition of clinical pipelines in *digital therapeutics* (loosely defined as evidence based therapeutics basedon software applications to prevent, manage or treat a disease or medical condition) (264), often related with data obtained from wearables and other subject-based sources.

Data standardization is becoming central not only in the medical research, and personalized clinical practice settings. It has been recently discussed how clinical trial data sharing is essential for reproducibility of the findings, for visibility of the results, to improve subsequent trails or advanced clinical trial stages, to perform digital comparisons of effectiveness (which are much faster and cheaper than their traditional counterparts); but also to speed results reporting, to enable continuous learning and even to support the emergence of startups or enterprise ventures, among other issues (265). In order for shared data to be optimally usable, there is an obvious need for standardization.

Data is, of course, not the only issue that needs to be assessed and validated toward the widespread implementation of AI/ML approaches in Precision Medicine. Eaneff and coworkers have recently argued for the need of *algorithmic stewardship* for AI/ML technologies in the medical setting. In this regard, an algorithmic steward would be a person or group within a healthcare or biomedical research institution responsible for tasks, such as creating and maintaining an algorithmic inventory of the methods used in the institution, monitoring ongoing clinical use and performance of such computational tools, evaluating the safety efficacy and fairness of the methods and so on (266).

Data and methods constitute the most visible items within the biomedical analytics ecosystem; metadata, is however, progressively gaining a more relevant role for AI/ML in Precision Medicine, as it contains, in many cases, hints for the automated *labeling* or classification (even if approximate) tasks that will be further improved by the use of computational intelligence and statistical learning approaches (87, 267). We will further discuss this issue in the next subsection.

5.2. An Ocean of Metadata

Metadata has become a central player in contemporary LSDA endeavors in many fields, including biomedicine; particularly relevant for AI/ML approaches. For this reason, aiming for high quality, well-formatted and standardized metadata has become quite relevant (268). Indeed, a number of biomedical data analysis teams and consortia are encouraging the use of standardized metadata guidelines, exemplified, for instance by a *checklist* of relevant issues to consider when building and publishing companion metadata (250, 269, 270); since such metadata could be instrumental to implement data analytics, as well as AI/ML toward a precision medicine approach (267, 271).

Metadata may result also quite useful to enhance the statistical analysis, probabilistic models and training of learning machines. Using metadata to generate best priors may improve the outcomes of query optimization by resampling and

bootstrapping (272–274), regularization of sparse datasets (275), as well as auxiliary source for multi-variate Bayesian analysis (200, 276, 277), multi-dimensional analyses on datasets with disparate dynamic ranges (278–281) among other instances (282–286).

Integrating multiple data and metadata sources takes even further the need to design, develop, and implement analysis algorithms able to handle heterogeneous data in the presence of noise accumulation, spurious correlations and incidental endogeneity, keeping a balance between statistical accuracy, computational efficiency, and interpretability (287–289).

LSDA approaches must be developed having in mind the presence of spurious correlations among unrelated covariates, challenging statistical inference by creating false positive findings (290). Incidental endogeneity occurs when a number of unrelated covariates become *correlated* via random correlations of their residual noises. A statistical approach to overcome some of these issues is the development of novel regularization methodologies (291–293) but also the use of outside cross-validation via independence screening tests (294, 295) that may be precluded by data unavailability from independent sources.

Taking these issues into account may require new models to implement metadata reporting standards (296, 297). Standardizing the way metadata is reported and retrieved in the biomedical and clinical settings will result critical for the development of generalistic machine learning approaches that make full use of these uniform data structures (298–300). It has been recently discussed that ignoring or bypassing such standards may jeopardize full research projects (301–303).

6. ETHICAL AND LEGAL CHALLENGES FOR COMPUTATIONAL MEDICINE

Aside from the methodologic and logistic issues already discussed, integrating data sources aiming at LSDA in the context of Precison Medicine also brings out concerns related to the ethical and legal problems that may arise, for instance related to privacy and confidentiality. Regarding the purely technological aspects of this problem, most of the members of the community of data analyst in healthcare and biomedicine are actually confident that these can be solved with security and encryption approaches already used to protect personal financial data (6, 46, 304). Aside from privacy concerns, managing sensitive data implies having several layers of access to the data. This is so since *some* sensitive personal data may be extremely useful for population level studies needed to develop personalized medicine. However, even if it is unlikely that full disclosure of sensitive biomedical and clinical information is needed, there is a fraction—that need to be determined and agreed-upon in advance—of potentially sensitive information that results fundamental for the development of personalized medicine, not just for the individual in particular but also population and sub-population-wise (305).

Then a conundrum arises as how to accommodate smooth clinical and biomedical data widespread with efficient privacy practices. The goal here is to implement stringent rules that maximize data yield while preserving anonymity and data protection. Data specialists have proposed several strategies to accomplish this goal. Currently one of the most favored is centered in *mining designs* based on the so-called *minimally-invasive queries* (MIQs) designed ex-profeso to preclude (and in due case disclose/document) any abuse of sensitive data (306). In some sense MIQ approaches mimic and extend the practices that have been long held by the international health insurance community while dealing with privacy in the EHRs via guidelines, such as the *Health Insurance Portability and Accountability Act* (HIPPA). Aside from its enormous legal and bioethical consequences, HIPPA adoption induced the development of data protocols in biomedical informatics that will result useful—even if as a starting point—for the LSDA under the Precision Medicine paradigm. Full implementation of optimized data usage/protection protocols is still underway, however, important advances have been made (307–310).

Reaching an optimal balance between information protection and efficient data mining outputs presents itself as a complex endeavor: some experts from the biomedical ethics community advocate for a careful case-by-case analysis, though admittedly this will be too complex to be implemented in general purpose LSDA workflows. As an alternative to this it has been suggested that multi-level data encryption (311, 312) can be applied in such a way that only authorized personnel will have the decoding keys to have access of the different levels of information (313).

In order to lessen the burden of encryption, encryption must be selective so that only personal identifiers and other private features (that may help disclose such identifiers) should be encrypted. Quasi-identifiers (QIDs), such as location, ethnic profiling, age and employment information, and highly-specific genomic data may be subject to certain low-level encryption by following *differential privacy* standards (314, 315). Some caution needs still to be taken since individual QIDs may not be informative enough to disclose identity, but there may be mining-integration procedures that may be able to do so by arranging coupled queries as it has been already discussed in the context of large scale genomic and transcriptomic studies (316–318).

Aside from genomic sources, other data types that may be used as potential QIDs in the context of biomedical informatics include, for instance, photographs: it has been discussed that from image (and imaging) data, AI approaches are able to infer *barcodes* from cranial and facial morphological features, skin pigmentation, eye color, retina patterns, iris structure, as well as hair type and color (108, 317, 319–322).

These are but a handful examples of how biomedical and clinical data features may turn into QIDs potentially posing ethical dilemmas to LSDA in the context of Precision Medicine. In this context and with the advent of powerful AI/ML approaches, a question arises as to which queries are *valid* and which ones are not from the standpoint of ethics, privacy and confidentiality. It is expected that as AI/ML methods become more powerful, methodological adjustments should evolve to balance safety and non-triviality of the queries with the impact of the analyses. This call for an organized implementation of such features via standardized *query tools* compliant with the agreed (potentially also evolving) ethical standards of the

community (313). This translates into further challenges for the computational tools for data mining and analysis that may be designed with hierarchical multi-layered data structures in mind from the start.

Protected health information (PHI) is a relevant issue in this regard since it potentially allow for individual identification. Developing methods to effectively de-identify sensible data, such as the one included in free-text clinical notes may become part of the solution to the ethical challenges of high throughput data mining in the clinical and biomedical settings. With this in mind, Norgeot and collaborators developed a customizable open source de-identification software called `Philter` (323). Philter https://github.com/BCHSI/philter-ucsf has shown to outperform well-known methods, such as the ones in the Physionet https://www.PhysioNet.org/physiotools/deid/ and Scrubber https://scrubber.nlm.nih.gov/files/ suites. Subject de-identification in clinical notes and similar documents since such corpora often contain detailed information about the state of individual patients, the evolution of their disease conditions, specific theraputics and outcomes. That kind of information that will result key for the development of Precision Medicine, but at the same time may pose privacy challenges unless effective de-identified.

In view of the advances in AI/ML and the ethical challenges that come as a consequence of these advances, design changes are needed not only in the analytics. Research protocols, clinical trials and documented medical procedures, for instance, must be revised since the personal decision to share or not personal healthcare information or participating in large scale biomedical research cohorts may change at the light of AI/ML advances. Hence, informed consent procedures may need to be adapted. This implies reframing the current paradigm for the protection of individual privacy and adopting ways to educate patients/participants on how the data collected may affect them and the what extent their data can or cannot be protected, contextualising this in terms of the potential benefits for them and for others (317).

It has been discussed that *re-educating* about the way they view their own data also implies increasing their involvement with how their data may be used to affect them and others. Indeed, one of the central tenets of Personalized Medicine is making healthcare, *personal*. In this regard, it is worth discussing the role that *data portability* will play in individual and collective decisions (324, 325). Integrating data analytics, privacy protection and data portability is, in brief, one of the current open problems in computational medicine and medical informatics (326–328).

Given all the twists and subtleties just discussed in the context of LSDA for Precision Medicine, it has been considered advantageous to document in all detail (or as comprehensively as possible given the particular context) how data is gathered, archived, processed, analyzed, disseminated, and used in each research study, clinical trial, or large-scale clinical follow-up. Guidelines have been currently advised as how to elaborate such a document termed a *data management plan* (DMP). We will briefly discuss on these matters in the next section.

7. THE IMPORTANCE OF A GOOD DATA MANAGEMENT PLAN

In view of all the complexities associated with projects managing and analyzing large amounts of potentially sensitive data, writing down a comprehensive document with all the associated information, a data management plan document is considered advantageous (329–332). The purpose of the DMP is to establish guidelines about how the data will be treated during the course of the project and even what will happen after the project is finished. The DMP considers what will be done with the data from its collection, throughout the organization, pre-processing, and analysis stages. It considers data quality controls, database preservation, and documentation techniques used, as well as usage restrictions and conditions for the further use, dissemination and sharing, embargoes, and limitations.

The DMP document has been established to be compliant with the legal requirements for all involved institutions and funding agencies. It should specify what types of data are to be collected, the recommended (sometimes preferred, sometimes mandatory) formats to handle and preserve the data. It results relevant to mention the software requirements and computational resources used to store, process, analyze, and visualize the data. The expected volume and structure of the databases, as well as its sources, traceability and metadata information (329). The DMP must also mention the intended data preservation strategies, database organization (e.g., naming conventions, dictionaries, reports' systems, etc.), identification and de-identification procedures. It is also advisable to establish guidelines for database curators—in some cases, even for auditors—(for instance regarding data integrity, quality controls and data standardization). All these entries of the DMP must be compliant with normative and organizational principles detailed in the so-called *Project Data Policy* (PDP) section of the DMP. The PDP may include information on legal, administrative and even ethical restrictions to be considered when managing the data. In some cases, this has to make it extensive to associated software and metadata (331).

The data dissemination policy section of the DMP states how, when and whom will have access to the data and under what circumstances. It is recommended that a subsection assigning personal roles and responsibilities of the associated personnel is included to ensure good data governance. The DMP is, in brief a dynamic instrument that plays a normative role, but also serves as a registered account on the whole data workflows and procedures throughout the project. Hence, a good DMP contributes to a secure and smooth functioning of the whole LSDA project (333, 334).

8. CONCLUSIONS AND PERSPECTIVES

Artificial Intelligence and Machine Learning (AI/ML) approaches have proven to be extremely relevant tools for the large scale analysis of biomedical and clinical data; central for the development of Personalized Medicine. Useful as they are, implementing AI/ML methods in the highly demanding

medical applications, it is not an easy endeavor. A number of caveats, shortcomings and subtle points have to be taken into account (and in many cases, circumvented) in order to provide appropriate solutions for the individual and public health care to fully benefit from these emerging paradigms.

In this work, we have discussed about some of the central challenges, problems, and drawbacks found in the applications of the methods and designs of large scale data analytics within clinical and biomedical environments, in particular under a Precision Medicine perspective.

Some relevant points can be briefly summarized as follows:

- Precision Medicine has been recently presented as an emergent paradigm to approach healthcare in a more predictive, preventative, personalized, participatory way (sometimes also called P4 Medicine). Precision Medicine has strong ties with data intensive approaches, as well as with machine learning and artificial intelligence.
- To deliver the promise of Precision Medicine, computational learning approaches are to be nurtured by well-curated and nifty integrated data ecosystems.
- Data resources in the biomedical research, clinical and healthcare environments are becoming extremely large, and are complex, unstructured and heterogeneous, hence difficult to deal with individually, even more so to be integrated into a coherent framework.
- The universe of diverse data sources needs to be collected, pre-processed, processed, modeled, and integrated to construct such coherent frameworks useful for Precision Medicine (see **Figure 1**). This is much easier said than done.
- In order for machine learning models to give good results their input needs to be *good* data. Transforming existing data into optimized forms for AI/ML is essential.
- If medicine is to become *personalized*, we must embrace diversity, heterogeneity, biases, class imbalance, and other intrinsic features of *individuals*. There is a need to develop methodologies to rigorously operate under these constraints.
- To develop, implement, optimize, and improve on these methods, a number of challenges needs to be overcome. These include technical limitations, computational aspects (both software and hardware/infrastructure), mathematical and modeling issues, and even ethical, legal, and policy matters.
- We have presented and discussed some of these challenges, aiming at showing the state of the art in these different fields.
- We have introduced the need for data intensive endeavors, from the research arena to the clinical setting and the healthcare institution level to design and implement a data management plan to consider the issues that may arise and planning ahead for their solution.

We are convinced that the development and implementation of tailor-made (or at least well-customized) approaches, in terms of robust statistical and computational algorithms, supported by optimized frameworks for data acquisition, storage, management, and analytics, but also by well-integrated software solutions and guided by solid ethical policies compliant with a deep respect for privacy, confidentiality, and individuality; is an ambitious but attainable goal. Hence, by combining state of the art computational learning methods and techniques with the best data acquisition and management practices the promise of AI/ML in Personalized Medicine may be delivered.

AUTHOR CONTRIBUTIONS

EH-L conceived the project. EH-L and MM-G performed research and drafted the manuscript. All authors reviewed and approved the manuscript.

FUNDING

This work was supported by CONACYT (grant no. 285544/2016 Ciencia Bsica, and grant no. 2115 Fronteras de la Ciencia), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (grant no. 232647/2014 CONACYT). EH-L acknowledges additional support from the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

REFERENCES

1. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. *BMC Med.* (2018) 16:1–15. doi: 10.1186/s12916-018-1122-7
2. Cirillo D, Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol.* (2019) 58:161–7. doi: 10.1016/j.copbio.2019.03.004
3. Suwinski P, Ong C, Ling MH, Poh YM, Khan AM, Ong HS. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Front Gen.* (2019) 10:49. doi: 10.3389/fgene.2019.00049
4. Shortreed SM, Cook AJ, Coley RY, Bobb JF, Nelson JC. Challenges and opportunities for using big health care data to advance medical science and public health. *Am J Epidemiol.* (2019) 188:851–61. doi: 10.1093/aje/kwy292
5. Fairchild G, Tasseff B, Khalsa H, Generous N, Daughton AR, Velappan N, et al. Epidemiological data challenges: planning for a more robust future through data standards. *Front Publ Health.* (2018) 6:336. doi: 10.3389/fpubh.2018.00336
6. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA.* (2013) 309:1351–2. doi: 10.1001/jama.2013.393
7. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international HapMap project. *Nature.* (2003) 426:789–96. doi: 10.1038/nature02168
8. Thorisson GA, Smith AV, Krishnan L, Stein LD. The international HapMap project web site. *Gen Res.* (2005) 15:1592–3. doi: 10.1101/gr.4413105
9. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. *Nat Gen.* (2013) 45:1113–20. doi: 10.1038/ng.2764
10. Cline MS, Craft B, Swatloski T, Goldman M, Ma S, Haussler D, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci Rep.* (2013) 3:2652. doi: 10.1038/srep02652
11. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol.* (2015) 19:A68. doi: 10.5114/wo.2014.47136
12. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to

drive high-quality survival outcome analytics. *Cell.* (2018) 173:400–16. doi: 10.1016/j.cell.2018.02.052
13. Consortium GP, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* (2012) 491:56–65. doi: 10.1038/nature11632
14. Siva N. 1000 Genomes project. *Nat Biotechnol.* (2008) 26:256–7. doi: 10.1038/nbt0308-256b
15. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. *Nat Methods.* (2012) 9:459–62. doi: 10.1038/nmeth.1974
16. Via M, Gignoux C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med.* (2010) 2:1–3. doi: 10.1186/gm124
17. Consortium G, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* (2020) 369:1318–30. doi: 10.1126/science.aaz1776
18. Stranger BE, Brigham LE, Hasz R, Hunter M, Johns C, Johnson M, et al. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project. *Nat Gen.* (2017) 49:1664. doi: 10.1038/ng.3969
19. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. *Nat Gen.* (2013) 45:580–585. doi: 10.1038/ng.2653
20. Aviv R, Teichmann SA, Lander ES, Ido A, Christophe B, Ewan B, et al. The human cell atlas. *Elife.* (2017) 6:e27041.
21. Hon CC, Shin JW, Carninci P, Stubbington MJ. The human cell atlas: technical approaches and challenges. *Briefings Funct. Gen.* (2018) 17:283–94. doi: 10.1093/bfgp/elx029
22. Dawber TR, Meadors GF, Moore Jr FE. Epidemiological Approaches to Heart Disease: The Framingham Study*. *Amer J Publ Health Nat Health.* (1951) 41:279–86.
23. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet.* (2014) 383:999–1008. doi: 10.1016/S0140-6736(13)61752-3
24. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* (2007) 447:661–78. doi: 10.1038/nature05911
25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z
26. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* (2015) 12:e1001779. doi: 10.1371/journal.pmed.1001779
27. Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol.* (2012) 1:123–6. doi: 10.1016/j.hlpt.2012.07.003
28. Palmer LJ. UK Biobank: bank on it. *Lancet.* (2007) 369:1980–2. doi: 10.1016/S0140-6736(07)60924-6
29. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Gen.* (2018) 50:1593–9. doi: 10.1038/s41588-018-0248-z
30. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci.* (2016) 19:1523–36. doi: 10.1038/nn.4393
31. Fawns-Ritchie C, Deary IJ. Reliability and validity of the UK Biobank cognitive tests. *PLoS One.* (2020) 15:e0231627. doi: 10.1371/journal.pone.0231627
32. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Amer J Epidemiol.* (2017) 186:1026–34. doi: 10.1093/aje/kwx246
33. Hamburg MA, Collins FS. The path to personalized medicine. *New Engl J Med.* (2010) 363:301–4. doi: 10.1056/NEJMp1006304
34. Collins FS, Varmus H. A new initiative on precision medicine. *New Engl J Med.* (2015) 372:793–5. doi: 10.1056/NEJMp1500523
35. O'Driscoll A, Daugelaite J, Sleator RD. Big data, Hadoop and cloud computing in genomics. *J Biomed Informat.* (2013) 46:774–81. doi: 10.1016/j.jbi.2013.07.001
36. van Dijk WB, Fiolet AT, Schuit E, Sammani A, Groenhof TKJ, van der Graaf R, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study. *J Clin Epidemiol.* (2021) 132:97–105. doi: 10.1016/j.jclinepi.2020.11.014
37. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs) a survey. *ACM Comput Surveys (CSUR).* (2018) 50:1–40. doi: 10.1145/3127881
38. Ferrão JC, Oliveira MD, Janela F, Martins HM, Gartner D. Can structured EHR data support clinical coding? a data mining approach. *Health Syst.* (2021) 10:138–61. doi: 10.1080/20476965.2020.1729666
39. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Gen.* (2012) 13:395–405. doi: 10.1038/nrg3208
40. Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34, New York, NY (2020). p. 606–13.
41. Razzaque A, Hamdan A. Artificial intelligence based multinational corporate model for EHR interoperability on an e-health platform. In: *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications.* Cham: Springer (2021). p. 71–81.
42. Wu S, Liu S, Wang Y, Timmons T, Uppili H, Bedrick S, et al. Intrainstitutional EHR collections for patient-level information retrieval. *J Assoc Inf Sci Technol.* (2017) 68:2636–48. doi: 10.1002/asi.23884
43. Stevens LA, DiAngi YT, Schremp JD, Martorana MJ, Miller RE, Lee TC, et al. Designing an individualized EHR learning plan for providers. *Appl Clin Inf.* (2017) 8:924–35. doi: 10.4338/ACI-2017-04-0054
44. Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J. EHR-independent predictive decision support architecture based on OMOP. *Appl Clin Inf.* (2020) 11:399–404. doi: 10.1055/s-0040-1710393
45. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. *Cell.* (2019) 177:58–69. doi: 10.1016/j.cell.2019.02.039
46. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Int Med.* (2013) 28:660–5. doi: 10.1007/s11606-013-2455-8
47. Emmert-Streib F, Dehmer M. A machine learning perspective on personalized Medicine: an automized, comprehensive knowledge base with ontology for pattern recognition. *Mach Learn Knowl Extract.* (2019) 1:149–56. doi: 10.3390/make1010009
48. Schork NJ. Artificial intelligence and personalized medicine. In: *Precision Medicine in Cancer Therapy.* Cham: Springer (2019). p. 265–83.
49. Papadakis GZ, Karantanas AH, Tsiknakis M, Tsatsakis A, Spandidos DA, Marias K. Deep learning opens new horizons in personalized medicine. *Biomed Rep.* (2019) 10:215–7. doi: 10.3892/br.2019.1199
50. Rodriguez F, Scheinker D, Harrington RA. Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. *Circ Res.* (2018) 123:1282–4. doi: 10.1161/CIRCRESAHA.118.314119
51. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. *Cell.* (2020) 181:92–101. doi: 10.1016/j.cell.2020.03.022
52. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. *Genes.* (2019) 10:87. doi: 10.3390/genes10020087
53. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods.* (2016) 111:21–31. doi: 10.1016/j.ymeth.2016.08.014
54. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* (2018) 1:1–10. doi: 10.1038/s41746-018-0029-1
55. Spicker JS, Brunak S, Frederiksen KS, Toft H. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. *Toxicol Sci.* (2008) 102:444–54. doi: 10.1093/toxsci/kfn001
56. Gönen M, Alpaydın E. Multiple kernel learning algorithms. *J Mach Learn Res.* (2011) 12:2211–68.

57. Hasanin T, Khoshgoftaar TM, Leevy JL, Bauder RA. Investigating class rarity in big data. *J Big Data.* (2020) 7:1–17. doi: 10.1186/s40537-020-00301-0
58. Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. *Mol Oncol.* (2021) 15:817–29. doi: 10.1002/1878-0261.12920
59. Eddy DM, Billings J. The quality of medical evidence and medical practice: March 1987. *Am J Ophthalmol.* (2021) 225:189–205. doi: 10.1016/j.ajo.2020.08.034
60. Faria L, Oliveira-Lima JAd, Almeida-Filho N. Evidence-based medicine: a brief historical analysis of conceptual landmarks and practical goals for care. *História Ciências Saúde-Manguinhos.* (2021) 28:59–78. doi: 10.1590/s0104-59702021000100004
61. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Database Syst Rev.* (2019) 10:ED000142. doi: 10.1002/14651858.ED000142
62. Croskerry P. Medical decision making. In: *The Routledge International Handbook of Thinking and Reasoning.* London, UK: Routledge (2017). p. 109–29.
63. Group EBMW, et al. Evidence-based medicine. a new approach to teaching the practice of medicine. *JAMA.* (1992) 268:2420. doi: 10.1001/jama.268.17.2420
64. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet.* (2017) 390:415–23. doi: 10.1016/S0140-6736(16)31592-6
65. Oliver K, Pearce W. Three lessons from evidence-based medicine and policy: increase transparency, balance inputs and understand power. *Palgrave Commun.* (2017) 3:1–7. doi: 10.1057/s41599-017-0045-9
66. Cairney P, Oliver K. Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? *Health Res Policy Syst.* (2017) 15:1–11. doi: 10.1186/s12961-017-0192-x
67. Ioannidis JP. Hijacked evidence-based medicine: stay the course and throw the pirates overboard. *J Clin Epidemiol.* (2017) 84:11–3. doi: 10.1016/j.jclinepi.2017.02.001
68. De Maria Marchiano R, Di Sante G, Piro G, Carbone C, Tortora G, Boldrini L, et al. Translational research in the era of precision medicine: where we are and where we will go. *J Pers Med.* (2021) 11:216. doi: 10.3390/jpm11030216
69. Chow N, Gallo L, Busse JW. Evidence-based medicine and precision medicine: complementary approaches to clinical decision-making. *Precis Clin Med.* (2018) 1:60–4. doi: 10.1093/pcmedi/pby009
70. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnol.* (2012) 29:613–24. doi: 10.1016/j.nbt.2012.03.004
71. Abrahams E, Silver M. The case for personalized medicine. *J Diabetes Sci Technol.* (2009) 3:680–4. doi: 10.1177/193229680900300411
72. Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. *Gene Therapy.* (2017) 24:551–61. doi: 10.1038/gt.2017.77
73. Ginsburg GS, Phillips KA. Precision medicine: from science to value. *Health Affairs.* (2018) 37:694–701. doi: 10.1377/hlthaff.2017.1624
74. Katsios C, Roukos DH. Individual genomes and personalized medicine: life diversity and complexity. *Pers Med.* (2010) 7:347–50. doi: 10.2217/pme.10.30
75. Joyner MJ, Paneth N, et al. Promises, promises, and precision medicine. *J Clin Investigat.* (2019) 129:946–8. doi: 10.1172/JCI126119
76. Weinshilboum RM, Wang L. Pharmacogenomics: precision medicine and drug response. *Mayo Clin Proc.* (2017) 92: 1711–22. doi: 10.1016/j.mayocp.2017.09.001
77. Sandhu C, Qureshi A, Emili A. Panomics for precision medicine. *Trends Mol Med.* (2018) 24:85–101. doi: 10.1016/j.molmed.2017.11.001
78. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. *Int J Med Inf.* (2018) 114:57–65. doi: 10.1016/j.ijmedinf.2018.03.013
79. Kaur J, Mann KS. AI based healthcare platform for real time, predictive and prescriptive analytics using reactive programming. *J Phys Conf Series.* (2017) 933:012010. doi: 10.1088/1742-6596/933/1/012010
80. Kamble SS, Gunasekaran A, Goswami M, Manda J. A systematic perspective on the applications of big data analytics in healthcare management. *Int J Healthcare Manag.* (2018) 2:226–40. doi: 10.1080/20479700.2018.1531606
81. Majnarić LT, Babič F, OSullivan S, Holzinger A. AI and big data in healthcare: towards a more comprehensive research framework for multimorbidity. *J Clin Med.* (2021) 10:766. doi: 10.3390/jcm10040766
82. Cesario A, DOria M, Calvani R, Picca A, Pietragalla A, Lorusso D, et al. The Role of Artificial Intelligence in Managing Multimorbidity and Cancer. *J Personal Med.* (2021) 11:314. doi: 10.3390/jpm11040314
83. Hassaine A, Salimi-Khorshidi G, Canoy D, Rahimi K. Untangling the complexity of multimorbidity with machine learning. *Mech Ageing Develop.* (2020) 190:111325. doi: 10.1016/j.mad.2020.111325
84. Onder G, Bernabei R, Vetrano DL, Palmer K, Marengoni A. Facing multimorbidity in the precision medicine era. *Mech Ageing Develop.* (2020) 190:111287. doi: 10.1016/j.mad.2020.111287
85. Singh SP, Karkare S, Baswan SM, Singh VP. Unsupervised machine learning for co/multimorbidity analysis. *Int J Stat Probab.* (2018) 7:23. doi: 10.5539/ijsp.v7n6p23
86. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Machine Learning for Healthcare Conference.* Ann Arbor, MI: PMLR (2019). p. 359–80.
87. Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inf Decis Making.* (2017) 17:1–13. doi: 10.1186/s12911-017-0556-8
88. Alber M, Tepole AB, Cannon WR, De S, Dura-Bernal S, Garikipati K, et al. Integrating machine learning and multiscale modeling perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digit Med.* (2019) 2:1–11. doi: 10.1038/s41746-019-0193-y
89. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst.* (2018) 6:1–12. doi: 10.1007/s13755-018-0046-0
90. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. *J Biomed Inf.* (2020) 108:103500. doi: 10.1016/j.jbi.2020.103500
91. Witt DR, Kellogg RA, Snyder MP, Dunn J. Windows into human health through wearables data analytics. *Curr Opin Biomed Eng.* (2019) 9:28–46. doi: 10.1016/j.cobme.2019.01.001
92. Nair LR, Shetty SD, Shetty SD. Applying spark based machine learning model on streaming big data for health status prediction. *Comput Elect Eng.* (2018) 65:393–9. doi: 10.1016/j.compeleceng.2017.03.009
93. Denny JC, Collins FS. Precision medicine in 2030seven ways to transform healthcare. *Cell.* (2021) 184:1415–9. doi: 10.1016/j.cell.2021.01.015
94. Weintraub WS, Fahed AC, Rumsfeld JS. Translational medicine in the era of big data and machine learning. *Circul Res.* (2018) 123:1202–4. doi: 10.1161/CIRCRESAHA.118.313944
95. Sevakula RK, Au-Yeung WTM, Singh JP, Heist EK, Isselbacher EM, Armoundas AA. State-of-the-Art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. *J Am Heart Assoc.* (2020) 9:e013924. doi: 10.1161/JAHA.119.013924
96. Bland J, Minich D, Eck B. A systems medicine approach: translating emerging science into individualized wellness. *Adv Med.* (2017) 2017:1718957. doi: 10.1155/2017/1718957
97. Hood L, Lovejoy JC, Price ND. Integrating big data and actionable health coaching to optimize wellness. *BMC Med.* (2015) 13:1–4. doi: 10.1186/s12916-014-0238-7
98. Dolley S. Big data's role in precision public health. *Front Publ Health.* (2018) 6:68. doi: 10.3389/fpubh.2018.00068
99. Imran S, Mahmood T, Morshed A, Sellis T. Big data analytics in healthcare-a systematic literature review and roadmap for practical implementation. *IEEE/CAA J Autom Sinica.* (2020) 8:1–22. doi: 10.1109/JAS.2020.1003384
100. Wang F, Casalino LP, Khullar D. Deep learning in medicine promise, progress, and challenges. *JAMA Int Med.* (2019) 179:293–4. doi: 10.1001/jamainternmed.2018.7117

101. Mifsud J, Gavrilovici C. Big data in healthcare and the life sciences. In: *Ethics and Integrity in Health and Life Sciences Research.* Warrington, UK: Emerald Publishing Limited (2018).
102. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* New York, NY: Hachette (2019).
103. Mathur S, Sutton J. Personalized medicine could transform healthcare. *Biomed Rep.* (2017) 7:3–5. doi: 10.3892/br.2017.922
104. Tyler J, Choi SW, Tewari M. Real-time, personalized medicine through wearable sensors and dynamic predictive modeling: a new paradigm for clinical medicine. *Curr Opin Syst Biol.* (2020) 20:17–25. doi: 10.1016/j.coisb.2020.07.001
105. Blasiak A, Khong J, Kee T. CURATE. AI: optimizing personalized medicine with artificial intelligence. *SLAS Technol Transl Life Sci Innov.* (2020) 25:95–105. doi: 10.1177/2472630319890316
106. De Georgia M, Loparo K. *Neurocritical Care Informatics: Translating Raw Data Into Bedside Action.* Berlin: Springer (2019).
107. Committee on A Framework for Developing a New Taxonomy of Disease NRCUC, et al. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. *Nat Acad Press (US)* (2011) 21–39. doi: 10.17226/13284
108. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell.* (2012) 148:1293–307. doi: 10.1016/j.cell.2012.02.009
109. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Human Mol Gen.* (2010) 19:ddq416. doi: 10.1093/hmg/ddq416
110. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. *J Med Internet Res.* (2019) 21:e13043. doi: 10.2196/13043
111. Becker M, Schultze H, Bresniker K, Singhal S, Ulas T, Schultze JL. A novel computational architecture for large-scale genomics. *Nat Biotechnol.* (2020) 38:1239–41. doi: 10.1038/s41587-020-0699-5
112. Kocheturov A, Pardalos PM, Karakitsiou A. Massive datasets and machine learning for computational biomedicine: trends and challenges. *Ann Oper Res.* (2019) 276:5–34. doi: 10.1007/s10479-018-2891-2
113. Mardis ER. The $1,000 genome, the $100,000 analysis. *Gen Med.* (2010) 2:84. doi: 10.1186/gm205
114. Lu C, Schneider MT, Gubbins P, Leach-Kemon K, Jamison D, Murray CJ. Public financing of health in developing countries: a cross-national systematic analysis. *Lancet.* (2010) 375:1375–87. doi: 10.1016/S0140-6736(10)60233-4
115. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *New Engl J Med.* (2012) 366:489–91. doi: 10.1056/NEJMp1114866
116. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion.* (2019) 50:71–91. doi: 10.1016/j.inffus.2018.09.012
117. Fan Z, He X, Wang L, Lv J, Kang Y. Research on entity relationship extraction for diabetes medical literature. In: *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Vol. 9. (Chongqing: IEEE). 2020. p. 424–430.
118. Bai T, Ge Y, Yang C, Liu X, Gong L, Wang Y, et al. BERST: An engine and tool for exploring biomedical entities and relationships. *Chinese J Electron.* (2019) 28:797–804. doi: 10.1049/cje.2019.05.007
119. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs.* (2014) 33:1163–70. doi: 10.1377/hlthaff.2014.0053
120. Panahiazar M, Taslimitehrani V, Jadhav A, Pathak J. Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases. In: *Big Data (Big Data), 2014 IEEE International Conference on.* Washington, DC: IEEE (2014). p. 790–5.
121. Sadman N, Tasneem S, Haque A, Islam MM, Ahsan MM, Gupta KD. Can NLP techniques be utilized as a reliable tool for medical science?Ĭ-Building a NLP Framework to Classify Medical Reports. In: *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).* Vancouver, BC: IEEE (2020). p. 0159–66.
122. Majewska O, Collins C, Baker S, Björne J, Brown SW, Korhonen A, et al. BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine. *J Biomed Semantics.* (2021) 12:1–13. doi: 10.1186/s13326-021-00247-z
123. Chiu B, Pyysalo S, Vulić I, Korhonen A. Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC Bioinf.* (2018) 19:1–13. doi: 10.1186/s12859-018-2039-z
124. Jovanović J, Bagheri E. Semantic annotation in biomedicine: the current landscape. *J Biomed Semantics.* (2017) 8:1–18. doi: 10.1186/s13326-017-0153-x
125. Cimino JJ, Shortliffe EH, Chiang MF, Blumenthal D, Brennan PF, Frisse M, et al. The future of informatics in biomedicine. In: *Biomedical Informatics.* London, UK: Springer (2021). p. 987–1016.
126. Yang T, Zhao Y. Application of cloud computing in biomedicine big data analysis cloud computing in big data. In: *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET).* Chennai: IEEE (2017). p. 1–3.
127. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Gen.* (2015) 8:33. doi: 10.1186/s12920-015-0108-y
128. Sobeslav V, Maresova P, Krejcar O, Franca TC, Kuca K. Use of cloud computing in biomedicine. *J Biomol Struct Dyn.* (2016) 34:2688–97. doi: 10.1080/07391102.2015.1127182
129. Calabrese B, Cannataro M. Cloud computing in healthcare and biomedicine. *Scalable Comput Pract Exp.* (2015) 16:1–18. doi: 10.12694/scpe.v16i1.1057
130. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat Rev Genet.* (2011) 12:224. doi: 10.1038/nrg2857-c2
131. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearbook Med Inf.* (2014) 23:42–7. doi: 10.15265/IY-2014-0018
132. Marx V. Biology: The big challenges of big data. *Nature.* (2013) 498:255–60. doi: 10.1038/498255a
133. Hiltemann S, Mei H, de Hollander M, Palli I, van der Spek P, Jenster G, et al. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. *GigaScience.* (2014) 3:1. doi: 10.1186/2047-217X-3-1
134. Liu B, Madduri RK, Sotomayor B, Chard K, Lacinski L, Dave UJ, et al. Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *J Biomed Inf.* (2014) 49:119–33. doi: 10.1016/j.jbi.2014.01.005
135. Zheng G, Li H, Wang C, Sheng Q, Fan H, Yang S, et al. A platform to standardize, store, and visualize proteomics experimental data. *Acta Biochimica et Biophysica Sinica.* (2009) 41:273–9. doi: 10.1093/abbs/gmp010
136. Harrow J, Hancock J, Community EE, Blomberg N, Blomberg N, Brunak S, et al. ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future. *EMBO J.* (2021) 40:e107409. doi: 10.15252/embj.2020107409
137. Mora-Márquez F, Vázquez-Poletti JL, de Heredia UL. NGScloud2: optimized bioinformatic analysis using Amazon Web Services. *PeerJ.* (2021) 9:e11237. doi: 10.7717/peerj.11237
138. Moreno P, Pireddu L, Roger P, Goonasekera N, Afgan E, Van Den Beek M, et al. Galaxy-Kubernetes integration: scaling bioinformatics workflows in the cloud. *BioRxiv.* (2018) 488643.
139. Yuan DY, Wildish T. Bioinformatics application with kubeflow for batch processing in clouds. In: *International Conference on High Performance Computing.* Frankfurt: Springer (2020). p. 355–67.
140. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, et al. A view of cloud computing. *Commun ACM.* (2010) 53:50–8. doi: 10.1145/1721654.1721672
141. Lahami M, Krichen M, Alroobaea R. Towards a test execution platform as-a-service: application in the e-health domain. In: *2018 International Conference on Control, Automation and Diagnosis (ICCAD).* Marrakech: IEEE (2018). p. 1–6.
142. Davoody N, Koch S, Krakau I, Hägglund M. Accessing and sharing health information for post-discharge stroke care through a national health information exchange platform-a case study. *BMC Med Inf Decis Making.* (2019) 19:1–16. doi: 10.1186/s12911-019-0816-x

143. Wang L, Lu Z, Van Buren P, Ware D. SciApps: a cloud-based platform for reproducible bioinformatics workflows. *Bioinformatics.* (2018) 34:3917–20. doi: 10.1093/bioinformatics/bty439
144. Namasudra S. Data access control in the cloud computing environment for bioinformatics. *Int J Appl Res Bioinf (IJARB).* (2021) 11:40–50. doi: 10.4018/IJARB.2021010105
145. Thirunavukkarasu GS, Champion B, Horan B, Seyedmahmoudian M, Stojcevski A. Iot-based system health management infrastructure as a service. In: Proceedings of the 2018 International Conference on Cloud Computing and Internet of Things; 2018. p. 55–61.
146. Yustim B. Implementation analysis on society-based hospital concept with software-as-a-service (SaaS) technology. *Int J Eng Technol.* (2018) 7:228–31. doi: 10.14419/ijet.v7i4.33.23565
147. Lakshmisri S. Software as a service in cloud computing. *Int J Creative Res Thoughts (IJCRT).* (2019) 7:2320–2882.
148. Lahami M, Krichen M, Alroobaea R. TEPaaS: test execution platform as-a-service applied in the context of e-health. *Int J Auton Adapt Commun Syst.* (2019) 12:264–83. doi: 10.1504/IJAACS.2019.10022473
149. Soh J, Copeland M, Puca A, Harris M. Overview of azure infrastructure as a service (IaaS) services. In: *Microsoft Azure.* Berkeley, CA: Springer (2020). p. 21–41.
150. Casalicchio E, Iannucci S. The state-of-the-art in container technologies: application, orchestration and security. *Concurrency Comput Pract Exp.* (2020) 32:e5668. doi: 10.1002/cpe.5668
151. Sahni S, Khanna A, Rodrigues JJ. Analysis of biological information using statistical techniques in cloud computing. In: *Applications of Cloud Computing.* Boca Raton, FL: Chapman and Hall/CRC (2020). p. 1–24.
152. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinform.* (2012) 13:1. doi: 10.1186/1471-2105-13-42
153. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Gen Biol.* (2010) 11:R86. doi: 10.1186/gb-2010-11-8-r86
154. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. *Gen Res.* (2002) 12:1611–8. doi: 10.1101/gr.361602
155. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res.* (1997) 25:3389–402. doi: 10.1093/nar/25.17.3389
156. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* (2004) 5:R80. doi: 10.1186/gb-2004-5-10-r80
157. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics.* (2007) 23:673–9. doi: 10.1093/bioinformatics/btm009
158. Thompson JD, Gibson T, Higgins DG, et al. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protocols Bioinf.* (2002) 1:2–3. doi: 10.1002/0471250953.bi0203s00
159. Navale V, Bourne PE. Cloud computing applications for biomedical science: a perspective. *PLoS Comput Biol.* (2018) 14:e1006144. doi: 10.1371/journal.pcbi.1006144
160. Oh M, Park S, Kim S, Chae H. Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Briefings Bioinf.* (2021) 22:66–76. doi: 10.1093/bib/bbaa032v
161. Bartold SP, Hannigan GG. DXplain. *J Med Lib Assoc.* (2002) 90:267.
162. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, Hernandez-Torres I, Rivero-Lopez C, Spicer T, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. *Diagnosis.* (2018) 5:71–76. doi: 10.1515/dx-2017-0045
163. Petiwala FF, Shukla VK, Vyas S. IBM watson: redefining artificial intelligence through cognitive computing. In: *Proceedings of International Conference on Machine Intelligence and Data Science Applications.* Singapore: Springer (2021). p. 173–85.
164. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* (2019) 56:24–31. doi: 10.1109/MSPEC.2019.8678513
165. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? a multicentre experimental study of Isabel. *BMJ Qual Safety.* (2021) 1–8. doi: 10.1136/bmjqs-2021-013493
166. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence—assisted symptom checker: cross-sectional survey study. *J Med Internet Res.* (2020) 22:e14679. doi: 10.2196/14679
167. Davies A, Mueller J, Hassey A, Moulton G. Development of a core competency framework for clinical informatics. *BMJ Health Care Inf.* (2021) 28:e100356. doi: 10.1136/bmjhci-2021-100356
168. Scott PJ, Dunscombe R, Evans D, Mukherjee M, Wyatt JC. Learning health systems need to bridge the'two cultures' of clinical informatics and data science. *J Innov Health Inf.* (2018) 25:126–31. doi: 10.14236/jhi.v25i2.1062
169. Cancilla M, Canalini L, Bolelli F, Allegretti S, Carrión S, Paredes R, et al. The deephealth toolkit: a unified framework to boost biomedical applications. In: *2020 25th International Conference on Pattern Recognition (ICPR).* Milan: IEEE (2021). p. 9881–8.
170. Ping P, Hermjakob H, Polson JS, Benos PV, Wang W. Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine. *Circ Res.* (2018) 122:1290–301. doi: 10.1161/CIRCRESAHA.117.310967
171. Wilson G, Aruliah D, Brown CT, Hong NPC, Davis M, Guy RT, et al. Best practices for scientific computing. *PLoS Biol.* (2014) 12:e1001745. doi: 10.1371/journal.pbio.1001745
172. Cesario A, Auffray C, Russo P, Hood L. P4 medicine needs P4 education. *Curr Pharmaceutical Design.* (2014) 20:6071–2. doi: 10.2174/1381612820666140314145445
173. Hannay JE, MacLeod C, Singer J, Langtangen HP, Pfahl D, Wilson G. How do scientists develop and use scientific software? In: *Proceedings of the 2009 ICSE workshop on Software Engineering for Computational Science and Engineering.* Vancouver, BC: IEEE Computer Society (2009). p. 1–8.
174. Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics.* (2011) 27:1309–10. doi: 10.1093/bioinformatics/btr114
175. Schatz MC, Trapnell C, Delcher AL, Varshney A. High-throughput sequence alignment using graphics processing units. *BMC Bioinf.* (2007) 8:474. doi: 10.1186/1471-2105-8-474
176. Manavski SA, Valle G. CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinf.* (2008) 9:1. doi: 10.1186/1471-2105-9-S2-S10
177. McArt DG, Bankhead P, Dunne PD, Salto-Tellez M, Hamilton P, Zhang SD. cudaMap: a GPU accelerated program for gene expression connectivity mapping. *BMC Bioinf.* (2013) 14:1. doi: 10.1186/1471-2105-14-305
178. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Gen.* (2013) 14:333–46. doi: 10.1038/nrg3433
179. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. *F1000Res.* (2021) 10:33. doi: 10.12688/f1000research.29032.2
180. Larsonneur E, Mercier J, Wiart N, Le Floch E, Delhomme O, Meyer V. Evaluating workflow management systems: a bioinformatics use case. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* Madrid: IEEE (2018). p. 2773–5.
181. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucl Acids Res.* (2013) 41:gkt328. doi: 10.1093/nar/gkt328
182. Aubin MR, da Rosa Righi R, Valiati VH, da Costa CA, Antunes RS, Galante G. Helastic: on combining threshold-based and Serverless elasticity approaches for optimizing the execution of bioinformatics applications. *J Comput Sci.* (2021) 53:101407. doi: 10.1016/j.jocs.2021.101407
183. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *J Ameri Med Inf Assoc.* (2014) 21:969–75. doi: 10.1136/amiajnl-2013-002155

184. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Gen Res.* (2010) 20:1297–1303. doi: 10.1101/gr.107524.110
185. Ahmad T, Ahmed N, Al-Ars Z, Hofstee HP. Optimizing performance of GATK workflows using Apache Arrow In-Memory data framework. *BMC Gen.* (2020) 21:1–14. doi: 10.1186/s12864-020-07013-y
186. Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinf.* (2016) 11:S1. doi: 10.1186/1471-2105-11-S12-S1
187. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* (2013) 23:228–35. doi: 10.1101/gr.141382.112
188. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, et al. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucl Acids Res.* (2013) 41:e142. doi: 10.1093/nar/gkt481
189. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol.* (2011) 29:644. doi: 10.1038/nbt.1883
190. Nanni L, Pinoli P, Canakoglu A, Ceri S. PyGMQL: scalable data extraction and analysis for heterogeneous genomic datasets. *BMC Bioinf.* (2019) 20:1–11. doi: 10.1186/s12859-019-3159-9
191. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Amer J Hum Gen.* (2007) 81:559–75. doi: 10.1086/519795
192. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Gen.* (2006) 38:500–1. doi: 10.1038/ng0506-500
193. Kuehn H, Liberzon A, Reich M, Mesirov JP. Using GenePattern for gene expression analysis. *Curr Protocols Bioinf.* (2008) 22:7–12. doi: 10.1002/0471250953.bi0712s22
194. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protocols Mol Biol.* (2010) 89:19–10. doi: 10.1002/0471142727.mb1910s89
195. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* (2005) 15:1451–5. doi: 10.1101/gr.4086505
196. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Gen.* (2000) 25:25–9. doi: 10.1038/75556
197. Consortium GO, et al. The Gene Ontology (GO) database and informatics resource. *Nucl Acids Res.* (2004) 32:D258–D261. doi: 10.1093/nar/gkh036
198. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* (2004) 20:578–580. doi: 10.1093/bioinformatics/btg455
199. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* (2005) 21:3448–9. doi: 10.1093/bioinformatics/bti551
200. Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. *Nucl Acids Res.* (2013) 41:gkt752. doi: 10.1093/nar/gkt752
201. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303
202. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, et al. Integrating biological data–the distributed annotation system. *BMC bioinf.* (2008) 9:1. doi: 10.1186/1471-2105-9-S8-S3
203. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* (2018) 16:e2005970. doi: 10.1371/journal.pbio.2005970
204. Bray MA, Carpenter AE. Quality control for high-throughput imaging experiments using machine learning in cellprofiler. In: *High Content Screening*. New York, NY: Springer (2018). p. 89–112.
205. Lau YS, Xu L, Gao Y, Han R. Automated muscle histopathology analysis using CellProfiler. *Skeletal Muscle.* (2018) 8:1–9. doi: 10.1186/s13395-018-0178-6
206. Gómez-Romero L, Tovar H, Moreno-Contreras J, Espinoza MA, de Anda-Jáuregui G. Automated reverse transcription polymerase chain reaction data analysis for sars-CoV-2 detection. *Revista de Investigacion Clinica; Organo del Hospital de Enfermedades de la Nutricion.* (2021) 73:339–46. doi: 10.24875/RIC.21000189
207. Santus E, Marino N, Cirillo D, Chersoni E, Montagud A, Chadha AS, et al. Artificial Intelligence–Aided Precision Medicine for COVID-19: Strategic Areas of Research and Development. *J Med Internet Res.* (2021) 23:e22453. doi: 10.2196/22453
208. Cavelaars M, Rousseau J, Parlayan C, de Ridder S, Verburg A, Ross R, et al. OpenClinica. *J Clin Bioinf.* (2015) 5:1–2. doi: 10.1186/2043-9113-5-S1-S2
209. Bauer C, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. The integrated data repository toolkit (IDRT): accelerating translational research infrastructures. *J Clin Bioinf.* (2015) 5:1–2. doi: 10.1186/2043-9113-5-S1-S6
210. Gilotay C, Lejeune S. VISTA Trials. *J Clin Bioinf.* (2015) 5:1–2. doi: 10.1186/2043-9113-5-S1-S4
211. Moni MA, Liò P. comoR: a software for disease comorbidity risk assessment. *J Clin Bioinf.* (2014) 4:1–11. doi: 10.1155/2014/472045
212. Noll J, Beecham S, Seichter D. A qualitative study of open source software development: the open EMR project. In: *2011 International Symposium on Empirical Software Engineering and Measurement.* Banff, AB: IEEE (2011). p. 30–9.
213. Bashiri A, Ghazisaeedi M. Open MRS softwares: effective approaches in management of patients' health information. *Int J Commun Med Publ Health.* (2017) 4:3948–51. doi: 10.18203/2394-6040.ijcmph20174803
214. Jones B, Yuan X, Nuakoh E, Ibrahim K. Survey of open source health information systems. *Health Inform.* (2014) 3:23–31. doi: 10.5121/hiij.2014.3102
215. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Int Med.* (2018) 178:1544–7. doi: 10.1001/jamainternmed.2018.3763
216. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inf Assoc.* (2018) 25:1419–28. doi: 10.1093/jamia/ocy068
217. Choi L, Beck C, McNeer E, Weeks HL, Williams ML, James NT, et al. Development of a system for postmarketing population pharmacokinetic and pharmacodynamic studies using real-world data from electronic health records. *Clin Pharmacol Therapeutics.* (2020) 107:934–43. doi: 10.1002/cpt.1787
218. Choi L, Carroll RJ, Beck C, Mosley JD, Roden DM, Denny JC, et al. Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects. *Bioinformatics.* (2018) 34:2988–96. doi: 10.1093/bioinformatics/bty306
219. Springate DA, Parisi R, Olier I, Reeves D, Kontopantelis E. rEHR: An R package for manipulating and analysing Electronic Health Record data. em PLoS ONE. (2017) 12:e0171784. doi: 10.1371/journal.pone.0171784
220. Lawton T, McCooe M. A novel modelling technique to predict resource requirements in critical care–a case study. *Future Healthcare J.* (2019) 6:17. doi: 10.7861/futurehosp.6-1-17
221. Cornelissen G. Cosinor-based rhythmometry. *Theor Biol Med Model.* (2014) 11:1–24. doi: 10.1186/1742-4682-11-16
222. Russell PH, Ghosh D. Radtools: r utilities for convenient extraction of medical image metadata. *F1000Res.* (2018) 7:1–12. doi: 10.12688/f1000research.17139.1
223. Wang C, Li H, Chen WC, Lu N, Tiwari R, Xu Y, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *J Biopharmaceutical Stat.* (2019) 29:731–48. doi: 10.1080/10543406.2019.1657133
224. Chen WC, Wang C, Li H, Lu N, Tiwari R, Xu Y, et al. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *J Biopharmaceutical Stat.* (2020) 30:508–20. doi: 10.1080/10543406.2020.1730877
225. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol.* (2014) 8:I1. doi: 10.1186/1752-0509-8-S2-I1
226. Hernández-de Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, Conesa A. STATegra EMS: an experiment management system

for complex next-generation omics experiments. *BMC Syst Biol.* (2014) 8:1. 10.1186/1752-0509-8-S2-S9
227. Conesa A, Mortazavi A. The common ground of genomics and systems biology. *BMC Syst Biol.* (2014) 8:S1. doi: 10.1186/1752-0509-8-S2-S1
228. Attwood T, Bongcam-Rudloff E, Gisel A. SEQAHEAD-COST action BM1006: next generation sequencing data analysis network. *EMBnet J.* (2011) 17:7. doi: 10.14806/ej.17.1.218
229. Bernasconi A, Canakoglu A, Masseroli M, Ceri S. The road towards data integration in human genomics: players, steps and interactions. *Briefings Bioinf.* (2021) 22:30–44. doi: 10.1093/bib/bbaa080
230. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature.* (2020) 583:699–710. doi: 10.1038/s41586-020-2493-4
231. Saunders G, Baudis M, Becker R, Beltran S, Béroud C, Birney E, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Gen.* (2019) 20:693–701. doi: 10.1038/s41576-019-0156-9
232. Vazquez M, Valencia A. Patient Dossier: Healthcare queries over distributed resources. *PLoS Comput Biol.* (2019) 15:e1007291. doi: 10.1371/journal.pcbi.1007291
233. Shaibi GQ, Kullo IJ, Singh DP, Hernandez V, Sharp RR, Cuellar I, et al. Returning genomic results in a Federally Qualified Health Center: the intersection of precision medicine and social determinants of health. *Gen Med.* (2020) 22:1552–9. doi: 10.1038/s41436-020-0806-5
234. Rajewsky N, Almouzni G, Gorski SA, Aerts S, Amit I, Bertero MG, et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature.* (2020) 587:377–86. doi: 10.1038/s41586-020-2715-9
235. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database.* (2020) 2020:1–35. doi: 10.1093/database/baaa010
236. Arrison T, Weidman S, et al. *Steps Toward Large-Scale Data Integration in the Sciences:: Summary of a Workshop.* Washington, DC: National Academies Press (2010).
237. Ziegler P, Dittrich KR. Three decades of data integration-All problems solved? In: *IFIP Congress Topical Sessions.* Toulouse: Springer (2004). p. 3–12.
238. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. *Clin Transl Sci.* (2021) 14:86–93. doi: 10.1111/cts.12884
239. Akil H, Martone ME, Van Essen DC. Challenges and opportunities in mining neuroscience data. *Science (New York, NY).* (2011) 331:708. doi: 10.1126/science.1199305
240. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag.* (1996) 17:37.
241. Abugessaisa I. Knowledge discovery in road accidents database-integration of visual and automatic data mining methods. *Int J Publ Inf Syst.* (2008) 4:59–85.
242. Morgenstern JD, Rosella LC, Daley MJ, Goel V, Schünemann HJ, Piggott T. AI's gonna have an impact on everything in society, so it has to have an impact on public health: a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. *BMC Publ Health.* (2021) 21:1–14. doi: 10.1186/s12889-020-10030-x
243. Rundo L, Pirrone R, Vitabile S, Sala E, Gambino O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *J Biomed Inf.* (2020) 108:103479. doi: 10.1016/j.jbi.2020.103479
244. Čyras K, Oliveira T, Karamlou A, Toni F. Assumption-based argumentation with preferences and goals for patient-centric reasoning with interacting clinical guidelines. *Argument Comput.* (2020) 1–41.
245. Leonelli S. Introduction: making sense of data-driven research in the biological and biomedical sciences. *Stud Hist Philos Biol Biomed Sci.* (2012) 43:1–3. doi: 10.1016/j.shpsc.2011.10.001
246. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME) toward standards for microarray data. *Nat Gen.* (2001) 29:365–371. doi: 10.1038/ng1201-365
247. Brazma A. Minimum information about a microarray experiment (MIAME)–successes, failures, challenges. *Sci World J.* (2009) 9:420–423. doi: 10.1100/tsw.2009.57
248. Simoneau J, Dumontier S, Gosselin R, Scott MS. Current RNA-seq methodology reporting limits reproducibility. *Briefings Bioinf.* (2021) 22:140–5. doi: 10.1093/bib/bbz124
249. Füllgrabe A, George N, Green M, Nejad P, Aronow B, Fexova SK, et al. Guidelines for reporting single-cell RNA-seq experiments. *Nat Biotechnol.* (2020) 38:1384–6. doi: 10.1038/s41587-020-00744-z
250. Marble HD, Huang R, Dudgeon SN, Lowe A, Herrmann MD, Blakely S, et al. A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients. *J Pathol Inf.* (2020) 11:22. doi: 10.4103/jpi.jpi_27_20
251. Hinkson IV, Davidsen TM, Klemm JD, Chandramouliswaran I, Kerlavage AR, Kibbe WA. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front Cell Develop Biol.* (2017) 5:83. doi: 10.3389/fcell.2017.00083
252. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med.* (2020) 26:1320–4. doi: 10.1038/s41591-020-1041-y
253. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inf Assoc.* (2014) 21:578–82. doi: 10.1136/amiajnl-2014-002747
254. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building data infrastructure to evaluate and improve quality: PCORnet. *J Oncol Pract.* (2015) 11:204–6. doi: 10.1200/JOP.2014.003194
255. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet®). *Egems.* (2018) 6:1–9. doi: 10.5334/egems.199
256. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Amer Med Inf Assoc.* (2010) 17:124–30. doi: 10.1136/jamia.2009.000893
257. Klann JG, Joss MA, Embree K, Murphy SN. Data model harmonization for the all of us research program: transforming i2b2 data into the OMOP common data model. *PLoS One.* (2019) 14:e0212463. doi: 10.1371/journal.pone.0212463
258. Bucalo M, Gabetta M, Chiudinelli L, Larizza C, Bellasi A, Zambelli A, et al. i2b2 to optimize patients enrollment. *Stud Health Technol Inf.* (2021) 281:506–7. doi: 10.3233/SHTI210217
259. Ravikumar K, Rastegar-Mojarad M, Liu H. BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database.* (2017) 2017:baw156. doi: 10.1093/database/baw156
260. Touré V, Flobak Å, Niarakis A, Vercruysse S, Kuiper M. The status of causality in biological databases: data resources and data retrieval possibilities to support logical modeling. *Briefings Bioinf.* (2021) 22:bbaa390. doi: 10.1093/bib/bbaa390
261. Guryanova S, Guryanova A. sbv IMPROVER: modern approach to systems biology. In: *Biological Networks and Pathway Analysis.* New York, NY: Springer (2017). p. 21–9.
262. Mahendra M, Luo Y, Mills H, Schenk G, Butte AJ, Dudley RA. Impact of Different Approaches to Preparing Notes for Analysis With Natural Language Processing on the Performance of Prediction Models in Intensive Care. *Crit Care Explor.* (2021) 3:e0450. doi: 10.1097/CCE.0000000000000450
263. Hong JC, Butte AJ. Assessing Clinical Outcomes in a Data-Rich World A Reality Check on Real-World Data. *JAMA Netw Open.* (2021) 4:e2117826. doi: 10.1001/jamanetworkopen.2021.17826
264. Patel NA, Butte AJ. Characteristics and challenges of the clinical pipeline of digital therapeutics. *NPJ Digit Med.* (2020) 3:1–5. doi: 10.1038/s41746-020-00370-8
265. Butte AJ. Trials and Tribulations 11 Reasons Why We Need to Promote Clinical Trials Data Sharing. *JAMA Netw Open.* (2021) 4:e2035043. doi: 10.1001/jamanetworkopen.2020.35043

266. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA.* (2020) 324:1397–1398. doi: 10.1001/jama.2020.9371
267. Harvey H, Glocker B. A standardised approach for preparing imaging data for machine learning tasks in radiology. In: *Artificial Intelligence in Medical Imaging.* Cham: Springer (2019). p. 61–72.
268. Özdemir V, Kolker E, Hotez PJ, Mohin S, Prainsack B, Wynne B, et al. Ready to put metadata on the post-2015 development agenda? linking data publications to responsible innovation and science diplomacy. *Omics J Integr Biol.* (2014) 18:1–9. doi: 10.1089/omi.2013.0170
269. Snyder M, Mias G, Stanberry L, Kolker E. Metadata checklist for the integrated personal omics study: proteomics and metabolomics experiments. *Big Data.* (2013) 1:202–6. doi: 10.1089/big.2013.0040
270. Snyder M, Mias G, Stanberry L, Kolker E. Metadata checklist for the integrated personal OMICS study: proteomics and metabolomics experiments. *Omics J Integr Biol.* (2014) 18:81–5. doi: 10.1089/omi.2013.0148
271. Kolker E, Özdemir V, Martens L, Hancock W, Anderson G, Anderson N, et al. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics J Integr Biol.* (2014) 18:10–14. doi: 10.1089/omi.2013.0149
272. Park H, Sakaori F, Konishi S. Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria. *J Stat Comput Simulat.* (2014) 84:1596–607. doi: 10.1080/00949655.2012.755532
273. Bühlmann P, Van De Geer S. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Berlin: Springer Science & Business Media (2011).
274. Hand DJ. Deconstructing statistical questions. *J Roy Stat Soc Series A (Stat Soc).* (1994) 157:317–356. doi: 10.2307/2983526
275. Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann Stat.* (2008) 36:1567–94. doi: 10.1214/07-AOS520
276. Isci S, Dogan H, Ozturk C, Otu HH. Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics.* (2014) 30:860–7. doi: 10.1093/bioinformatics/btt643
277. Reshetova P, Smilde AK, van Kampen AH, Westerhuis JA. Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. *BMC Syst Biol.* (2014) 8:S2. doi: 10.1186/1752-0509-8-S2-S2
278. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinf.* (2014) 15:1. doi: 10.1186/1471-2105-15-162
279. Fagan A, Culhane AC, Higgins DG. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics.* (2007) 7:2162–71. doi: 10.1002/pmic.200600898
280. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci USA.* (2003) 100:3351–6. doi: 10.1073/pnas.0530258100
281. Yeung ES. Genome-wide correlation between mrna and protein in a single cell. *Angewandte Chemie Int Edn.* (2011) 50:583–5. doi: 10.1002/anie.201005969
282. Van den Bulcke T, Lemmens K, Van de Peer Y, Marchal K. Inferring transcriptional networks by mining 'omics' data. *Curr Bioinf.* (2006) 1:301–13. doi: 10.2174/157489306777827991
283. Yoo S, Huang T, Campbell JD, Lee E, Tu Z, Geraci MW, et al. MODMatcher: Multi-Omics Data Matcher for Integrative Genomic Analysis. *PLoS Comput Biol.* (2014) 10:e1003790. doi: 10.1371/journal.pcbi.1003790
284. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. *PLoS One.* (2012) 7:e29348. doi: 10.1371/journal.pone.0029348
285. Wang XF. Joint generalized models for multidimensional outcomes: a case study of neuroscience data from multimodalities. *Biometric J.* (2012) 54:264–80. doi: 10.1002/bimj.201100041
286. Hu P, Greenwood CM, Beyene J. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinf.* (2005) 6:128. doi: 10.1186/1471-2105-6-128
287. Fan J, Han F, Liu H. Challenges of big data analysis. *Nat Sci Rev.* (2014) 1:293–314. doi: 10.1093/nsr/nwt032
288. Fan J, Fan Y. High dimensional classification using features annealed independence rules. *Ann Stat.* (2008) 36:2605. doi: 10.1214/07-AOS504
289. Hall P, Pittelkow Y, Ghosh M. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J Roy Stat Soc Series B (Stat Methodol).* (2008) 70:159–73. doi: 10.1111/j.1467-9868.2007.00631.x
290. Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J Roy Stat Soc Series B (Stat Methodol).* (2012) 74:37–65. doi: 10.1111/j.1467-9868.2011.01005.x
291. Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n. *Ann Stat.* (2007) 35:2313–51. doi: 10.1214/009053606000001523
292. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat.* (2010) 894–942. doi: 10.1214/09-AOS729
293. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Stat Assoc.* (2001) 96:1348–60. doi: 10.1198/016214501753382273
294. Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. *J Comput Graph Stat.* (2012) 18:533–50. doi: 10.1198/jcgs.2009.08041
295. Genovese CR, Jin J, Wasserman L, Yao Z. A comparison of the lasso and marginal regression. *J Mach Learn Res.* (2012) 13:2107–43.
296. Alfayez R. Do Healthcare Metadata Models Designed for Web Publishing Meet the Accreditation Standards? A Case Study in the Healthcare and Medical Education. *Electron J e-Learn.* (2020) 18:356–69. doi: 10.34190/EJEL.20.18.4.008
297. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. *BMC Med Res Methodol.* (2016) 16:1–9. doi: 10.1186/s12874-016-0164-9
298. Swedlow JR, Kankaanpää P, Sarkans U, Goscinski W, Galloway G, Malacrida L, et al. A global view of standards for open image data formats and repositories. *Nat Methods.* (2021) 18:1–7. doi: 10.1038/s41592-021-01113-7
299. Badawy R, Hameed F, Bataille L, Little MA, Claes K, Saria S, et al. Metadata concepts for advancing the use of digital health technologies in clinical research. *Digit Biomarkers.* (2019) 3:116–132. doi: 10.1159/000502951
300. El-Achkar TM, Eadon MT, Menon R, Lake BB, Sigdel TK, Alexandrov T, et al. A multimodal and integrated approach to interrogate human kidney biopsies with rigor and reproducibility: guidelines from the Kidney Precision Medicine Project. *Physiol Genomics.* (2021) 53:1–11. doi: 10.1152/physiolgenomics.00104.2020
301. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data.* (2020) 7:1–4. doi: 10.1038/s41597-020-0524-5
302. Rajesh A, Chang Y, Abedalthagafi MS, Wong-Beringer A, Love MI, Mangul S. Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* (2021) 22:106. doi: 10.1186/s13059-021-02332-z
303. Bittner MI. Rethinking data and metadata in the age of machine intelligence. *Patterns.* (2021) 2:100208. doi: 10.1016/j.patter.2021.100208
304. Schneeweiss S. Learning from big health care data. *New Engl J Med.* (2014) 370:2161–3. doi: 10.1056/NEJMp1401111
305. Bizer C, Boncz P, Brodie ML, Erling O. The meaningful use of big data: four perspectives–four challenges. *ACM SIGMOD Record.* (2012) 40:56–60. doi: 10.1145/2094114.2094129
306. Tene O, Polonetsky J. Privacy in the age of big data: a time for big decisions. *Stanford Law Rev Online.* (2012) 64:63.
307. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. *BioMed Central.* (2019) 20:128. doi: 10.1186/s13059-019-1741-0
308. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. *Nat Biotechnol.* (2018) 36:547–51. doi: 10.1038/nbt.4108
309. Hie B, Cho H, Berger B. Realizing private and practical pharmacological collaboration. *Science.* (2018) 362:347–50. doi: 10.1126/science.aat4807
310. Rahmouni HB, Solomonides T, Mont MC, Shiu S. Modelling and enforcing privacy for medical data disclosure across Europe. In: *MIE.* Bristol, UK (2009). p. 695–9.

311. Kim JW, Edemacu K, Jang B. MPPDS: multilevel privacy-preserving data sharing in a collaborative eHealth system. *IEEE Access.* (2019) 7:109910–23. doi: 10.1109/ACCESS.2019.2933542
312. Jana B, Poray J, Mandal T, Kule M. A multilevel encryption technique in cloud security. In: *2017 7th International Conference on Communication Systems and Network Technologies (CSNT).* Nagpur: IEEE (2017). p. 220–4.
313. Servos D. *A Role and Attribute Based Encryption Approach To Privacy and Security in Cloud Based Health Services.* Thunder Bay, ON: Lakehead University (2012).
314. Friedman A, Schuster A. Data mining with differential privacy. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY (2010). p. 493–502.
315. Hassan MU, Rehmani MH, Chen J. Differential privacy techniques for cyber physical systems: a survey. *IEEE Commun Surveys Tutorials.* (2019) 22:746–89. doi: 10.1109/COMST.2019.2944748
316. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Gen.* (2012) 44:603–8. doi: 10.1038/ng.2248
317. Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol.* (2012) 8:612. doi: 10.1038/msb.2012.47
318. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* (2008) 4:e1000167. doi: 10.1371/journal.pgen.1000167
319. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. *Lancet.* (2010) 375:1525–35. doi: 10.1016/S0140-6736(10)60452-7
320. Ormond KE, Wheeler MT, Hudgins L, Klein TE, Butte AJ, Altman RB, et al. Challenges in the clinical application of whole-genome sequencing. *Lancet.* (2010) 375:1749–51. doi: 10.1016/S0140-6736(10)60599-5
321. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* (2011) 7:e1002280. doi: 10.1371/journal.pgen.1002280
322. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA.* (2014) 311:1035–45. doi: 10.1001/jama.2014.1717
323. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit. Med.* (2020) 3:1–8. doi: 10.1038/s41746-020-0258-y
324. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med.* (2018) 15:e1002689. doi: 10.1371/journal.pmed.1002689
325. Yoon HJ. Medical big data for smart healthcare. *Ann Hepato Biliary Pancreatic Surgery.* (2021) 25:S27. doi: 10.14701/ahbps.BP-SY-3-2
326. Dimitrov DV. Blockchain applications for healthcare data management. *Healthcare Inf Res.* (2019) 25:51–6. doi: 10.4258/hir.2019.25.1.51
327. Spencer A, Patel S. Applying the data protection act 2018 and general data protection regulation principles in healthcare settings. *Nurs Manag.* (2019) 26:34–40. doi: 10.7748/nm.2019.e1806
328. Ballantyne A, Stewart C. Big data and public-private partnerships in healthcare and research. *Asian Bioethics Rev.* (2019) 11:315–26. doi: 10.1007/s41649-019-00100-7
329. Michener WK. Ten simple rules for creating a good data management plan. *PLoS Comput Biol.* (2015) 11:e1004525. doi: 10.1371/journal.pcbi.1004525
330. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. *Ten Simple Rules for Responsible Big Data Research.* San Francisco, CA: Public Library of Science (2017).
331. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. *PLoS Comput Biol.* (2014) 10:e1003542. doi: 10.1371/journal.pcbi.1003542
332. Mietchen D. The transformative nature of transparency in research funding. *PLoS Biol.* (2014) 12:e1002027. doi: 10.1371/journal.pbio.1002027
333. Miksa T, Cardoso J, Borbinha J. Framing the scope of the common data model for machine-actionable data management plans. In: *2018 IEEE International Conference on Big Data (Big Data).* Seattle, WA: IEEE (2018). p. 2733–42.
334. Gu W, Hasan S, Rocca-Serra P, Satagopam VP. Road to effective data curation for translational research. *Drug Disc Today.* (2020) 26:626–30. doi: 10.1016/j.drudis.2020.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Medicine

ORIGINAL RESEARCH
published: 24 February 2022
doi: 10.3389/fmed.2022.737662

Development of a Method for Quantitative Evaluation of Facial Swelling in a Rat Model of Cerebral Ischemia by Facial Image Processing

Yanfei Liu[1,2], *Hui Huang*[3], *Yiwen Li*[1], *Jing Cui*[1], *Tiejun Tong*[4], *Hongjun Yang*[5]* *and Yue Liu*[1]*

[1] National Clinical Research Centre for Chinese Medicine Cardiology, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing, China, [2] Second Department of Geriatrics, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing, China, [3] Beijing Duan-Dian Pharmaceutical Research & Development Co., Ltd., Beijing, China, [4] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China, [5] Medical Experimental Center, China Academy of Chinese Medical Sciences, Beijing, China

A quantitative method for the evaluation of facial swelling in rats with middle cerebral artery occlusion (MCAO) was established using a mathematical method for the first time. The rat model of MCAO was established *via* bilateral common carotid artery ligation. Three groups of rats with the same baseline were selected (model group, positive drug group, and control group) according to their behavioral score and body weight 24 h after surgery. Drug administration was initiated on post-MCAO day 8 and was continued for 28 days. Mobile phones were used to collect facial images at different time points after surgery. In facial image analysis, the outer canthi of both eyes were used as the facial dividing line, and the outer edge of the rat's face was framed using the marking method, and the framed part was regarded as the facial area (S) of the rats. The histogram created with Photoshop CS5 was used to measure the face area in pixels. The distance between the outer canthi of both eyes (Le) and vertical line from the tip of the nose to the line joining the eyes was recorded as H1, and the line from the tip of the nose to the midpoint of the line joining the eyes was recorded as H2. The facial area was calibrated based on the relationship between H1 and H2. The distance between the eyes was inversely proportional to the distance between the rats and mobile phone such that the face area was calibrated by unifying Le. The size of Le between the eyes was inversely proportional to the distance between the rats and mobile phone. This was used to calibrate the face area. When compared with the control group, the facial area of the model group gradually increased from postoperative day 1 to day 7, and there was a significant difference in the facial area of the model group on postoperative day 7. Hence, positive drugs exhibited the effect of improving facial swelling. H1 and H2 can reflect the state of turning the head and raising the head of the rats, respectively. Facial area was calibrated according to the relationship between H1 and H2, which had no obvious effect on the overall conclusion. Furthermore, mobile phone lens was used to capture the picture of rat face, and the distance between the eyes and H1 and H2 was used to calibrate the facial area. Hence, this method is convenient and can be used to evaluate subjective judgment of the human eyes *via* a quantitative method.

Keywords: ischemic stroke, occlusion of the middle cerebral artery, artificial intelligence, facial swelling, rat models

OPEN ACCESS

Edited by:
Yuguang Wang,
Shanghai Jiao Tong University, China

Reviewed by:
Wenyue Su,
Western University of Health Sciences, United States
Zhengbang Li,
Central China Normal University, China

***Correspondence:**
Hongjun Yang
hjyang@icmm.ac.cn
Yue Liu
liuyueheart@hotmail.com

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *07 July 2021*
Accepted: *01 February 2022*
Published: *24 February 2022*

Citation:
Liu Y, Huang H, Li Y, Cui J, Tong T, Yang H and Liu Y (2022) Development of a Method for Quantitative Evaluation of Facial Swelling in a Rat Model of Cerebral Ischemia by Facial Image Processing. Front. Med. 9:737662. doi: 10.3389/fmed.2022.737662

INTRODUCTION

An important pathogenesis of ischemic stroke, which is a major disease that threatens human health, is the cascade of cerebral artery embolism and consequent inflammatory response (1, 2). The middle cerebral artery occlusion (MCAO) model is a clinically common simulation of ischemic stroke, which is less invasive and exhibits the closest resemblance to human ischemic stroke (3). Specifically, bilateral common carotid artery ligation and reperfusion is often employed in rats or mice to establish MCAO rat models for simulating the clinical features of ischemic stroke and performing pharmacodynamic evaluation. In the course of a routine rat MCAO model establishment and drug evaluation experiment, we determined that facial swelling occured in each group of rats after surgery, and the changes in facial swelling in each group exhibited certain characteristics as the duration of drug intervention increased. To avoid subjective evaluation via gross examination, we attempted to develop a convenient method for quantitatively evaluating facial swelling characteristics of the model rats. Furthermore, we employed some mathematical methods to maximally reduce the bias due to human manipulation to provide a multi-dimensional quantitative index for future pharmacodynamic evaluation (4, 5).

MATERIALS AND METHODS

Materials

Thirty male specific-pathogen free (SPF)-grade 10-week-old Sprague-Dawley rats weighing 220–270 g were purchased from Beijing Vital River Laboratory Animal Technology Co., Ltd. (license number: SCXK (Beijing) 2016-0006) and housed in SPF-grade animal facilities. Donepezil hydrochloride tablets were purchased from Zhejiang Huahai Pharmaceutical Co., Ltd. (NMPA approval no. H20183417, lot number: 1426J20004). The positive drug used in this study was donepezil hydrochloride tablets, which was often used as a positive drug in vascular dementia and cerebral ischemia experiments (6).

Methodology

MCAO Procedure and New Findings

The rats were anesthetized with 1 ml/100 g of 4% chloral hydrate via intraperitoneal injection and 1-cm incisions were made on the left and right regions of the neck. Blunt dissection of the superficial fascia was performed wherein the superficial fascia and intermuscular space among the digastric, sternocleidomastoid, and omohyoid muscles were separated. The bilateral common carotid arteries and vagus nerve were exposed. Furthermore, the common carotid artery and vagus nerve were carefully separated and two sutures were passed through the common carotid artery at the proximal and distal ends. The sutures were retained on the lateral side of the wound. The wound was sutured and ligation was maintained for 10 min, followed by 10 min of reperfusion. These steps were repeated three times. After the last reperfusion, the sutures were removed from the wound and the common carotid artery was permanently ligated with double sutures, and the right common carotid artery was ligated in the same manner as the left one. During the period from postoperative day 1 to day 35 at the end of the experiment, the model group showed significant facial swelling compared with the normal group that did not undergo surgery (**Figure 1**).

Grouping and Drug Administration

Twenty-four rats that underwent MCAO were bifactorally grouped according to behavioral scores and body weight after postoperative 24 h. After excluding rats with different baseline values, three groups of six rats, each with the same baseline values, were selected and classified into model group and positive drug group, and six rats that did not undergo MCAO were assigned to the control group. In the positive drug group, donepezil hydrochloride tablet was administered by gavage at a dose of 0.5 mg/(kg.d), and the model and control groups were provided equal volumes of distilled water via gavage daily. Administration was commenced on post-MCAO day 8 and was continued for 28 days.

Facial Image Acquisition and Analysis Process

To further analyze the characteristics of facial swelling, we acquired facial images using a camera at different time points after each group of rats recovered autonomous behavior after performing MCAO. The experiment was divided into two stages. The first stage was from MCAO to the period before the administration of drugs. This stage lasted a total of 7 days, and facial images were collected on postoperative days 1, 3, 5, and 7. The second phase was from grouping and administration to the end of the experiment, and facial images were acquired on postoperative days 8, 12, 16, 20, 24, and 28. The drug administration was commenced on day 8 (**Figure 2**).

The body and head of the rat were fixed with both hands during acquisition, and attempts were made such that the face of the rat faced the camera during photography. In particular, the head elevation angle and head rotation restriction were maintained as consistent to the maximum extent each time. Three images were acquired for each rat, and the image with the best angle and clarity was selected for calculation during analysis. The acquisition device was HUAWEI YAL-AL10, a mobile phone camera. The resolution of the camera at the time of photography was set as 72 × 72 DPI, and the image size was 3,000 × 4,000 pixels. The acquired images were imported into Photoshop CS5. The facial images were analyzed using the outer canthi of both eyes as the facial segmentation line. Furthermore, the outer edges of the face of the rat were boxed using markers and the boxed portion was considered as the total facial area (facial area, S) of the rats. Finally, the facial area was measured in pixels using the histogram in Photoshop CS5 (**Figure 3**). The distance between the outer canthi of both eyes (Le) and vertical line from the tip of the nose to the line joining the eyes was recorded as H1, and the line from the tip of the nose to the midpoint of the line joining the eyes was recorded as H2. The length was measured in pixels using the histogram tool in the software (4, 5). Thus, by following this method (4, 5), we acquired facial images of each group of rats at each postoperative time point (**Figure 3**).

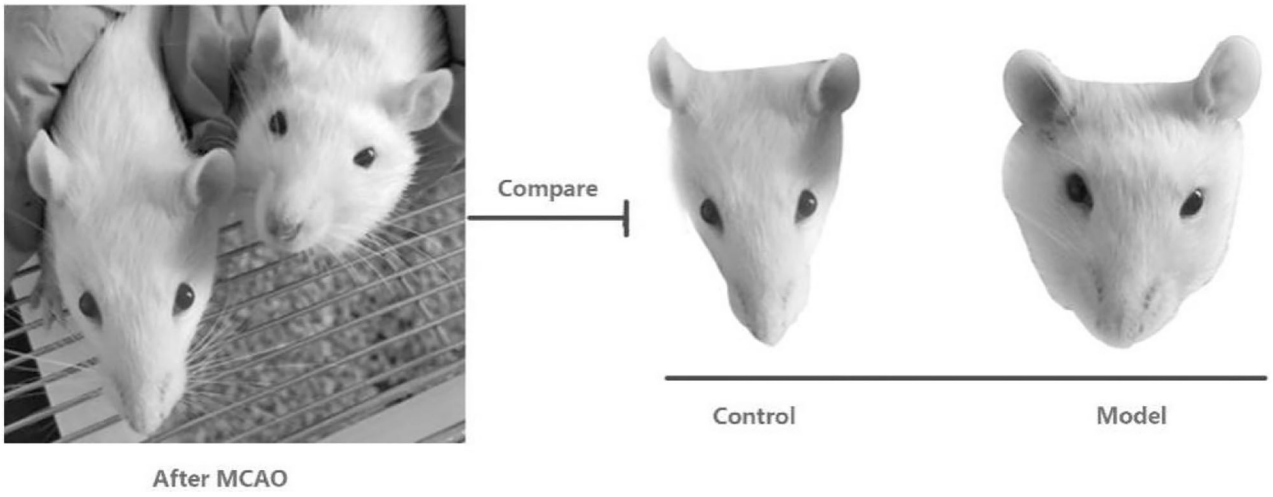

FIGURE 1 | Comparative pictures of facial swelling after surgery in rats treated with middle cerebral artery occlusion.

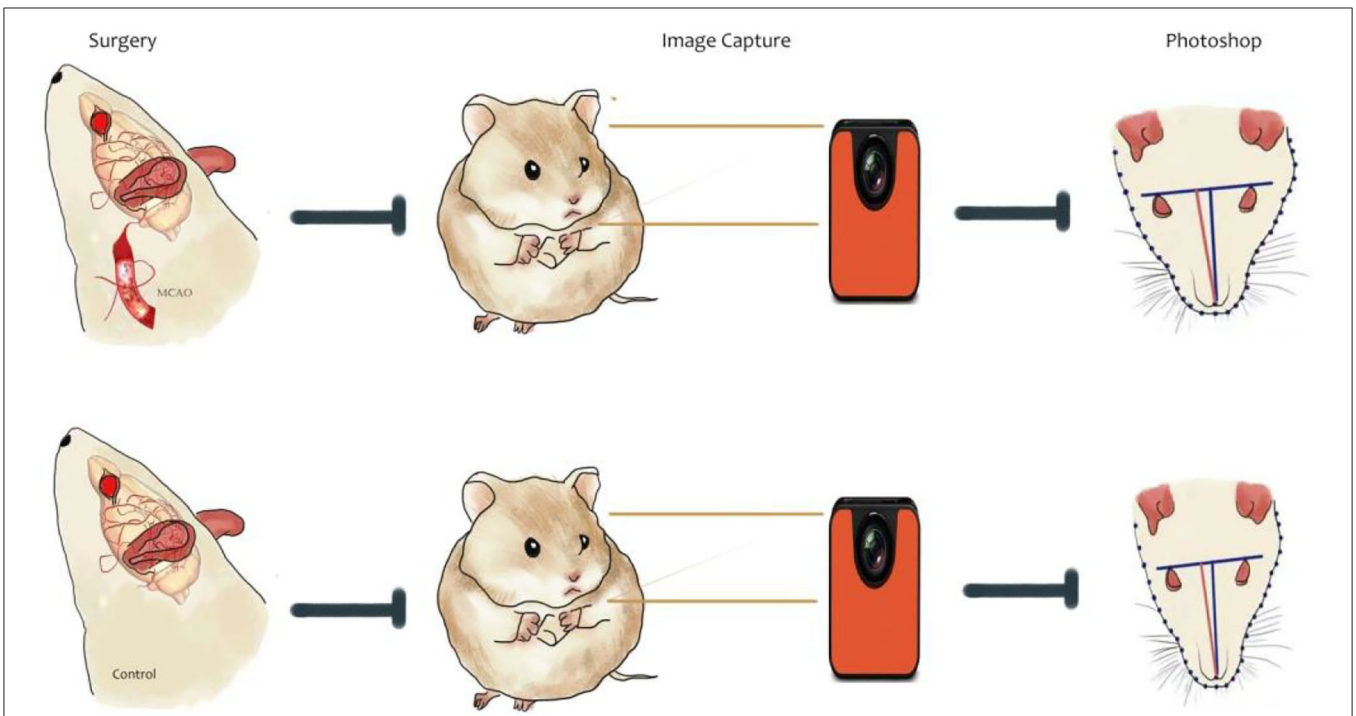

FIGURE 2 | Schematic of the flow of facial image acquisition and analysis after cerebral ischemia in rats.

Facial Image Calibration Methods

We didn't anesthetize the rats or install any assembly equipment to make the whole process as easy as possible. Therefore, in the process of taking photos, there would be uncontrollable factors such as the distance between the lens and the target and the head swing of rats in different directions, but we found that these problems can be corrected by simple mathematical methods. During the facial image acquisition process, we determined that the distance of the camera lens from the target object, the lifting or lowering of the head of the rat, and frontal and head-turned images affected the facial image acquisition and results. To minimize this interference, we utilized a simple mathematical principle (4, 5) for calibration (**Figure 4**). When the lens is turned away from the face of the rat, the distance Le between the outer canthi of the eyes decreases and point S decreases accordingly. When the face of the rat is lifted,

FIGURE 3 | Images of facial recognition at different time points after surgery in the model group of rats.

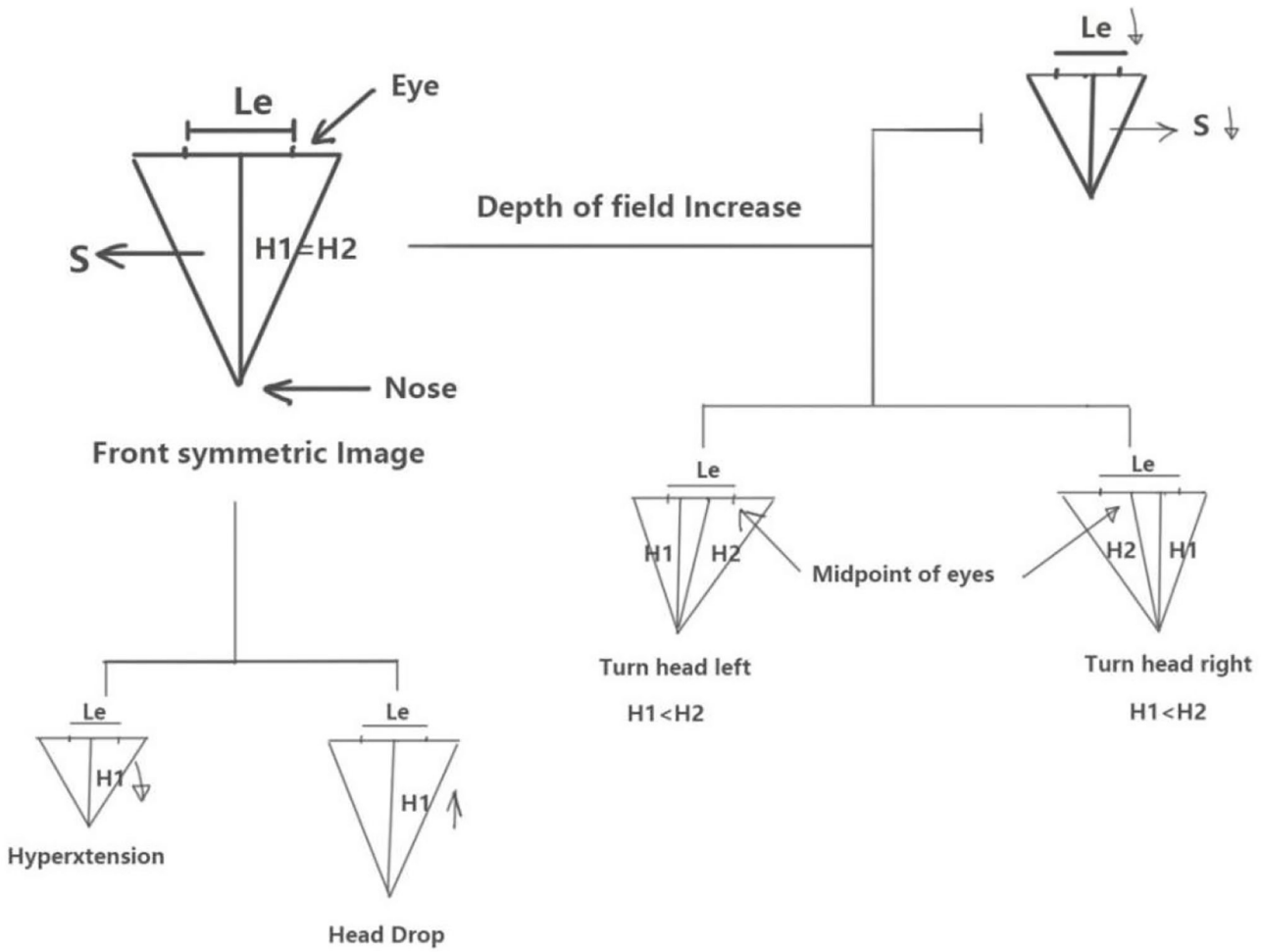

FIGURE 4 | Schematic of the calibration method using mathematical interpretation of head–facial variables.

H1 shortens and S decreases, while Le is assumed to be constant. Conversely, when the head of the rat is lowered, H1 and S increase while Le is assumed as constant. When the rat is facing the lens, H1 = H2, and when the rat turns its head, either to the left or to the right, H2 > H1. When this occurs, the facial area S of the rat also appears to increase

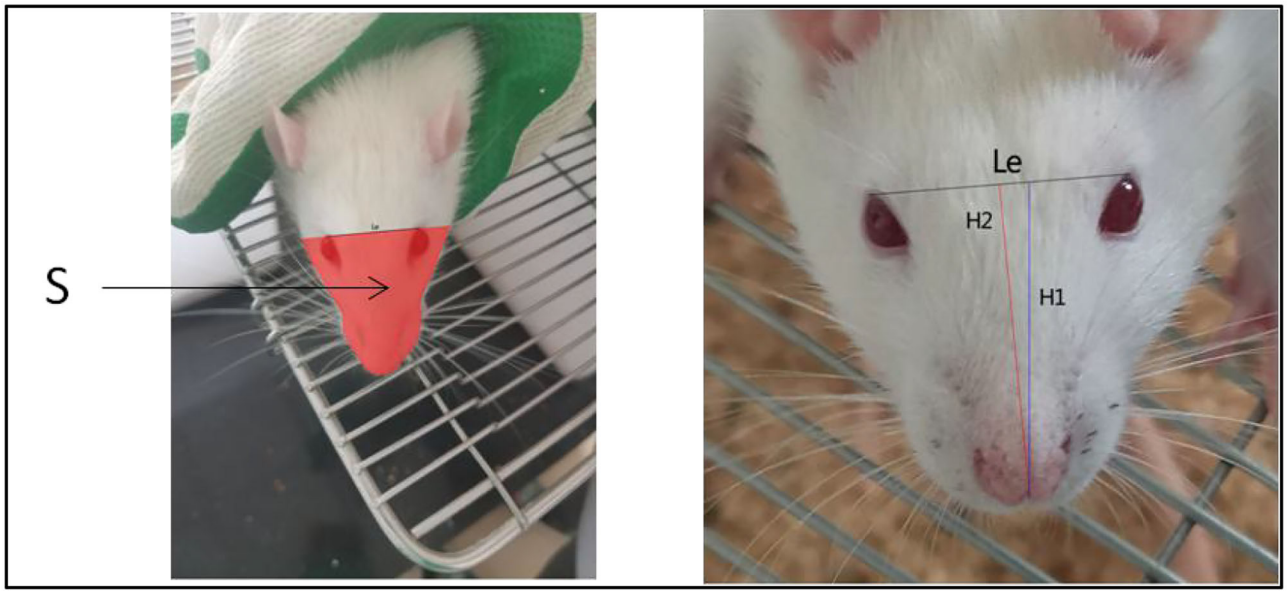

FIGURE 5 | Schematic of each parameter of rat face. Red denotes the area of the face recognized (S).

or decrease with respect to H1 and H2, while Le is assumed as constant.

As described above, the vertical length H1 from the tip of the nose to the line connecting the two eyes is indicated by the blue line, and the length H2 from the midpoint between the two eyes to the tip of the nose is indicated by the red line as shown in **Figure 5**.

Statistical Analysis

Statistical analysis was performed with SPSS (version 22.0). All data were expressed as mean ± SD. The comparisons between multiple groups were analyzed by one-way ANOVA, and group comparisons were analyzed using Student's *t* test. A $P < 0.05$ was considered statistically significant.

RESULTS

Calibration of the Distance Between the Lens and Target Object

As the distance between the camera and subject decreases, the subject's face area increases. Conversely, as the distance between the camera and object increases, the face area decreases. To ensure that the technique is adaptable for use in experiments, the operator did not fix the distance between the lens and target when capturing. Hence, Le, H1, and S decreased as the lens moved away from the target. Specifically, in this case, we considered the basic principles of digital and aesthetic anatomy wherein the interocular distance is determined by the brow bone and is fixed for the same types of rats. Therefore, we can determine the distance of the lens from the target object while taking a picture based on the size of the Le of the same type of rats. When the lens is turned away from the face of the rat, the distance Le between the outer canthi of the eyes decreases and point S decreases accordingly. Similarly, we can normalize Le of the same type of rats and use it to convert the area under the same Le to eliminate the changes in the absolute value of facial area due to the varying distance of the lens.

As described in **Figure 6**, we normalized the interocular Le of the same rats at different time points and converted the facial area of the rats as described above. The results indicated that there was no significant change in trend between the groups at different time points after calibration compared with that before calibration (**Figures 6A–D**). However, the absolute value of facial area changed, which is also consistent with our description above. After performing MCAO, the model group showed an increase in facial area from postoperative day 3 compared with the control group. Furthermore, a significant difference ($p < 0.05$) was observed on postoperative day 7 as the postoperative duration increased (**Figures 6E,F**).

The experimental results in **Figure 7** indicate that the S values of the different groups during the administration differ before and after Le calibration. On day 1 after dosing (day 8 after performing MCAO), the model group showed a significant increase in facial area compared with the normal group ($p < 0.05$). On day 5 after administration (day 12 after performing MCAO), the model and positive drug groups exhibited a significant increase in facial area compared with the normal group ($p < 0.001$). However, the active control group exhibited a lower facial area than the model group. From day 9 (day 16 after MCAO) to day 21 (day 28 after MCAO), the facial area of the model group was higher than that of the normal group. This indicates that bilateral common carotid artery ligation during MCAO can lead to facial swelling in rats. This can be relieved after the administration to the positive drug group.

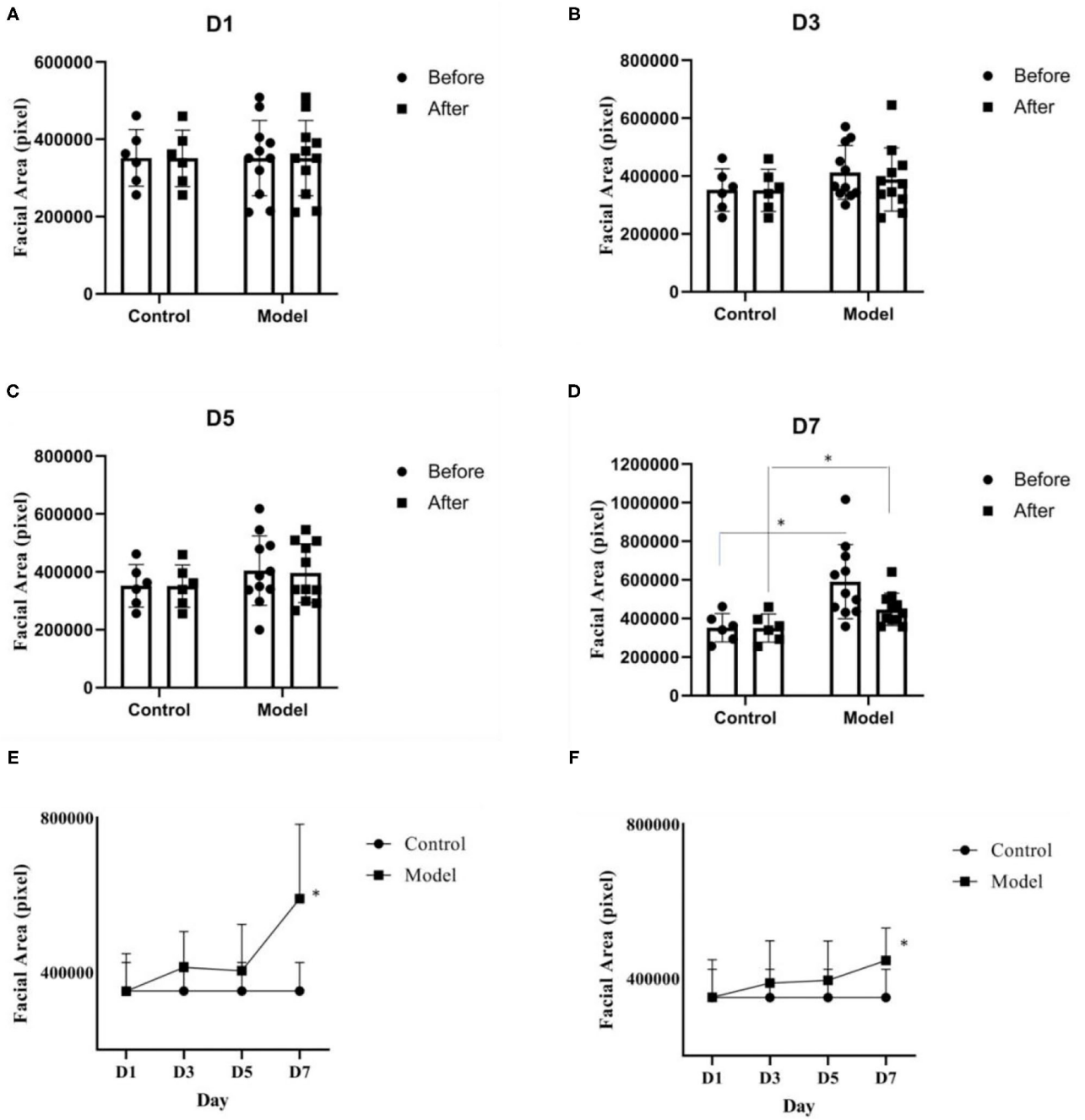

FIGURE 6 | Comparison of parameter S in each group before and after Le calibration in the postoperative to pre-dosing phase. Comparison of S values in each group before and after Le calibration: **(A)** on day 1 after surgery, **(B)** on day 3 after surgery, **(C)** on day 5 after surgery, **(D)** on day 7 after surgery. Curves of changes in S values at different time points in the two groups: **(E)** before calibration and **(F)** after calibration. Model group compared with control group, $^{*}p < 0.05$. control group: $n = 6$, model group: $n = 11$.

Calibration of Head Rotation

When grasping rats, head rotation often occurs. In this case, we considered the occurrence of head rotation when H1 < H2 according to **Figure 4**. **Figure 8** shows pictures of model group No. 4 rat acquired at different time points. The H1 value of the rat is very close its H2 value. The pictures in first line of **Figure 4** indicate that the lower part of the rats is more symmetrical at each time point. Furthermore, we considered that the No. 4 rat's face was facing the camera lens in this case.

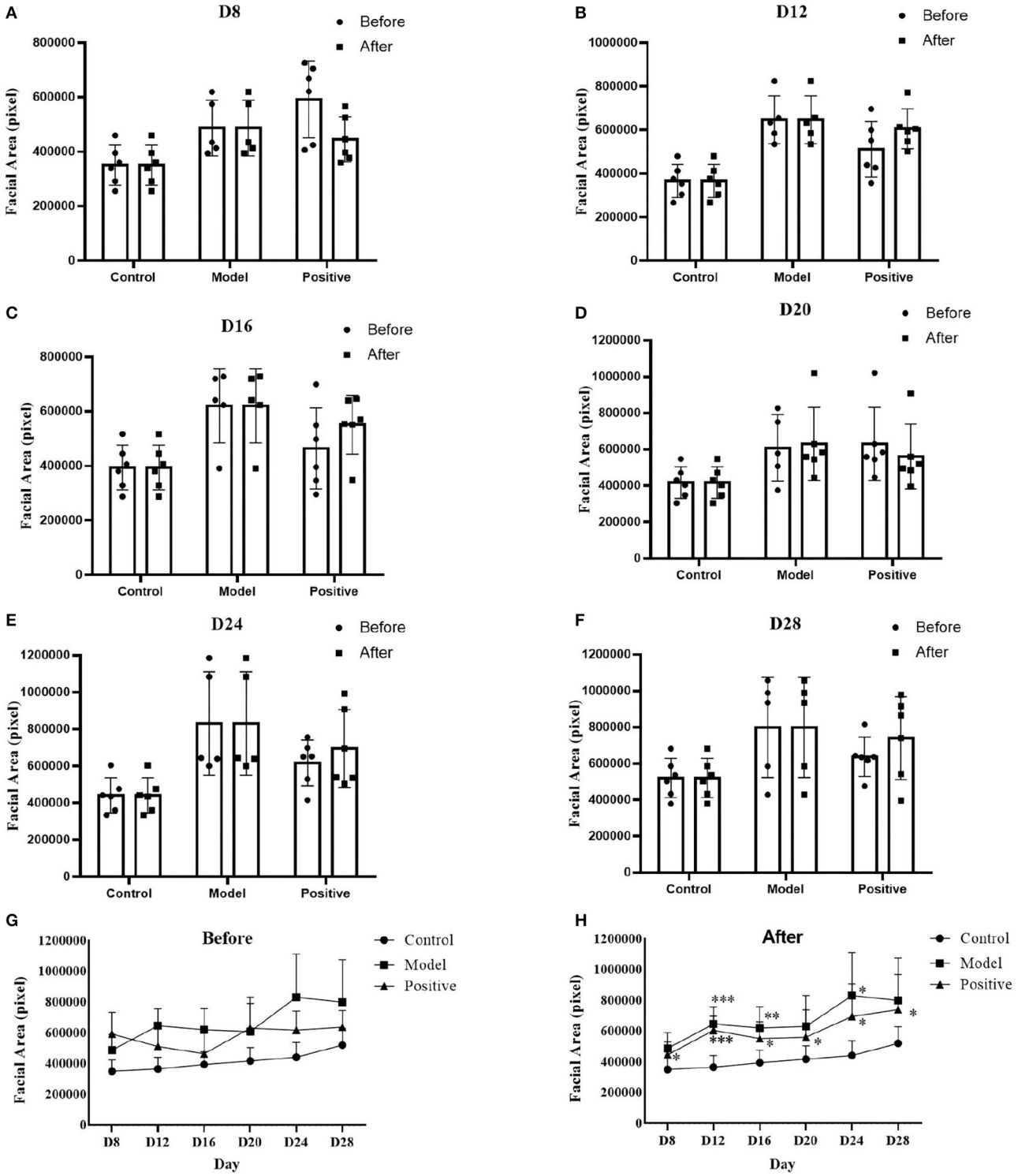

FIGURE 7 | Comparison of parameter S before and after Le calibration in each group after dosing to the end of the experiment. Comparison of S values before and after Le calibration: **(A)** on experiment day 8 (day 1 after dosing), **(B)** on experiment day 12 (day 5 after dosing), **(C)** on experiment day 16 (day 9 after dosing), **(D)** on experiment day 20 (day 13 after dosing), **(E)** on experiment day 24 (day 17 after dosing), **(F)** on experiment day 28 (day 21 after dosing); S value change curves of different groups, **(G)** before Le calibration, and **(H)** after Le calibration. Compared with the normal group, $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Model group: $n = 5$, control group: $n = 6$, positive drug group: $n = 6$.

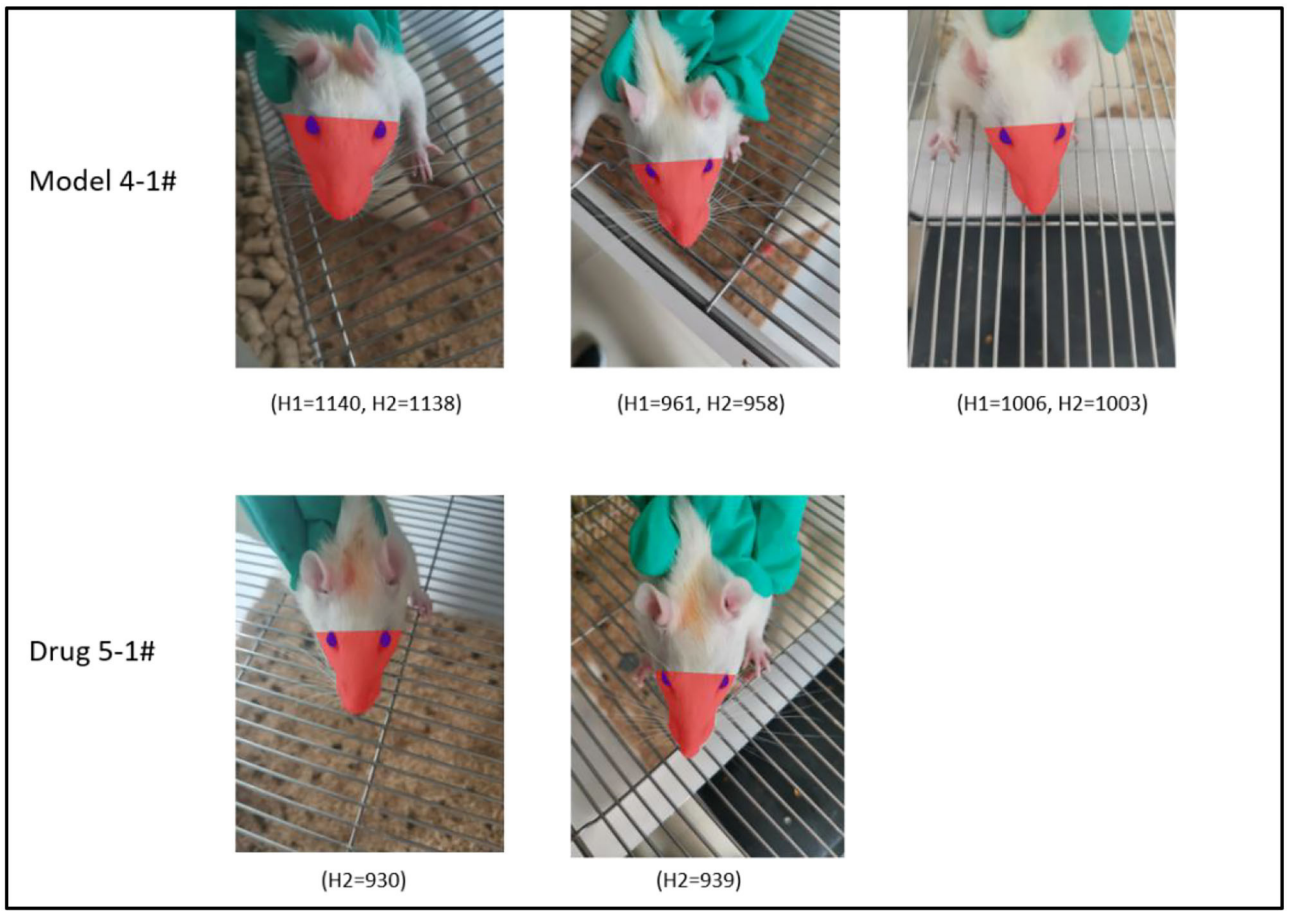

FIGURE 8 | Demonstration of the head angle calibration effect. Row 1 shows rat #4 in the model group, H1 and H2 are close at all time points, and the face of the rat faces the front in all cases. Row 2 shows rat #5 from the investigational drug group, H2 is close at various time points, and the head angle is approximate.

We measured H1 and H2 parameters for each rat in the model group after surgery until drug administration. By acquiring pictures at successive time points, we observed that the mean values of H1 and H2 did not differ significantly at each time point (**Figure 9A**). Therefore, we concluded that the calibration of H1 and H2 was weaker than that of Le described above. To further assess the effect of this parameter on facial recognition, we measured H1 and H2 from the model and active control groups at each time point between the administration of the drug and end of the experiment. A comparison of the results indicate that H1 and H2 were identical (**Figures 9B,C**).

Lifted Head vs. Lowered Head Calibration

The distance between the tip of the nose and line between the eyes decreased when the head was tilted upward. Conversely, the distance between the tip of the nose and line between the eyes increased when the head was tilted downward. During this time, the distance between the eyes remained constant (assuming the lens was at the same distance from the target) (**Figure 8**). As shown in **Figure 8**, rat #5 in the active control group exhibited similar H2 values at the two different time points, and the angle of head elevation was also very close.

We calibrated the facial area of each rat in the model group from the postoperative to the pre-dosing phase in the order of mirror depth (Le), head rotation (H1 vs. H2), and head lifting (H1). Specifically, first, Le was calibrated based on the uncalibrated facial area (S) to obtain the calibrated area (leftmost bar of each part of **Figure 10**). Then, S data were calibrated according to the relationship between H1 and H2, and each rat was calibrated such that the area of each rat corresponded to the area when it was facing the camera (middle bar of each figure in **Figure 10**). Finally, based on S calibrated in the second step, H1 of all rats was normalized to obtain the final calibrated facial area of rats (H1 in each part of **Figure 10**). The results indicate that the deviation of S for each rat in the group decreased as calibration was progressively performed, thereby indicating that the calibration led to further representation of the objective facial area of the rat.

DISCUSSION

In this study, during the process of MCAO model construction and subsequent drug evaluation, we unintentionally determined that the model rats exhibited facial swelling compared with the

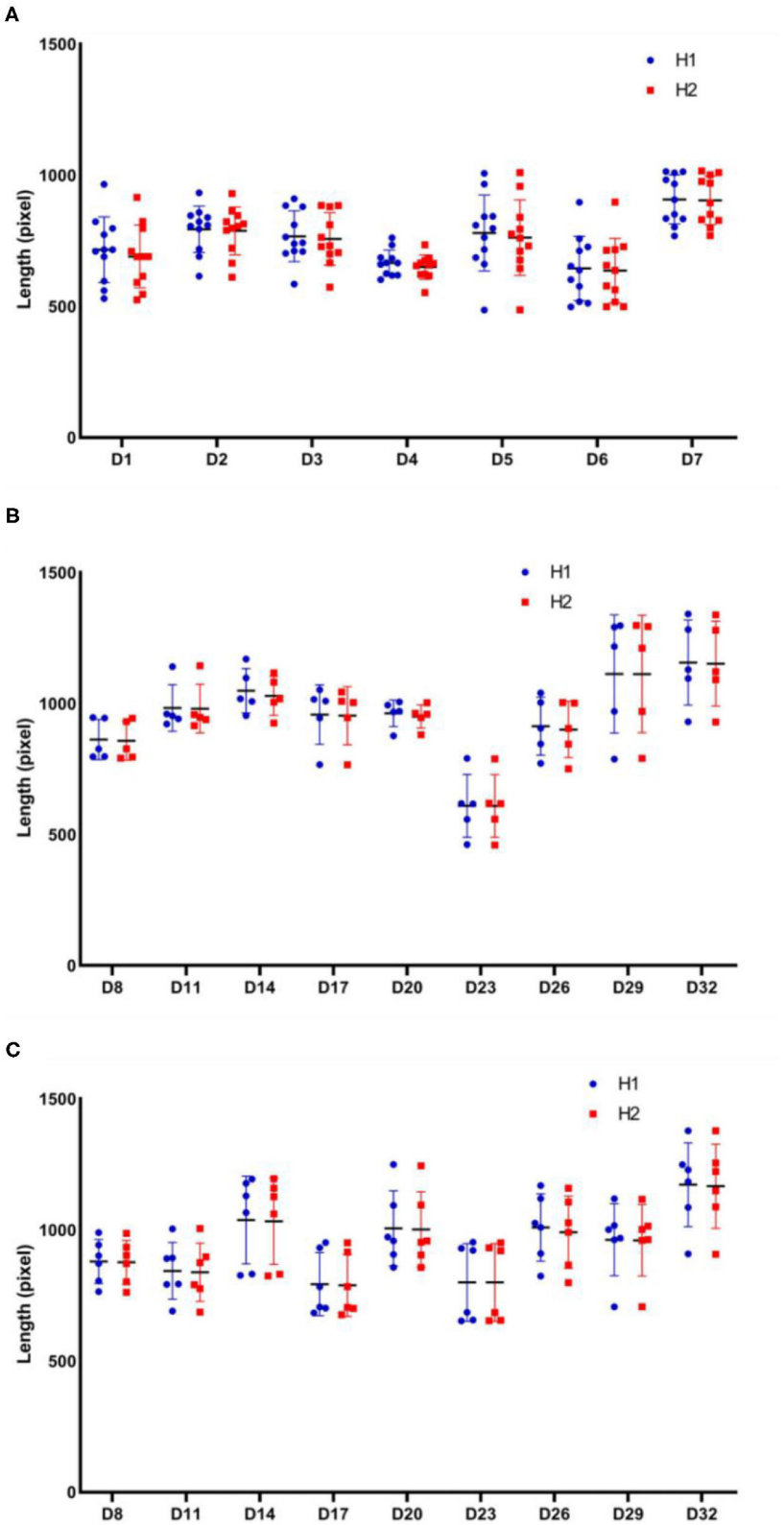

FIGURE 9 | Comparison of H1 and H2 parameters for each group at different stages of the experiment. H1 and H2 dynamics curves for each rat: **(A)** in the model group after surgery and before administration; **(B)** in the model group between the beginning of administration and end of the experiment; **(C)** in the active control group between the beginning of administration and end of the experiment. Model group: $n = 5$, positive drug group: $n = 6$.

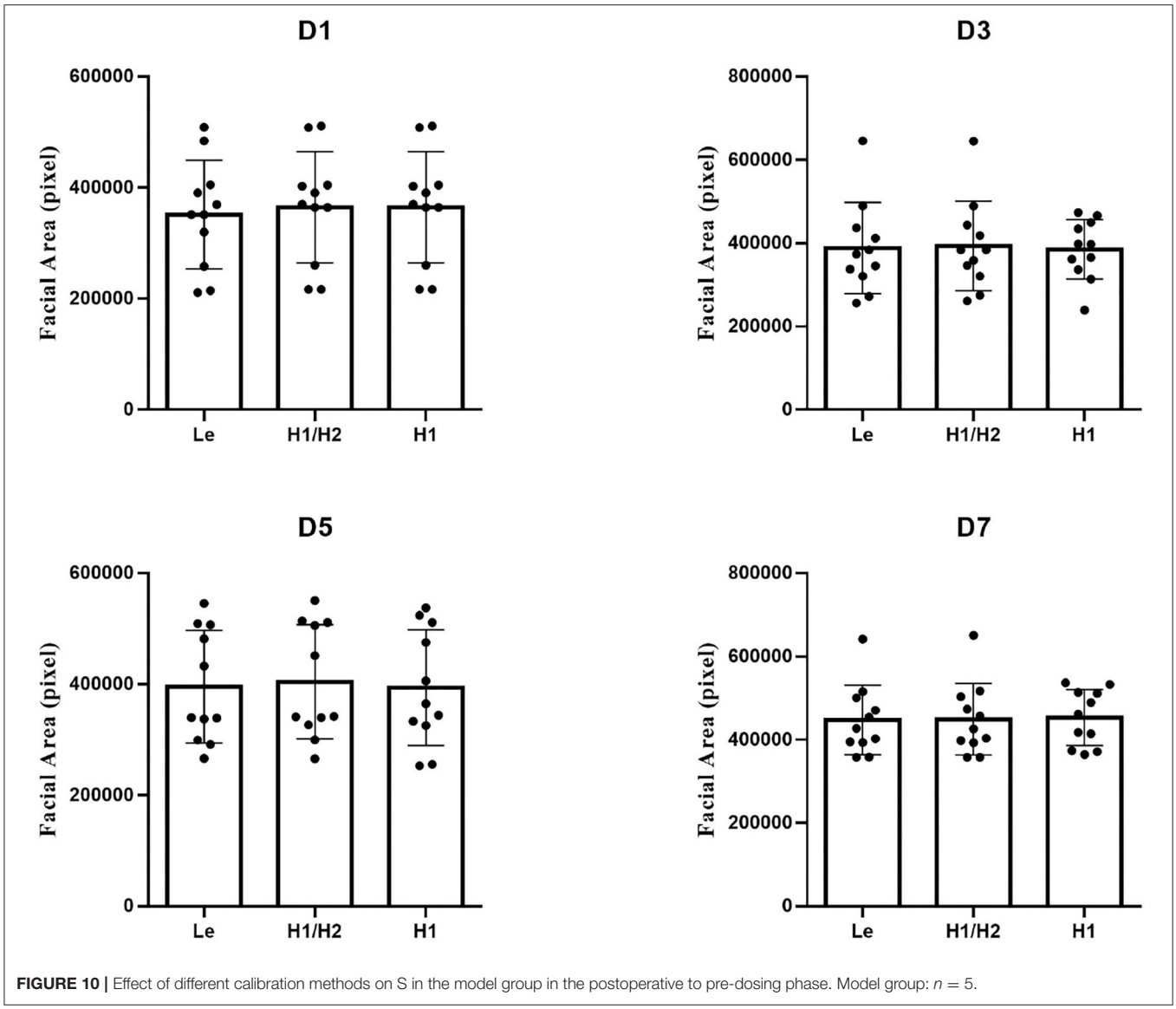

FIGURE 10 | Effect of different calibration methods on S in the model group in the postoperative to pre-dosing phase. Model group: $n = 5$.

control rats. To objectively evaluate the degree of facial swelling in rats, we used the most common photography method (mobile phone photography), which is flexible and can be performed at any time during the experiment. After image acquisition, we used Photoshop, common commercially available image processing software, for parameter measurements.

Anatomy is the basis of medicine and biology. With advances in technology, computer techniques have been increasingly applied to anatomy (7) such as digital pathology, image digitization, and three-dimensional scanning of the head. Furthermore, computer techniques are applied in plastic surgery and treatment of vascular diseases in the maxillofacial region (8). To enhance the representation of facial swelling features, we referred to anatomical and plastic surgery-related concepts to compute and analyze acquired images *via* custom parameter settings and calibration principles. Regarding the acquired images of the variable rat faces, we utilized the anatomy of the skull to capture the invariant brow bone of the same rats and used the distance between the eyes as one of the calibration methods to eliminate differences due to the distance between the lens and target. Simultaneously, we also adopted basic mathematical principles to determine the deviation due to the head tilting or head rotation of rats to perform a simple quantitative analysis. By performing a series of calibration analyses and comparisons, we concluded that the size of S was most closely related to the depth of the lens. Furthermore, as the distance between the rat and lens decreased, the interocular distance Le between the eyes increased and S increased at that time. Therefore, Le calibration is extremely critical in data analysis. Conversely, the calibration of H1 and H2 slightly affects the change in the size of S.

A literature search was conducted to address the biological explanation of the phenomenon of facial swelling after MCAO (9). However, to the best of our knowledge, there are no reports on the phenomenon of facial swelling in rats after MCAO.

Several studies reported that shoulder-hand syndrome occurs after stroke, with one of its typical features corresponding to hand swelling. It has also been reported that stroke patients tend to exhibit deep vein thrombosis (10), and one of its classical clinical signs involves swelling of the affected limb. Based on these results, we hypothesized that facial swelling can occur in patients with cerebral ischemia. After reviewing the literature and based on our previous research experience, we believe that the causes of facial swelling after cerebral ischemia are as follows: (1) a sharp decrease in cerebral blood flow due to bilateral common carotid artery ligation (11), which in turn increases intracranial pressure. Some clinical trials have shown that inadequate venous drainage triggered by bilateral radical neck dissection can cause intracranial hypertension, which leads to facial swelling. It is hypothesized that carotid artery ligation affects venous return. This in turn results in facial swelling. (2) Bilateral common carotid artery ligation can lead to cell swelling and tissue edema (12), and patients with hypoxic–ischemic brain damage are more likely to exhibit cerebral edema (13). Additionally, acute intracranial pressure elevation (14) can cause periventricular leukomalacia. (3) Clinically, the middle cerebral artery (MCA) trunk exhibits a higher chance of stenosis or even occlusion than the anterior and posterior cerebral arteries (15). This is mainly because the MCA trunk has a higher blood flow and is more prone to atherosclerotic plaques and mural thrombi. Hence, this results in luminal narrowing (16). Conversely, occlusion of the superior cortical branch of the MCA can lead to contralateral involvement and impaired circulation (17). (4) The craniofacial and temporal fascia contain rich blood supply (18), which is derived from the common carotid artery, superficial temporal artery, facial artery, and maxillary artery, which are accompanied by veins and intertwined into a network at the terminal branches of the internal carotid artery. Therefore, some patients with severe stenosis of the extracranial segment of the internal carotid artery (more than 70% stenosis) can be treated by mandibular carotid endarterectomy (19, 20). The swelling of the maxillofacial region, which is observed using contrast techniques, is associated with compensatory thickening of the facial arteries (21). (5) Swelling of the maxillofacial region is closely associated with the onset of inflammation. Furthermore, facial swelling is observed in chronic angioneurotic edema (22), which is mainly due to capillary dilation, congestion, and exudation in deep connective tissue, and it is accompanied by inflammatory cell infiltration (23). Conversely, the tissues of the eyelids, upper and lower lips, and cheeks are relatively loose and are easily observed when edema occurs.

The positive drug used in this study was donepezil hydrochloride tablets. They are routinely used in clinical practice and can reversibly inhibit acetylcholine hydrolysis by acetylcholinesterase, thereby increasing the concentration of acetylcholine and exerting therapeutic effects by enhancing the function of cholinergic nerves. During the 28-day period of control administration to MCAO-treated rats, we observed that the drug had some ameliorative effect on facial swelling in the model rats after surgery. However, its effector mechanism is not known.

In addition, we analyzed the ocular characteristics of rats in the acquired images, including the eye area and proportion of the face occupied by the eyes (data not shown in this paper) and observed that the eyes of rats can protrude early and atrophy later after MCAO. This is also related to the fact that the blood supply to the eyes mainly comes from the branches of the internal carotid artery and cases of exophthalmos in stroke patients have been reported. Bilateral common carotid artery ligation leads to an increase in intraocular pressure of the body, which results in protrusion and atrophy of the eyes.

In summary, in this study, we established a simple and easy method to significantly replace the existing subjective scoring methods for edema and provide new ideas for future applications based on the analysis of facial swelling in stroke patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The animal study was reviewed and approved by Xiyuan Hospital of China Academy of Chinese Medical Sciences.

AUTHOR CONTRIBUTIONS

YuL and HY contributed to the design of experiments, manuscript revision, and decisions involving submission for publication and co-corresponding authors. YaL and HH performed the experiments and prepared the manuscript. YiL, JC, and TT analyzed the data. All authors edited the manuscript and provided final approval for the final version for publication.

FUNDING

This study was supported by the Special Project for Outstanding Young Talents of China Academy of Chinese Medical Sciences (ZZ15-YQ-017 and ZZ13-YQ-001-A1).

REFERENCES

1. Kumar A, Misra S, Nair P, Algahtany M. Epigenetics mechanisms in ischemic stroke: a promising avenue? *J Stroke Cerebrovasc Dis.* (2021) 30:105690. doi: 10.1016/j.jstrokecerebrovasdis.2021.105690
2. Eltzschig HK, Eckle T. Ischemia and reperfusion–from mechanism to translation. *Nat Med.* (2011) 17:1391–401. doi: 10.1038/nm.2507
3. Stanzione R, Forte M, Cotugno M, Bianchi F, Marchitti S, Rubattu S. Role of DAMPs and of leukocytes infiltration in ischemic stroke: Insights from animal models and translation to the human disease. *Cell Mol Neurobiol.* (2020). doi: 10.1007/s10571-020-00966-4. [Epub ahead of print].

4. Toennies KD. *Guide to Medical Image Analysis: Methods and Algorithms, 2nd Edition*. Magdeburg: Springer. (2017).
5. Montgomery DC, Runger GC. *Applied Statistics and Probability for Engineers, 7th Edition.* Tempe, AZ: John Wiley and Sons (2019).
6. Sugimoto H. Donepezil hydrochloride: a treatment drug for Alzheimer's disease. *Chem Rec*. (2001)1:63–73. doi: 10.1002/1528-0691(2001)1:1<63::AID-TCR9>3.0.CO;2-J
7. Assouline SL, Meyer C, Weber E, Chatelain B, Barrabe A, Sigaux N, et al. How useful is intraoperative cone beam computed tomography in maxillofacial surgery? An overview of the current literature. *Int J Oral Maxillofac Surg.* (2021) 50:198–204. doi: 10.1016/j.ijom.2020.05.006
8. Ding M, Kang Y, Yuan Z, Shan X, Cai Z. Detection of facial landmarks by a convolutional neural network in patients with oral and maxillofacial disease. *Int J Oral Maxillofac Surg*. (2021) 50:1443–49. doi: 10.1016/j.ijom.2021.01.002
9. Revuelta JM, Zamarrón Á, Fortes J, Rodríguez-Boto G, Vaquero J, Gutiérrez-González R. Experimental rat model of chronic cerebral hypoperfusion-reperfusion mimicking normal perfusion pressure breakthrough phenomenon. *Neurocirugia*. (2020) 31:209–15. doi: 10.1016/j.neucir.2019.11.002
10. Liu XC, Chen XW, Li ZL, Wang SC, Chen C. Anatomical distribution of lower-extremity deep venous thrombosis in patients with acute stroke. *J Stroke Cerebrovasc Dis.* (2020) 29:104866. doi: 10.1016/j.jstrokecerebrovasdis.2020.104866
11. Song J, Cheon SY, Lee WT, Park KA, Lee JE. The effect of ASK1 on vascular permeability and edema formation in cerebral ischemia. *Brain Res.* (2015) 1595:143–55. doi: 10.1016/j.brainres.2014.11.024
12. Ji C, Yu X, Xu W, Lenahan C, Tu S, Shao A. The role of glymphatic system in the cerebral edema formation after ischemic stroke. *Exp Neurol.* (2021) 340:113685. doi: 10.1016/j.expneurol.2021.113685
13. Alquisiras-Burgos I, Ortiz-Plata A, Franco-Pérez J, Millán A, Aguilera P. Resveratrol reduces cerebral edema through inhibition of de novo SUR1 expression induced after focal ischemia. *Exp Neurol.* (2020) 330:113353. doi: 10.1016/j.expneurol.2020.113353
14. Jia SW, Liu XY, Wang SC, Wang YF. Vasopressin hypersecretion-associated brain edema formation in ischemic stroke: Underlying mechanisms. *J Stroke Cerebrovasc Dis.* (2016) 25:1289–300. doi: 10.1016/j.jstrokecerebrovasdis.2016.02.002
15. Raeiq A, Lee S, Knuckey N, Honeybul S. Phillips, TJ. Vertebrobasilar dolichoectasia causing obstructive hydrocephalus and cerebellar edema due to fourth ventricular obstruction. *Interdiscip Neurosurg.* (2021) 25:101135. doi: 10.1016/j.inat.2021.101135
16. Ozaki CK, Sobieszczyk PS, Ho KJ, McPhee JT, Gravereaux EC. Evidence-based carotid artery-based interventions for stroke risk reduction. *Curr Probl Surg.* (2014) 51:198–242. doi: 10.1067/j.cpsurg.2014.01.002
17. Han M, Kwon I, Ha J, Kim J, Cha MJ, Kim YD, et al. Collateral augmentation treatment with a combination of acetazolamide and head-down tilt in a rat ischemic stroke model. *J Clin Neurosci.* (2020) 73:252–8. doi: 10.1016/j.jocn.2020.01.079
18. Zhou M, Zhong A, Chen J, Sun Y, Wang Z, Xiong L, et al. Superficial muscular aponeurotic system-pedicled flaps for the reconstruction of facial defects: clinical application and anatomical basis. *J Plast Reconstr Aesthet Surg.* (2020) 73:1318–25. doi: 10.1016/j.bjps.2020.02.009
19. Yan Z, Wei J, Wu W, Yang X, Sun M, Wang W, et al. Embolization and sclerotherapy of maxillofacial arteriovenous malformations with the use of fibrin glue combined with pingyangmycin. *Oral Surg Oral Med Oral Pathol Oral Radiol.* (2020) 130:25–31. doi: 10.1016/j.oooo.2020.02.003
20. Treat-Jacobson DJ, Rich K, DeVeaux T, Fitzgerald K, Flood A, Gilpin V, et al. Society for vascular nursing clinical practice guideline (CPG) For Carotid Artery Stenting. *J Vasc Nurs.* (2013) 31:32–55. doi: 10.1016/j.jvn.2012.12.003
21. Pallagatti S, Sheikh S, Puri N, Mittal A, Singh B. To evaluate the efficacy of ultrasonography compared to clinical diagnosis, radiography and histopathological findings in the diagnosis of maxillofacial swellings. *Eur J Radiol.* (2012) 81:1821–7. doi: 10.1016/j.ejrad.2011.04.065
22. Uzun T. Management of patients with hereditary angio-oedema in dental, oral, and maxillofacial surgery: a review. *by J Oral Maxillofac Surg.* (2019) 57:992–7. doi: 10.1016/j.bjoms.2019.09.008
23. Wienholtz NKF, Christensen CE, Coskun H, Zhang DG, Ghanizada H, Egeberg A, et al. Infusion of pituitary adenylate cyclase-activating polypeptide-38 in patients with rosacea induces flushing and facial edema that can be attenuated by sumatriptan. *J Invest Dermatol*. (2021) 141:1687–98. doi: 10.1016/j.jid.2021.02.002

Conflict of Interest: HH was employed by the company Beijing Duan-Dian Pharmaceutical Research & Development Co., Ltd., Beijing, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 01 April 2022
doi: 10.3389/fmed.2022.807382

Development of a Machine Learning-Based Predictive Model for Lung Metastasis in Patients With Ewing Sarcoma

Wenle Li[1,2†], Tao Hong[3†], Wencai Liu[4†], Shengtao Dong[5], Haosheng Wang[6], Zhi-Ri Tang[7], Wanying Li[2], Bing Wang[2], Zhaohui Hu[8], Qiang Liu[1], Yong Qin[9*] and Chengliang Yin[10*]*

[1] Department of Orthopedics, Xianyang Central Hospital, Xianyang, China, [2] Clinical Medical Research Center, Xianyang Central Hospital, Xianyang, China, [3] Department of Cardiac Surgery, Fuwai Hospital Chinese Academy of Medical Sciences, Shenzhen, Shenzhen, China, [4] Department of Orthopaedic Surgery, the First Affiliated Hospital of Nanchang University, Nanchang, China, [5] Department of Spine Surgery, Second Affiliated Hospital of Dalian Medical University, Dalian, China, [6] Department of Orthopaedics, The Second Hospital of Jilin University, Changchun, China, [7] School of Physics and Technology, Wuhan University, Wuhan, China, [8] Department of Spinal Surgery, Liuzhou People's Hospital, Liuzhou, China, [9] Department of Orthopedics Surgery, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [10] Faculty of Medicine, Macau University of Science and Technology, Macau, Macau SAR, China

Background: This study aimed to develop and validate machine learning (ML)-based prediction models for lung metastasis (LM) in patients with Ewing sarcoma (ES), and to deploy the best model as an open access web tool.

Methods: We retrospectively analyzed data from the Surveillance Epidemiology and End Results (SEER) Database from 2010 to 2016 and from four medical institutions to develop and validate predictive models for LM in patients with ES. Patient data from the SEER database was used as the training group (n = 929). Using demographic and clinicopathologic variables six ML-based models for predicting LM were developed, and internally validated using 10-fold cross validation. All ML-based models were subsequently externally validated using multiple data from four medical institutions (the validation group, n = 51). The predictive power of the models was evaluated by the area under receiver operating characteristic curve (AUC). The best-performing model was used to produce an online tool for use by clinicians to identify ES patients at risk from lung metastasis, to improve decision making and optimize individual treatment.

Results: The study cohort consisted of 929 patients from the SEER database and 51 patients from multiple medical centers, a total of 980 ES patients. Of these, 175 (18.8%) had lung metastasis. Multivariate logistic regression analysis was performed with survival time, T-stage, N-stage, surgery, and bone metastasis providing the independent predictive factors of LM. The AUC value of six predictive models ranged from 0.585 to 0.705. The Random Forest (RF) model (AUC = 0.705) using 4 variables was identified as the best predictive model of LM in ES patients and was employed to construct an online tool to assist clinicians in optimizing patient treatment. (https://share.streamlit.io/liuwencai123/es_lm/main/es_lm.py).

OPEN ACCESS

Edited by:
Chris Hodge,
The University of Sydney, Australia

Reviewed by:
Steven Christopher Smith,
Virginia Commonwealth University Health System, United States
Kun Liu,
Sir Run Run Shaw Hospital, China
Bing Yang,
Tianjin Medical University, China

***Correspondence:**
Chengliang Yin
chengliangyin@163.com
Yong Qin
qinyong0125@126.com
Qiang Liu
m13992079668@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *02 November 2021*
Accepted: *07 March 2022*
Published: *01 April 2022*

Citation:
Li W, Hong T, Liu W, Dong S, Wang H, Tang Z-R, Li W, Wang B, Hu Z, Liu Q, Qin Y and Yin C (2022) Development of a Machine Learning-Based Predictive Model for Lung Metastasis in Patients With Ewing Sarcoma. Front. Med. 9:807382. doi: 10.3389/fmed.2022.807382

Conclusions: Machine learning were found to have utility for predicting LM in patients with Ewing sarcoma, and the RF model gave the best performance. The accessibility of the predictive model as a web-based tool offers clear opportunities for improving the personalized treatment of patients with ES.

Keywords: Ewing sarcoma, lung metastasis, machine learning algorithms, multicenter, web calculator

INTRODUCTION

Ewing sarcoma (ES) is an aggressive sarcoma with a high propensity for local recurrence and distant metastasis in children and adolescents (1, 2). ES is the second most common primary bone malignancy, accounting for 5% of all child and adolescent cancers (3). ES frequently involves the diaphysis region of long bones (4). Despite the development of new treatment regimens, ES has a high likelihood of tumor metastasis, leading to a worsening prognosis and resulting in a poor 5-year survival rate of only 20–45% (4, 5). In a retrospective study of 975 patients with ES, 5-year survival and 5-year relapse-free survival rates for patients with localized disease were 70 and 55%, respectively, but only 33 and 21% for those with distant metastasis disease (6).

Although diagnostic imaging techniques have improved dramatically during the past 30 years, metastatic status can only be detected in approximately 20–25% of ES patients

TABLE 1 | Baseline of patients with SEER database and multicenter data.

	level	Overall (*N* = 980)	Multicenter (validation group, *N* = 51)	SEER (training group, *N* = 929)	p
Race (%)	Black	39 (4.0)	0 (0.0)	39 (4.2)	<0.001
	Other	126 (12.9)	51 (100.0)	75 (8.1)	
	White	815 (83.2)	0 (0.0)	815 (87.7)	
Age [mean (SD)]	NA	22.39 (16.45)	24.96 (18.97)	22.25 (16.30)	0.252
Sex (%)	Female	418 (42.7)	23 (45.1)	395 (42.5)	0.828
	Male	562 (57.3)	28 (54.9)	534 (57.5)	
Primary. Site (%)	Axis bone	431 (44.0)	27 (52.9)	404 (43.5)	0.394
	Limb bone	317 (32.3)	13 (25.5)	304 (32.7)	
	other	232 (23.7)	11 (21.6)	221 (23.8)	
Laterality (%)	left	374 (38.2)	21 (41.2)	353 (38.0)	0.894
	Not a paired site	296 (30.2)	15 (29.4)	281 (30.2)	
	right	310 (31.6)	15 (29.4)	295 (31.8)	
T (%)	T1	351 (35.8)	20 (39.2)	331 (35.6)	0.008
	T2	429 (43.8)	25 (49.0)	404 (43.5)	
	T3	39 (4.0)	5 (9.8)	34 (3.7)	
	TX	161 (16.4)	1 (2.0)	160 (17.2)	
N (%)	N0	841 (85.8)	44 (86.3)	797 (85.8)	0.312
	N1	80 (8.2)	6 (11.8)	74 (8.0)	
	NX	59 (6.0)	1 (2.0)	58 (6.2)	
M (%)	M0	662 (67.6)	30 (58.8)	632 (68.0)	0.225
	M1	318 (32.4)	21 (41.2)	297 (32.0)	
surgery (%)	No	413 (42.1)	25 (49.0)	388 (41.8)	0.381
	Yes	567 (57.9)	26 (51.0)	541 (58.2)	
Radiation (%)	No	757 (77.2)	29 (56.9)	728 (78.4)	0.001
	Yes	223 (22.8)	22 (43.1)	201 (21.6)	
Chemotherapy (%)	No/Unknown	58 (5.9)	0 (0.0)	58 (6.2)	0.125
	Yes	922 (94.1)	51 (100.0)	871 (93.8)	
Bone.metastases (%)	No	831 (84.8)	40 (78.4)	791 (85.1)	0.271
	Yes	149 (15.2)	11 (21.6)	138 (14.9)	
Lung.metastases (%)	No	795 (81.1)	41 (80.4)	754 (81.2)	1
	Yes	185 (18.9)	10 (19.6)	175 (18.8)	
times [mean (SD)]	NA	30.56 (22.65)	29.71 (22.40)	30.61 (22.67)	0.782

TABLE 2 | Baseline table of patients in the Ewing sarcoma lung metastasis group vs. the no lung metastasis group.

	Level	Overall (*N* = 929)	No (*N* = 754)	Yes (*N* = 175)	*p*
Race (%)	Black	39 (4.2)	27 (3.6)	12 (6.9)	0.105
	Other	75 (8.1)	64 (8.5)	11 (6.3)	
	White	815 (87.7)	663 (87.9)	152 (86.9)	
Age [mean (SD)]	NA	22.25 (16.30)	22.10 (16.35)	22.88 (16.10)	0.569
Sex (%)	Female	395 (42.5)	329 (43.6)	66 (37.7)	0.18
	Male	534 (57.5)	425 (56.4)	109 (62.3)	
Primary.Site (%)	Axis bone	404 (43.5)	316 (41.9)	88 (50.3)	0.13
	Limb bone	304 (32.7)	253 (33.6)	51 (29.1)	
	other	221 (23.8)	185 (24.5)	36 (20.6)	
Race (%)	Black	39 (4.2)	27 (3.6)	12 (6.9)	0.105
	Other	75 (8.1)	64 (8.5)	11 (6.3)	
	White	815 (87.7)	663 (87.9)	152 (86.9)	
T (%)	T1	331 (35.6)	304 (40.3)	27 (15.4)	<0.001
	T2	404 (43.5)	312 (41.4)	92 (52.6)	
	T3	34 (3.7)	20 (2.7)	14 (8.0)	
	TX	160 (17.2)	118 (15.6)	42 (24.0)	
N (%)	N0	797 (85.8)	676 (89.7)	121 (69.1)	<0.001
	N1	74 (8.0)	37 (4.9)	37 (21.1)	
	NX	58 (6.2)	41 (5.4)	17 (9.7)	
M (%)	M0	632 (68.0)	632 (83.8)	0 (0.0)	<0.001
	M1	297 (32.0)	122 (16.2)	175 (100.0)	
surgery (%)	No	388 (41.8)	271 (35.9)	117 (66.9)	<0.001
	Yes	541 (58.2)	483 (64.1)	58 (33.1)	
Radiation (%)	No	728 (78.4)	593 (78.6)	135 (77.1)	0.739
	Yes	201 (21.6)	161 (21.4)	40 (22.9)	
Chemotherapy (%)	No/Unknown	58 (6.2)	45 (6.0)	13 (7.4)	0.585
	Yes	871 (93.8)	709 (94.0)	162 (92.6)	
Bone.metastases (%)	No	791 (85.1)	672 (89.1)	119 (68.0)	<0.001
	Yes	138 (14.9)	82 (10.9)	56 (32.0)	
times [mean (SD)]	NA	30.61 (22.67)	32.40 (22.83)	22.89 (20.31)	<0.001

(3), with the lung being the most common metastatic site (5, 7, 8). Computed tomography (CT) scans of the chest are usually carried out to detect lung metastasis. However, given the high cost, radiation damage, and low efficiency of detection of metastatic nodules, new strategies are urgently required to accurately predict the development of lung metastasis in patients with ES (9, 10).

Machine learning (ML) has emerged as a powerful computer-based method of data mining and analysis and has been extensively applied as a "prediction tool" in a multitude of different scientific, engineering, and medical scenarios (11–15). ML has been shown to detect more interactions between variables, and to be more accurate than conventional statistical methods (14, 16). ML algorithms have been applied to model clinical outcome and to improve cognition of tumor growth and progression (17). However, although numerous ML-based predictive models of tumor development have been reported, no study has been conducted in predicting lung metastasis associated with Ewing Sarcoma.

The Surveillance Epidemiology and End Results (SEER) database contains data for around 26% of the United States population and is commonly used to study rare diseases since it overcomes the obstacle of inadequate case numbers (18–20). We constructed several ML-based models of LM in patients with ES, using the SEER database. External validation was subsequently performed using data from multiple medical centers to predict the probability of LM with the aim of improving individualized patient management. The best model was uploaded as a web-based tool.

MATERIALS AND METHODS

Study Population and Data Selection

Data were sourced from the SEER database and four medical institutions in China: Liuzhou People's Hospital, Second Affiliated Hospital of Jilin University, Xianyang Central Hospital, and Second Affiliated Hospital of Dalian Medical University, respectively. This

TABLE 3 | Univariate and multifactorial logistic regression analysis of risk factors for lung metastasis in patients with Ewing sarcoma.

Variables	Univariate OR (95% CI)	*p* value	Multivariate OR (95% CI)	*p* value
Age (years)	1.000 (0.991–1.010)	0.968	/	/
Survival time (month)	0.980 (0.973–0.988)	<0.001	0.988 (0.979–0.997)	<0.01
Race				
White	Ref	Ref	Ref	Ref
Black	1.939 (0.960–3.914)	0.065	/	/
Other	0.872 (0.529–1.439)	0.593	/	/
Sex				
Male	Ref	Ref	Ref	Ref
Female	0.804 (0.579–1.116)	0.192	/	/
Primary site				
Limb bones	Ref	Ref	Ref	Ref
Axis of a bone	1.359 (0.937–1.970)	0.106	/	/
other	0.924 (0.585–1.460)	0.735	/	/
Laterality				
Left	Ref	Ref	Ref	Ref
Right	1.148 (0.784–1.681)	0.479	/	/
Other	1.004 (0.676–1.491)	0.984	/	/
T				
T1	Ref	Ref	Ref	Ref
T2	3.461 (2.214–5.410)	<0.001	2.701 (1.690–4.317)	<0.001
T3	8.025 (3.8074–16.917)	<0.001	4.037 (1.773–9.194)	<0.01
TX	4.071 (2.415–6.864)	<0.001	3.146 (1.778–5.566)	<0.001
N				
N0	Ref	Ref	Ref	Ref
N1	5.570 (0.3457–8.975)	<0.001	5.102 (3.048–8.540)	<0.001
NX	2.255 (1.245–4.084)	<0.01	1.411 (0.734–2.715)	0.302
Surgery				
No	Ref	Ref	Ref	Ref
Yes	0.278 (0.196–0.394)	<0.001	0.451 (0.309–0.658)	<0.001
Radiation				
No	Ref	Ref	Ref	Ref
Yes	1.241 (0.858–1.795)	0.251	/	/
Chemotherapy				
No	Ref	Ref	Ref	Ref
Yes	0.794 (0.419–1.504)	0.479	/	/
Bone metastases				
No	Ref	Ref	Ref	Ref
Yes	3.403 (2.326–4.977)	<0.001	1.685 (1.090–2.605)	<0.05

retrospective study did not use personal identifying information and thus did not require informed patient consent or Institutional Ethics Committee Board approval.

Patients selected from the SEER database (2010–2016) who were diagnosed with ES originating in bone, as identified by ICD-O-3/WHO 2008 morphology code 9260d, composed the "training" group. Criteria for exclusion were more than one primary tumor and incomplete clinicopathological information. The "validation" group was composed of ES patient data obtained from four hospitals in different regions of China, from 2010 to 2018. All cases featured complete clinicopathological data and follow-up information and no other primary tumors. Demographic and clinicopathological variables included in both groups were: race, age, sex, primary site, laterality, T-stage, N-stage, M-stage, surgery, radiation, chemotherapy, bone metastasis, and survival times. For consistency with SEER database records, "race" in the Chinese medical records was classified as "other". Detailed treatments, such as surgery, radiation, and chemotherapy were classified as Yes or No, and were not recorded in the SEER database.

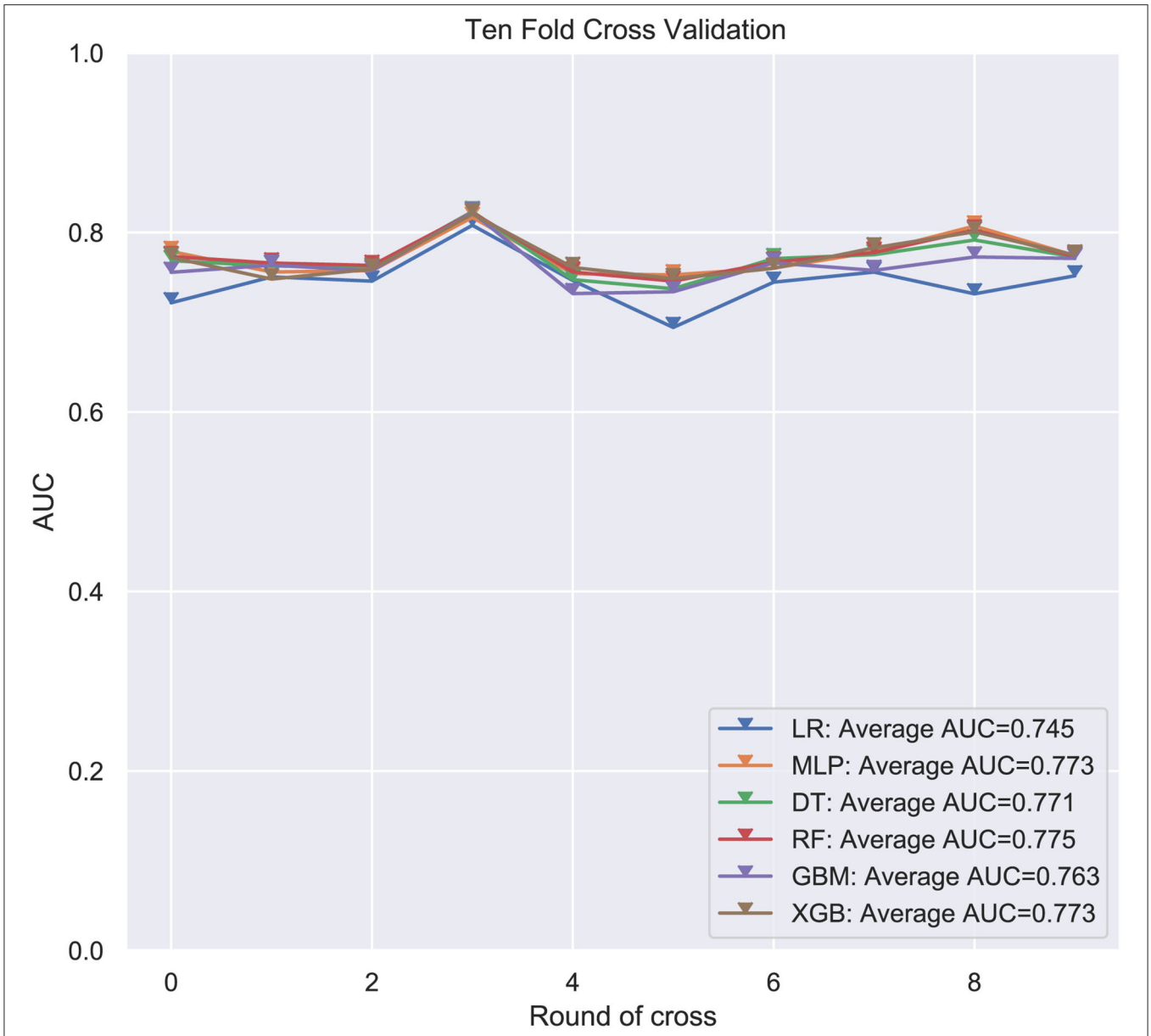

FIGURE 1 | Average area under the curve (AUC) values of 10-fold cross-validation. RF, Random forest predictive model; DT, Decision tree; XGB, Extreme gradient boosting; GBM, Gradient boosting machine; MLP, Multilayer perceptron; LR, Logistic regression; AUC used as an indicator of performance, RF model achieved the best predictive performance while the MLP model showed the lowest.

Establishment and Evaluation of Prediction Models

Using demographic and clinicopathological data, we explored the effect of variables ($p < 0.05$) in univariate analysis, in the multifactorial regression model, and in predictive models based on the ML algorithms. Six different ML algorithms were applied independently to develop predictive models of LM in patients with ES, as follows: Random Forest (RF), Logistic regression (LR), Extreme gradient boosting (XGB), Gradient boosting machine (GBM), Multilayer perceptron (MLP), and Decision tree (DT) (21, 22). For the training process of the ML algorithms using python (version 3.8), we employed 10-fold cross-validation to avoid overfitting (23). We also calculated the average value of the area under receiver operating characteristic curve (AUC) to evaluate the predictive power of each model.

The ML algorithms were subsequently applied to the validation group and the AUC was again calculated to evaluate the predictive performance of all models. The higher the AUC value, the better the model. Finally, the best-performing model was designed as a web-based tool for predicting the likelihood of LM in ES patients.

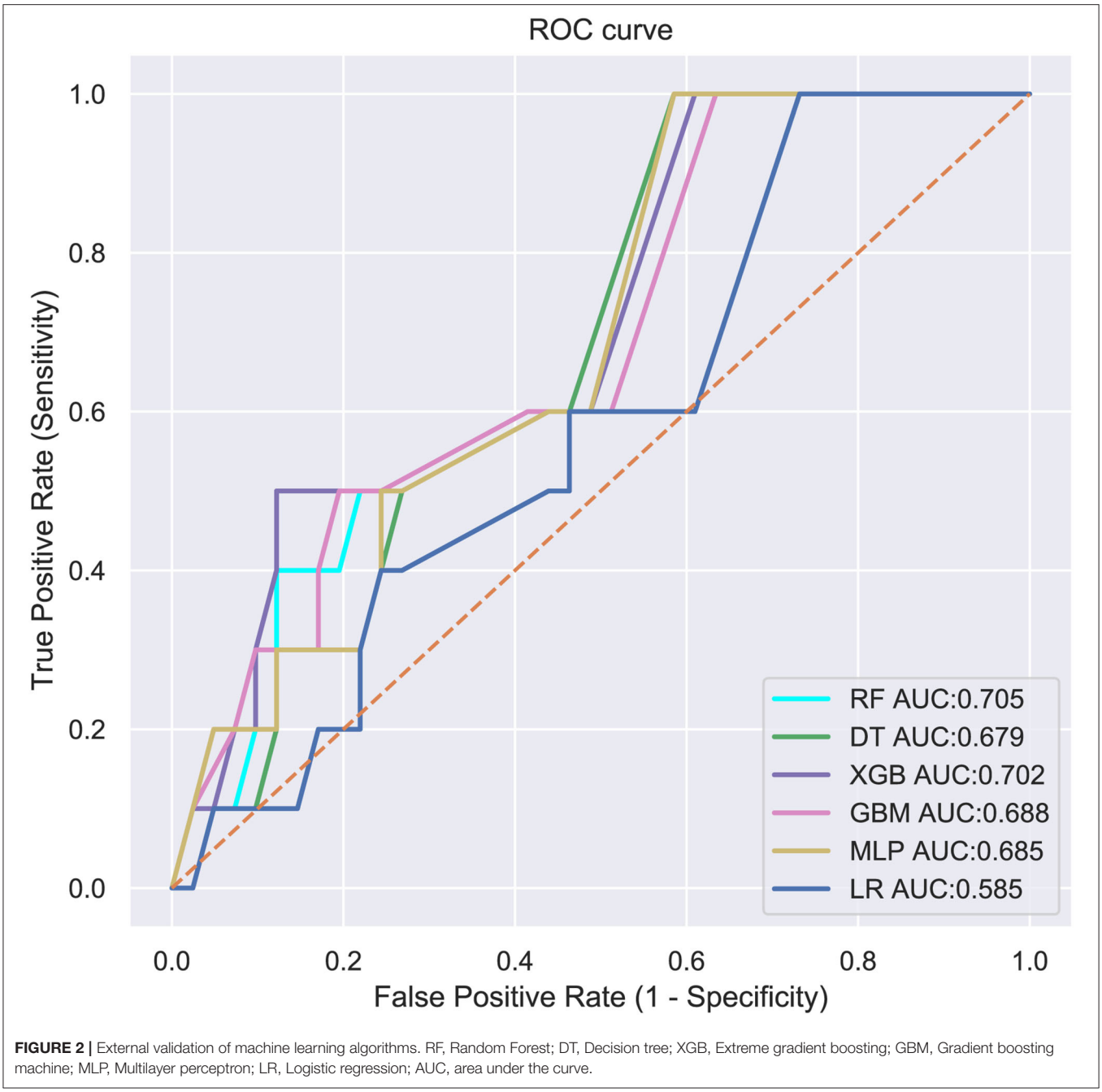

FIGURE 2 | External validation of machine learning algorithms. RF, Random Forest; DT, Decision tree; XGB, Extreme gradient boosting; GBM, Gradient boosting machine; MLP, Multilayer perceptron; LR, Logistic regression; AUC, area under the curve.

As a model inspection technique, permutation feature importance can be used for any fitted estimator (24–26). Thus, a total of 100 independent training simulation results were applied to assess the most important variables in each predictive model using permutation feature importance analysis. We further assessed the relative contribution of four key clinical variables to LM predictive models using spearman correlation of features analysis and plotted a correlation heat map.

Statistical Analysis

All data were extracted from the SEER database *via* the SEER * Stat software (version 8.3.6). All analyses were performed using python (version 3.8). The baseline variables between the training group and validation group were compared using Student's t tests and Pearson chi-square test. A two-sided $p < 0.05$ was deemed to have statistical significance.

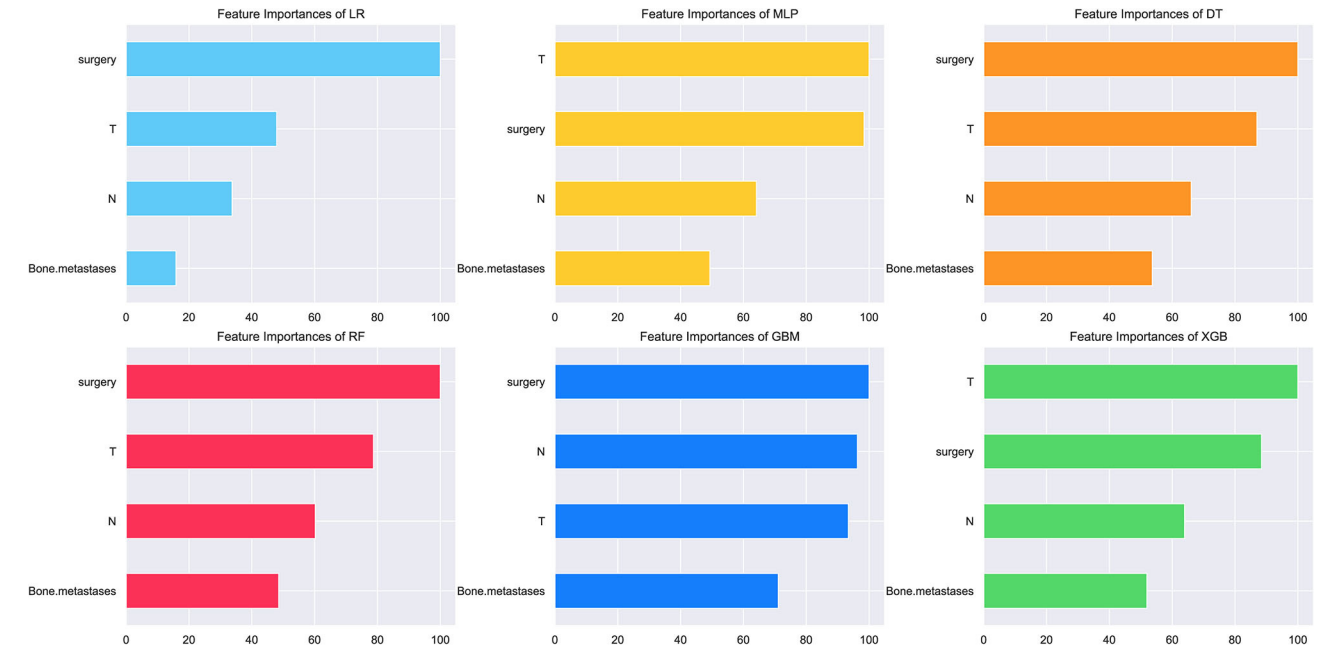

FIGURE 3 | The relative importance of variables for the prediction of LM using ML algorithms. Surgery, T-stage and N-stage ranked in the top three in all prediction models, with bone metastasis ranked fourth.

RESULTS

Baseline Characteristics

A total of 980 patients with ES were enrolled in this study; 929 patients originating from the SEER database were assigned to the training group; and 51 patients from four medical centers in China were assigned to the validation group (**Table 1**). There were significant differences between the two groups in terms of race, T-stage, and radiation ($p < 0.05$). In the validation group, all patients were classified under race as "others". The proportion of radiation was significantly higher in the validation group than in the training group. In addition, more patients were diagnosed as TX in the training group. The remaining variables were not significantly different in both groups (**Table 1**). Lung metastasis occurred in 185 (18.9%) cases, the median age of the patients was 22.25 years ($SD = 16.3$), more than 85% of the patients were Caucasian and 534 (57.5%) patients were male. Comparison of the baseline data between the lung metastasis group and no lung metastasis group, revealed significant differences for the following factors: T-stage, N-stage, M-stage, surgery, bone metastasis, and survival time ($p < 0.001$). The demographic and clinicopathological variables of all 980 patients are summarized in **Table 2**.

Univariate and Multifactorial LR Analysis of LM

The following variables were shown to have significant correlation with the development of LM in univariate analysis ($p < 0.05$): survival time, T-stage, N-stage, surgery, and bone metastasis ($p < 0.001$) (**Table 3**). Multifactorial LR analysis based on the variables ($p < 0.05$) in univariate analysis, demonstrated that T- stage (T2, $OR = 2.7018$, 95% $CI = 1.690–4.317$; T3, $OR = 4.0378$, 95% $CI = 1.773–9.194$; TX, $OR = 3.1468$, 95% $CI = 1.778–5.566$), N1 stage [*vs.* N0 stage, N1, ($OR = 5.102$, 95% $CI = 3.048–8.540$)], and bone metastasis ($OR = 1.685$, 95% $CI = 1.090–2.605$) were independent negative predictors of LM while survival time ($OR = 0.988$, 95% $CI = 0.979–0.997$) and surgery ($OR = 0.451$, 95% $CI = 0.309–0.658$) were positive predictors.

Predictive Performance of Machine Learning (ML) Algorithms

Six ML-based models for predicting LM in ES patients were developed based on the training group data. The average AUC of the six models determined by 10-fold cross-validation is shown in **Figure 1**, with the RF model achieving the best performance ($AUC = 0.775$). When the models established in training were subjected to external validation (**Figure 2**), the RF model still achieved the best performance ($AUC = 0.705$) in predicting LM and was accordingly selected as the design for a web-based, predictive tool.

Influence of Variables on Prediction Performance

In consideration of clinical utility (**Figure 3**), we focused on four variables (T-stage, N-stage, surgery, and bone metastasis) to construct ML-based predictive models for LM in ES patients. Although there were slight differences in the importance of variables identified by each model; three factors, such as surgery, T-stage and N-stage, consistently ranked in the top three, and bone metastasis ranked fourth. The relative importance of

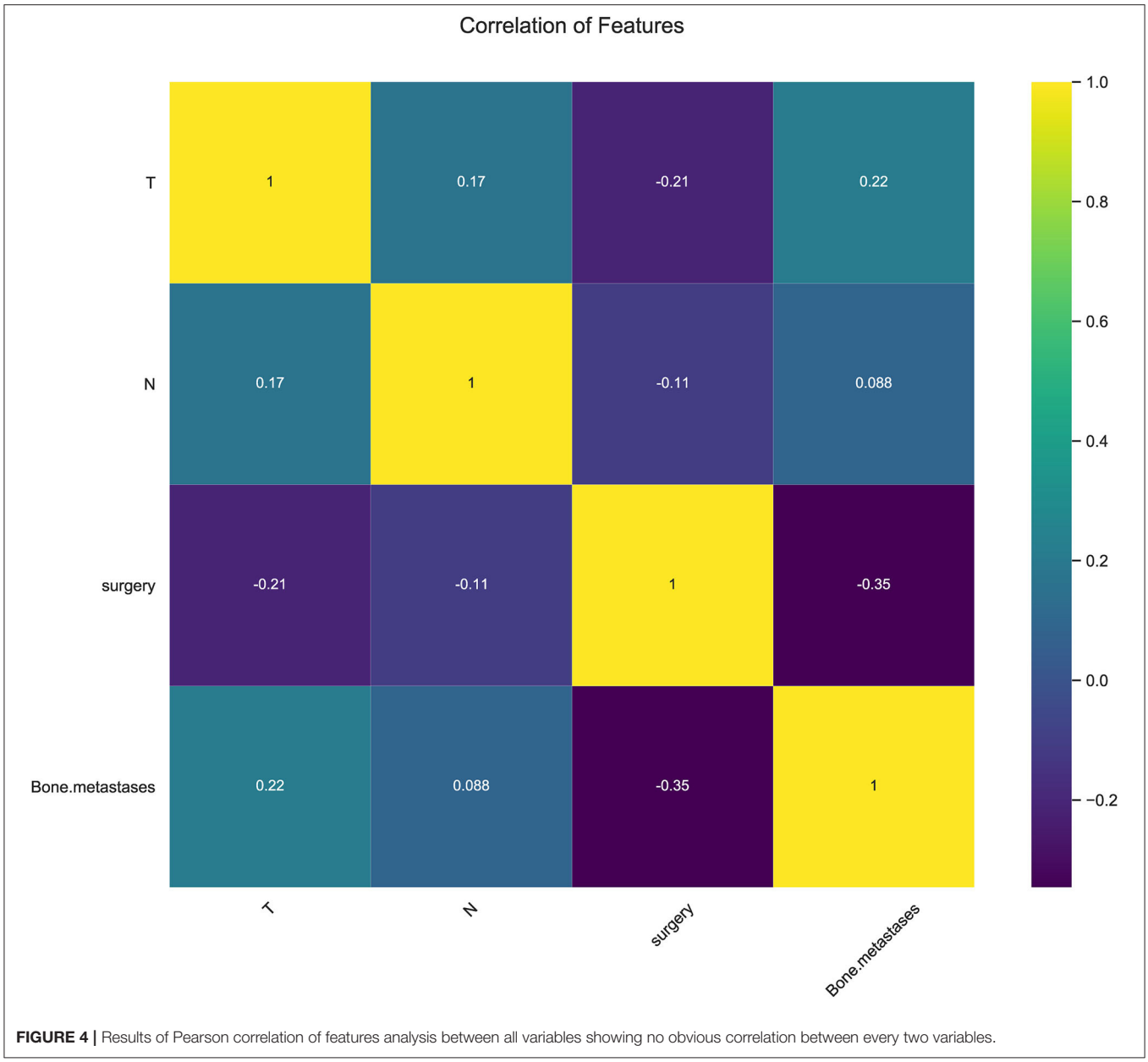

FIGURE 4 | Results of Pearson correlation of features analysis between all variables showing no obvious correlation between every two variables.

variables in predicting LM using the RF model decreased in the order: surgery > T-stage > N-stage > bone metastasis. Analysis using spearman correlation of features approach revealed no significant positive correlation between any variable, and a negative correlation between surgery and the other three variables, indicating that all variables were independent (**Figure 4**).

Design of a Web-Based Tool for Predicting LM in ES Patients

The best-performing RF model was used to design a web-based tool to assist clinicians in predicting lung metastasis in ES patients (https://share.streamlit.io/liuwencai123/es_lm/main/es_lm.py) (**Figure 5**).

DISCUSSION

Multi-modal therapy of metastatic disease based on chemotherapy, surgery, and radiation would be improved dramatically by the availability of reliable methods for predicting metastasis (27, 28). Many mathematical models of tumor malignancy employ multivariate regression or correlation analysis, which usually require the variables to be independent and linear (29–32). In addition to traditional univariate and multivariate analysis, we used multiple ML algorithms, which

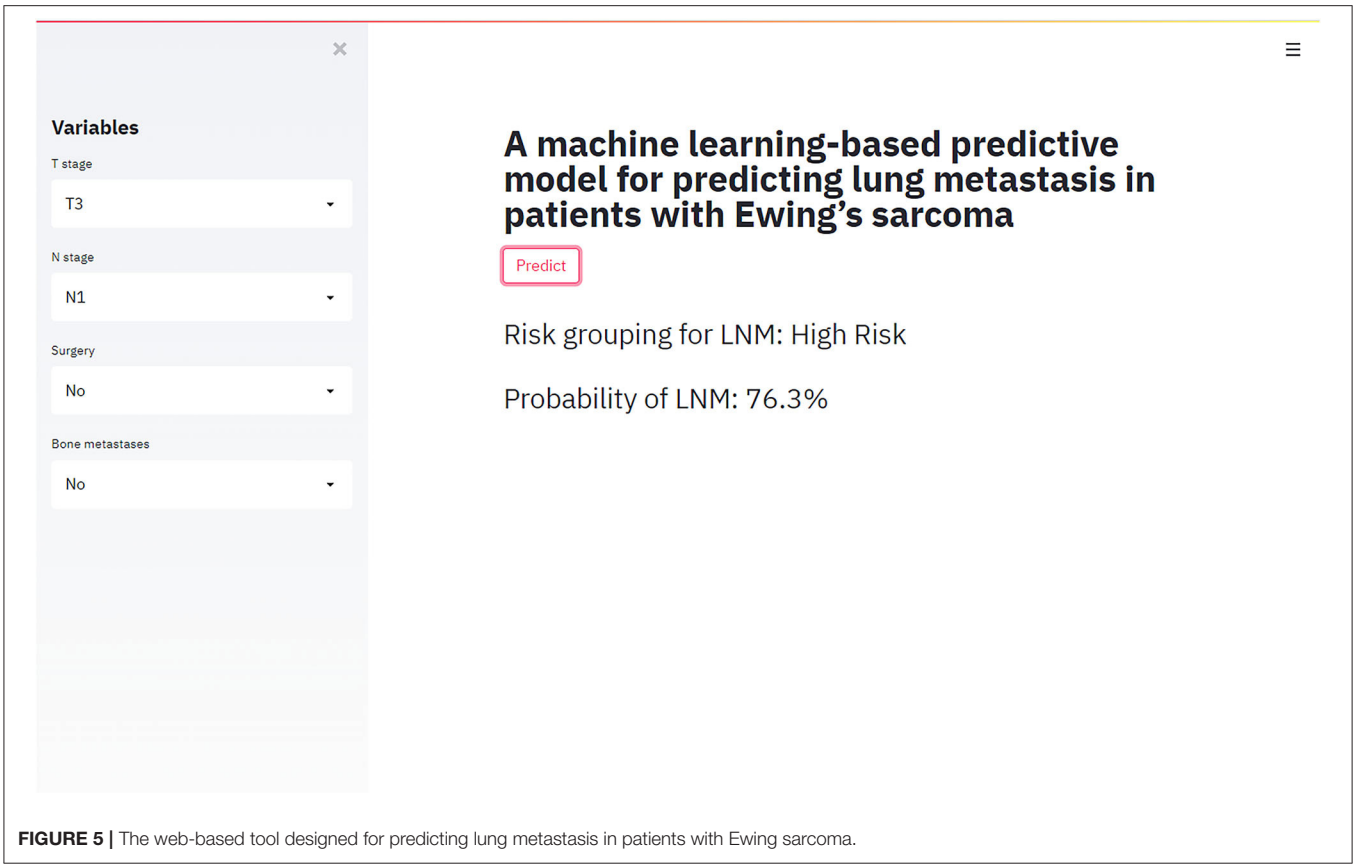

FIGURE 5 | The web-based tool designed for predicting lung metastasis in patients with Ewing sarcoma.

are widely applied in healthcare data analysis, to construct predictive models of LM in ES patients. We found that the RF model provided the best performance. RF is a commonly used ML algorithm that has a proven track record in handling large complex nonlinear datasets (33, 34). We subsequently designed a rapid web-based clinical tool, which is based on the RF model, for predicting lung metastasis in patients with ES.

Patient survival time was positively related to LM in univariate analysis. However, when considering clinical practice, survival time has no meaning for patients initially diagnosed with ES, and it is difficult to assess the survival time of a part of the patient population. Thus, survival time was not considered as a variable in ML models.

In the present study, four clinical variables: surgery, T-stage, N-stage, and bone metastasis were found to be the most important factors for predicting LM status by ML algorithms. We identified surgery as a protective factor against LM. To our knowledge, this factor has not been included previously in LM risk prediction models. Surgery is not only a vital form of treatment, but also plays a significant diagnostic role, which enables more accurate TNM staging and prognosis of ES patients. Surgery ranked first in order of importance in most of the predictive models developed in the present study, while T-stage (tumor size) ranked in the top two in all models investigated and was highly predictive of LM, similar to previous reports (35, 36). Large tumor volume indicates a longer growth cycle, resulting in a more proliferative and aggressive state, thus increasing the occurrence of lung metastasis. The correlation heat map showed that the T-stage correlated negatively with surgery since radical surgical treatment is difficult for large tumors, and lung metastasis is more likely.

Extensive investigations have consistently demonstrated that patients with regional node involvement were more prone to develop distant metastasis (37–41). Since the lung is associated with an abundance of lymphatic vessels, a tumor is more likely to metastasize to the lung when lymph nodes are positive. However, due to the scarcity of lymphatic vessels in bone tumor, it is conventionally accepted that dissemination to lymph nodes is uncommon (4, 42). Applebaum et al., for example, found that only 6.3% (91/1,452) of cases featured lymph node involvement (37). In contrast, our study revealed a much higher rate of lymph node metastasis, approximately 18.9% (185/980).

Importantly, our ML-based models revealed that bone metastasis was an important predictor of LM in ES patients, ranking fourth in importance behind surgery, T-stage and N-stage variables. Of the 138 patients in the two combined cohorts (training group and validation group) who had bone metastasis, 40.6% (56/138) also displayed lung metastasis. This figure was significantly higher than the number of patients who showed LM without bone metastasis (15%, 119/791).

Our present study of ML-based models for predicting LM in ES patients contained certain limitations which, nonetheless, serve as a guide for future improvements. Firstly, the information accessed from the SEER database was to a certain degree limited. Clinical information, such as the precise surgical treatment, surgical margin status, tumor marker, vascular invasion, radiation dosage, and chemotherapy modalities were unavailable, which limits the predictive value of the developed models. Secondly, the data from the SEER database was retrospective, which may introduce bias in data selection. However, while cognizant of these limitations, our study affirmed that ML-based prediction models can effectively identify the likelihood of LM in patients with ES by inspection of clinical factors such as surgery, N-stage, T-stage, and bone metastasis. The RF model performed best according to ROC analysis and was subsequently used to produce a web-based tool designed to help clinicians identify ES patients with lung metastasis, improve decision making and optimize individual treatment. Increased case data and multicenter studies are anticipated to lead to improvements in predictive performance.

CONCLUSION

Machine learning algorithms were applied to develop a prognostic tool for predicting the risk of LM in patients with ES. A RF model performed best and was engineered as a web-based tool for use by clinicians to improve patient diagnosis and treatment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

CY, QL, and YQ designed the study. WL and TH collected and evaluated the data and wrote the first draft of the manuscript. All authors contributed to the interpretation of the results and the final draft of the manuscript.

FUNDING

This study was supported by the National Clinical Research Center for Orthopedics, Sports Medicine and Rehabilitation, and the Jiangsu China-Israel Industrial Technical Research Institute Foundation, 2021-NCRC-CXJJ-ZH-11.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to EditSprings (https://www.editsprings.cn/) for providing linguistic services.

REFERENCES

1. Khan S, Abid Z, Haider G, Bukhari N, Zehra D, Hashmi M, et al. Incidence of Ewing's sarcoma in different age groups, their associated features, and its correlation with primary care interval. *Cureus.* (2021) 13:e13986. doi: 10.7759/cureus.13986
2. Yu H, Ge Y, Guo L, Huang L. Potential approaches to the treatment of Ewing's sarcoma. *Oncotarget.* (2017) 8:5523–39. doi: 10.18632/oncotarget.12566
3. Balamuth NJ. Womer RB. Ewing's sarcoma, *Lancet Oncol.* (2010) 11:184–92. doi: 10.1016/S1470-2045(09)70286-4
4. Shi J, Yang J, Ma X, Wang X. Risk factors for metastasis and poor prognosis of Ewing sarcoma: a population based study. *J Orthop Surg Res.* (2020) 15:88. doi: 10.1186/s13018-020-01607-8
5. Gaspar N, Hawkins DS, Dirksen U, Lewis IJ, Ferrari S, Le Deley MC, et al. Ewing sarcoma: current management and future approaches through collaboration. *J Clin Oncol.* (2015) 33:3036–46. doi: 10.1200/JCO.2014.59.5256
6. Cotterill SJ, Ahrens S, Paulussen M, Jürgens HF, Voûte PA, Gadner H, et al. Prognostic factors in Ewing's tumor of bone: analysis of 975 patients from the European Intergroup Cooperative Ewing's sarcoma study group. *J Clin Oncol.* (2000) 18:3108–14. doi: 10.1200/JCO.2000.18.17.3108
7. Esiashvili N, Goodman M, Marcus RB. Jr. Changes in incidence and survival of Ewing sarcoma patients over the past 3 decades: surveillance Epidemiology and End Results data. *J Pediatr Hematol Oncol.* (2008) 30:425–30. doi: 10.1097/MPH.0b013e31816e22f3
8. Arpaci E, Yetisyigit T, Seker M, Uncu D, Uyeturk U, Oksuzoglu B, et al. Prognostic factors and clinical outcome of patients with Ewing's sarcoma family of tumors in adults: multicentric study of the Anatolian Society of Medical Oncology. *Med Oncol.* (2013) 30:469. doi: 10.1007/s12032-013-0469-z
9. Völker T, Denecke T, Steffen I, Misch D, Schönberger S, Plotkin M. et al. Positron emission tomography for staging of pediatric sarcoma patients: results of a prospective multicenter trial. *J Clin Oncol.* (2007) 25:5435–41. doi: 10.1200/JCO.2007.12.2473
10. Mikulić D, Ilić I, Cepulić M, Giljević JS, Orlić D, Zupancić B, et al. Angiogenesis and Ewing sarcoma–relationship to pulmonary metastasis and survival. *J Pediatr Surg.* (2006) 41:524–9. doi: 10.1016/j.jpedsurg.2005.11.058
11. Mo X, Chen X, Ieong C, Zhang S, Li H, Li J, et al. Early prediction of clinical response to etanercept treatment in juvenile idiopathic arthritis using machine learning. *Front Pharmacol.* (2020) 11:1164. doi: 10.3389/fphar.2020.01164
12. Jin S, Kostka K, Posada JD, Kim Y, Seo SI, Lee DY, et al. Prediction of major depressive disorder following beta-blocker therapy in patients with cardiovascular diseases. *J Pers Med*. (2020) 10. doi: 10.3390/jpm10040288
13. Vey J, Kapsner LA, Fuchs M, Unberath P, Veronesi G, Kunz M. A toolbox for functional analysis and the systematic identification of diagnostic and prognostic gene expression signatures combining meta-analysis and machine learning. *Cancers (Basel).* (2019) 11. doi: 10.3390/cancers11101606
14. Stumpo V, Staartjes VE, Esposito G, Serra C, Regli L, Olivi A, et al. Machine learning and intracranial aneurysms: from detection to outcome prediction. *Acta Neurochir Suppl.* (2022) 134:319–31. doi: 10.1007/978-3-030-85292-4_36
15. Zilcha-Mano S, Roose SP, Brown PJ, Rutherford BR. A machine learning approach to identifying placebo responders in late-life depression trials. *Am J Geriatr Psychiatry.* (2018) 26:669–77. doi: 10.1016/j.jagp.2018.01.001
16. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol.* (2017) 2:204–9. doi: 10.1001/jamacardio.2016.3956
17. Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. *Cancers (Basel).* (2020) 12. doi: 10.3390/cancers12030603
18. Doll KM, Rademaker A, Sosa JA. Practical guide to surgical data sets: surveillance, epidemiology, and end results (SEER) database. *JAMA Surg.* (2018) 153:588–9. doi: 10.1001/jamasurg.2018.0501

19. Mao W, Deng F, Wang D, Gao L, Shi X. Treatment of advanced gallbladder cancer: a SEER-based study. *Cancer Med.* (2020) 9:141–50. doi: 10.1002/cam4.2679
20. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The surveillance, epidemiology, and end results (seer) program and pathology: toward strengthening the critical relationship. *Am J Surg Pathol.* (2016) 40:e94-94e102. doi: 10.1097/PAS.0000000000000749
21. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief Bioinform.* (2016) 17:33–42. doi: 10.1093/bib/bbv087
22. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery, *Lancet Oncol.* (2019) 20:e262-262e273. doi: 10.1016/S1470-2045(19)30149-4
23. Sturgiss EA, Rieger E, Haesler E, Ridd MJ, Douglas K. Galvin SL. Adaption and validation of the Working Alliance Inventory for General Practice: qualitative review and cross-sectional surveys. *Fam Pract.* (2019) 36:516–22. doi: 10.1093/fampra/cmy113
24. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics.* (2010) 26:1340–7. doi: 10.1093/bioinformatics/btq134
25. Mi X, Zou B, Zou F, Hu J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat Commun.* (2021) 12:3008. doi: 10.1038/s41467-021-22756-2
26. Yang JB, Shen KQ, Ong CJ. Li XP. Feature selection for MLP neural network: the use of random permutation of probabilistic outputs. *IEEE Trans Neural Netw.* (2009) 20:1911–22. doi: 10.1109/TNN.2009.2032543
27. Kibiş EY. Büyüktahtakin IE. Optimizing multi-modal cancer treatment under 3D spatio-temporal tumor growth. *Math Biosci.* (2019) 307:53–69. doi: 10.1016/j.mbs.2018.10.010
28. Liu Z, Zhong Y, Zhou X, Huang X, Zhou J, Huang D, et al. Inherently nitric oxide containing polymersomes remotely regulated by NIR for improving multi-modal therapy on drug resistant cancer. *Biomaterials.* (2021) 277:121118. doi: 10.1016/j.biomaterials.2021.121118
29. Pearce O, Delaine-Smith RM, Maniati E, Nichols S, Wang J, Böhm S, et al. Deconstruction of a metastatic tumor microenvironment reveals a common matrix response in human cancers. *Cancer Discov.* (2018) 8:304–19. doi: 10.1158/2159-8290.CD-17-0284
30. Arefan D, Hausler RM, Sumkin JH, Sun M, Wu S. Predicting cell invasion in breast tumor microenvironment from radiological imaging phenotypes. *BMC Cancer.* (2021) 21:370. doi: 10.1186/s12885-021-08122-x
31. Liu Z, Mi M, Li X, Zheng X, Wu G, Zhang L. A lncRNA prognostic signature associated with immune infiltration and tumour mutation burden in breast cancer. *J Cell Mol Med.* (2020) 24:12444–56. doi: 10.1111/jcmm.15762
32. Madekivi V, Boström P, Karlsson A, Aaltonen R, Salminen E. Can a machine-learning model improve the prediction of nodal stage after a positive sentinel lymph node biopsy in breast cancer. *Acta Oncol.* (2020) 59:689–95. doi: 10.1080/0284186X.2020.1736332
33. Lebedev AV, Westman E, Van Westen GJ, Kramberger MG, Lundervold A, Aarsland D. et al. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin.* (2014) 6:115–25. doi: 10.1016/j.nicl.2014.08.023
34. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl.* (2019) 134:93–101. doi: 10.1016/j.eswa.2019.05.028
35. Ye C, Dai M, Zhang B. Risk factors for metastasis at initial diagnosis with ewing sarcoma. *Front Oncol.* (2019) 9:1043. doi: 10.3389/fonc.2019.01043
36. Ramkumar DB, Ramkumar N, Miller BJ. Henderson ER. Risk factors for detectable metastatic disease at presentation in Ewing sarcoma - An analysis of the SEER registry. *Cancer Epidemiol.* (2018) 57:134–9. doi: 10.1016/j.canep.2018.10.013
37. Applebaum MA, Goldsby R, Neuhaus J. DuBois SG. Clinical features and outcomes in patients with Ewing sarcoma and regional lymph node involvement. *Pediatr Blood Cancer.* (2012) 59:617–20. doi: 10.1002/pbc.24053
38. van der Kamp MF, Muntinghe F, Iepsma RS, Plaat B, van der Laan B, Algassab A, et al. Predictors for distant metastasis in head and neck cancer, with emphasis on age. *Eur Arch Otorhinolaryngol.* (2021) 278:181–90. doi: 10.1007/s00405-020-06118-0
39. Javidiparsijani S, Brickman A, Lin DM, Rohra P, Ghai R, Bitterman P, et al. Is regional lymph node metastasis of head and neck paraganglioma a sign of aggressive clinical behavior: a clinical/pathologic review. *Ear Nose Throat J.* (2021) 100:447–53. doi: 10.1177/0145561319863373
40. Chu PY, Chen YF Li CY, Yang JS, King YA, Chiu YJ, et al. Factors influencing locoregional recurrence and distant metastasis in Asian patients with cutaneous melanoma after surgery: a retrospective analysis in a tertiary hospital in Taiwan. *J Chin Med Assoc.* (2021) 84:870–6. doi: 10.1097/JCMA.0000000000000586
41. Kilic C, Kimyon Comert G, Cakir C, Yuksel D, Codal B, Kilic F, et al. Recurrence pattern and prognostic factors for survival in cervical cancer with lymph node metastasis. *J Obstet Gynaecol Res.* (2021) 47:2175–84. doi: 10.1111/jog.14762
42. Edwards JR, Williams K, Kindblom LG, Meis-Kindblom JM, Hogendoorn PC, Hughes D, et al. Lymphatics and bone. *Hum Pathol.* (2008) 39:49–55. doi: 10.1016/j.humpath.2007.04.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Big Data

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fdata.2022.801998

When Patients Recover From COVID-19: Data-Driven Insights From Wearable Technologies

Muzhe Guo[1†], *Long Nguyen*[2†], *Hongfei Du*[1] *and Fang Jin*[1*]

[1] *Department of Statistics, The George Washington University, Washington, DC, United States,* [2] *Department of Computer Science and Data Science, School of Applied Computational Sciences, Meharry Medical College, Nashville, TN, United States*

Coronavirus disease 2019 (COVID-19) is known as a contagious disease and caused an overwhelming of hospital resources worldwide. Therefore, deciding on hospitalizing COVID-19 patients or quarantining them at home becomes a crucial solution to manage an extremely big number of patients in a short time. This paper proposes a model which combines Long-short Term Memory (LSTM) and Deep Neural Network (DNN) to early and accurately classify disease stages of the patients to address the problem at a low cost. In this model, the LSTM component will exploit temporal features while the DNN component extracts attributed features to enhance the model's classification performance. Our experimental results demonstrate that the proposed model achieves substantially better prediction accuracy than existing state-of-art methods. Moreover, we explore the importance of different vital indicators to help patients and doctors identify the critical factors at different COVID-19 stages. Finally, we create case studies demonstrating the differences between severe and mild patients and show the signs of recovery from COVID-19 disease by extracting shape patterns based on temporal features of patients. In summary, by identifying the disease stages, this research will help patients understand their current disease situation. Furthermore, it will also help doctors to provide patients with an immediate treatment plan remotely that addresses their specific disease stages, thus optimizing their usage of limited medical resources.

Keywords: COVID-19, wearable data, neural networks, uncertainty quantification, pattern extraction

OPEN ACCESS

Edited by:
Raghvendra Mall, St. Jude Children's Research Hospital, United States

Reviewed by:
Aniruddha Adiga, University of Virginia, United States
Oluwafemi A. Sarumi, Federal University of Technology, Nigeria

***Correspondence:**
Fang Jin
fangjin@gwu.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *26 October 2021*
Accepted: *28 March 2022*
Published: *28 April 2022*

Citation:
Guo M, Nguyen L, Du H and Jin F (2022) When Patients Recover From COVID-19: Data-Driven Insights From Wearable Technologies. Front. Big Data 5:801998. doi: 10.3389/fdata.2022.801998

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), manifests as a wide range of symptoms, including fever, cough, fatigue, breathing difficulties, loss of smell and taste, and pneumonia[1]. It spreads rapidly from infected people to others through close contact or small exhaled droplets. The pandemic is now causing havoc in countries around the world, with more than 282 million cases and around 5.41 million deaths, as of late December 2021 reported by WHO (2021). This deluge of patients is overwhelming hospitals everywhere, especially in some developing countries where vaccines are not sufficient, and it is difficult to cope with the need to conduct extensive disease testing programs and treat huge numbers of patients in a very short period. It is therefore vital for medical staff to be

[1]https://en.wikipedia.org/wiki/Coronavirus_disease_2019

able to identify patients COVID-19 disease stages before making the decision to hospitalize them. Severe patients need to be hospitalized quickly and receive a higher priority in dedicated treatment, while patients with milder symptoms might only need to self-quarantine at home. Fast and reliable techniques to detect and identify the disease stages are thus the focus of active research by scientists and medical technologists.

Vaira et al. found that anosmia and ageusia associated with fever (>37.5°C) are common onset symptoms that can be an early signal of a COVID-19 infection (discussed by Heerfordt and Heerfordt, 2020; Ortiz-Martínez et al., 2020; Vaira et al., 2020; Walker et al., 2020), therefore, investigated the use of Google Trends to study the loss of smell and smoking cessation and predicted COVID-19 incidence. Wang et al. (2020) built a deep convolutional neural network model to detect COVID-19 from chest X-ray images. Most of the existing work focused on early disease detection, but few works were proposed to identify the disease stages and develop useful insights for patients who must quarantine at home. We therefore propose to explore the problem of disease stage identification, because this will help doctors decide the most appropriate treatment plans for patients at each stage, allowing them to optimize their usage of scarce resources when the hospital is under pressure. Besides, since our work would help create a low-cost, efficient self-monitor solution that can be used by everyone, it is beneficial, especially for people who are quarantined at home.

Interestingly, there have been some huge improvements in wearable technologies over the last few years, with a number of wearable devices being widely introduced that enhance our everyday life. For example, smartwatches such as Fitbit[2] are helping us to track our sleep patterns and daily activities, encouraging us to maintain a healthier lifestyle. Smart Shirt is another example of this trend that is beginning to play an important role in our information infrastructure, supporting healthcare systems for monitoring vital signs efficiently and cost-effectively with the universal interface of clothing (Park and Jayaraman, 2003). The possibilities are seemingly unlimited: chip-integrated sensors are being used to monitor a number of physical medicine applications (Bonato, 2005). Sensors have already been developed specifically for COVID-19 applications, including an automatic sanitizer tunnel that detects a human being using an ultrasonic sensor from a distance of 1.5 feet and disinfects him/her using a sanitizer spray (Pandya et al., 2020). Quer et al. (2020) used wearable sensors to differentiate COVID-19 positive vs. negative cases in symptomatic individuals, pointing out that wearable devices are easy to access for most people. The fast development of wearable technologies makes it possible to be utilized to identify COVID-19 disease stages. However, existing studies are all either (i) mainly limited to the detection of COVID-19, with no attempt to identify the stages of the disease; (ii) not designed to analyze variations in the associated factors per COVID-19 stage; or (iii) unable to provide a comprehensive view of the disease for layman readers. Therefore, we seized this opportunity to investigate data-driven approaches to COVID-19 through wearable technologies in an attempt to bridge this gap. This paper introduces a wide-ranging set of data-driven approaches to identify infected patients' stages using wearable technologies. Specifically, this work aims to accurately and early infer from wearable data obtained from sensing devices attached to COVID-19 patients whether the COVID-19 patients are in mild, moderate, severe, or recovery stages in an earlier stage. We achieved this by introducing a model that utilizes a Long-short Term Memory (LSTM) network and a Deep Neural Network (DNN) to aggregate and jointly exploit temporal stream data from wearable devices and attribute stream from characteristics of patients. It is worth mentioning that our comprehensive experimental evaluation shows the improved performance achieved by our model compared to existing machine learning (ML) classification methods, which can only use one of the data streams. By identifying these patients in earlier stages, medical professionals will be able to take swift action if the patient requires early hospitalization or if it is safe for them to continue to self-quarantine at home. In addition, we also compare the lifestyles between severe and mild patients, allowing us to investigate and evaluate factors that impact the recovery of the patients. Specifically, the work aims to address the following three research questions (**RQs**):

- **RQ1:** Can we build an accurate ML model to predict COVID-19 stages and identify whether a patient will progress to a more severe stage in an earlier stage?
- **RQ2:** Which set of factors are associated with the severity of a patients symptoms? What can we learn from these factors in association with COVID-19 stages?
- **RQ3:** What signs signify recovery or deterioration in COVID-19 patients?

Overall, three novel contributions are made in this research:

1. We develop a classification model with uncertainty quantification to identify the major COVID-19 disease stages. Our model is able to recognize patients' disease stages in a timely manner because we utilize data from the wearable device, which is more responsive to disease stages than the subject's senses.
2. Our work provides useful insights into the progression of COVID-19 disease and vital indicators at each stage. The research input is from a data source (a wearable device like a smartwatch) that everyone can access and use on their own. Our approach is data-driven and can mitigate human bias substantially.
3. We investigate factors associated with COVID-19 severity and recovery. We also create case studies (1) demonstrating the differences between severe and mild patients and (2) showing the signs of recovery from COVID-19 disease using a shape-based pattern extraction model.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 discusses our methodology, including an overview of the data preparation, stage identification model, feature importance, and pattern extraction model. Section 4 shows our evaluation and experimental results. Section 5 presents

[2]https://www.fitbit.com

TABLE 1 | List of features specific to heart rate variability (HRV).

Feature name	Meaning
bpm	Heart rate
mxdmn	Difference between highest and lowest cardio interval values
sdnn	Standard deviation of normal heartbeat intervals
rmssd	Root mean square of successive differences for consecutive intervals
pnn50	Percent of RR-intervals that fall outside a 50 ms range of the average
mode	Most common cardio interval length
amo	Mode amplitude
lf	Power of low frequency waves
hf	Power of high frequency waves
vlf	Power of very low frequency waves
lfhf	Ratio of low to high frequency waves
total_power	Total power of HF, LF, and VLF waves generated by the heart
rr_data (time-based)	Intervals in milliseconds between consecutive heart beats

some limitation in our study. Finally, we offer conclusions in Section 6.

2. RELATED WORK

Here, we survey recent related studies on battling the COVID-19 crisis. These studies fall into two broad scientific areas: machine learning (ML) and remote monitoring utilizing the Internet of Things (IoT).

ML Research: Researchers have attempted different methods to battle COVID-19. Assaf et al. developed a model that used white blood cell count, time from symptoms to admission, oxygen saturation, and blood lymphocyte count to predict if a patient is at high risk for COVID-19. Their prediction model can be useful for efficient triage and in-hospital allocation, better prioritization of medical resources, and improving overall management (Assaf et al., 2020). Ahamad et al. (2020) developed a model that applies ML algorithms to reveal potential COVID-19 patients by analyzing their age, gender, fever, and history of travel. By extracting 11 blood indices through a random forest algorithm, Wu et al. (2020) built an assistant discrimination tool that can identify suspected patients using their blood test results. Barstugan et al. (2020) and Elaziz et al. (2020) choose to use image-based diagnosis (CT images) building Support Vector Machine and K-Nearest Neighbors algorithms for predicting suspected COVID-19 infection.

Remote monitoring research: However, these studies' data sources, such as CT images or blood test results, would often need to be collected by trained professionals. With COVID-19 patients number rising, we see a shortage of medical resources worldwide and make clinic visits bear more risk as suspected patients gather for examination. Therefore, many people prefer to use the Internet of Things (IoT) to diagnose COVID-19 to avoid the risk of infection. Singh et al. demonstrated that IoT implementation could help infected patients with COVID-19 identify symptoms rapidly and greatly reduce healthcare costs (Singh et al., 2020). Islam et al. (2020) suggested that wearable devices could provide real-time remote monitoring and contact tracing features, which can be used to improve healthcare systems' current management schemes. For example, Maghdid et al. (2020) designed an artificial intelligence-enabled framework that analyzes signals from a smartphone's sensor signal. It helped to diagnose the severity of pneumonia to predict the COVID-19 infection.

Most prior works were focusing on the early prediction or detection of COVID-19 infection. As the epidemic escalates dramatically every day, we want to further conserve healthcare resources by identifying different stages of COVID-19 patients. For example, diagnosed early and moderate stage patients could adopt self-quarantine treatment in time, saving valuable resources that can then be utilized by patients with severe COVID-19 stage.

3. METHOD

3.1. Data Preparation

3.1.1. Dataset Description

We used an open dataset provided by Welltory [3] The dataset comprises multivariate data records from 186 COVID-19 patients experiencing different stages. The data includes variables such as heart rate, sleeping patterns, daily activities, heart rate variability (HRV), blood pressure, patient demographics (age, gender, country, etc.), environmental information, and other patient facts (smoking, alcohol, other background diseases, etc.). We focus on the HRV information measured using wearable devices. HRV is also popular in many clinical and investigational research such as diabetes (Benichou et al., 2018), brain emotion, stress, anxiety (Goessl et al., 2017; Mather and Thayer, 2018), or cardiology related (Sessa et al., 2018). **Table 1** provides detailed descriptions of HRV specific features, where rr_data (intervals in milliseconds between consecutive heartbeats) is a sequence data with a length of 100. In addition, we also selected ordered categorical variables with values from 1 to 6 recording the intensity of seven common COVID-19 symptoms that were in the HRV survey dataset: breath, confusion, cough, fatigue, fever, pain, and bluish. We believe these variables can better assist in the task of prediction, but we only focus on the other HRV variables for the subsequent analysis.

Since each patient may be recorded multiple times, the stage of disease may be different from one recording period to the next. For example, some patients who were mild patients at the beginning of the record may become severe patients a week later. So, in the task of predicting the stage of disease, we remove the user code and predict the disease status for each record. All patients have a total of 1,480 complete records. Each record will be associated with a label by a survey from Welltory, identifying the corresponding patient's current stage. **Figure 1** summarizes the number of stages per disease stage category.

[3]https://github.com/Welltory/hrv-covid19

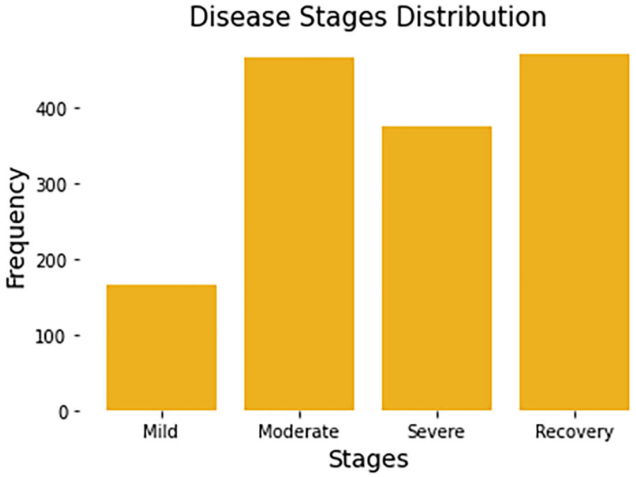

FIGURE 1 | Distribution of four disease stages. Each patient may span multiple disease stages due to the progression of the disease.

TABLE 2 | List of self-generated features (time-based and statistical features).

Domain	Feature name	Source
Time-based	*DI*	Lag difference of *rr* sequence
	DM	Absolute deviation from the mean of *rr* sequence
Statistical	*rr_var*	Variance of *rr* sequence
	rr_skew	Skewness of *rr* sequence
	rr_kurt	Kurtosis of *rr* sequence
	rr_max	Maximum of *rr* sequence
	rr_min	Minimum of *rr* sequence
	rr_median	Median of *rr* sequence
	rr_mean	Mean of *rr* sequence
	rr_iqr	Interquartile range of *rr* sequence
	morning	One-hot variables $I_{morning}$
	day	One-hot variables I_{day}
	evening	One-hot variables $I_{evening}$
	night	One-hot variables I_{night}
	daysa.t.	Number of days of COVID-19 infection

3.1.2. Feature Expansion

To make the most of the information in the data, we enrich our feature set based on temporal and statistical properties. First, for the variable time series, intervals in milliseconds between consecutive heartbeats (represented by *rr*), we computed a variety of statistics for this sequence, such as its variance (*rr_var*), skewness (*rr_skew*), kurtosis (*rr_kurt*), maximum (*rr_max*), minimum (*rr_min*), median (*rr_median*), mean (*rr_mean*), interquartile range (*rr_iqr*), etc. These features are popular and widely used in many research such as heart rate analysis (Bolanos et al., 2006) or brain waves recognition (Campisi and La Rocca, 2014). Besides, we divide each day into four periods and further create four one-hot variables: *morning*, *day*, *evening*, and *night*. That is, if a row of data for a patient is recorded in the morning, then the variable *morning* for this record is 1, while the other three variables are all 0. Another variable we created is called *day_after_test* (*days a.t.*), and its value depends on the number of days each patient has been infected with COVID-19.

In addition, we obtain two new temporal sequence data using the transformation of *rr_data*. Suppose the original heartbeat interval is $RR = \{x_1, x_2, ..., x_T\}$, we transform this time series by computing lag difference (*DI*) and the absolute deviation from the mean (*DM*), in order to remove temporal dependency and to eliminate the trend and seasonality of the time series. Mathematically, the two newly constructed time-series are as follows:

$$\begin{aligned} DI &= \{x_2 - x_1, x_3 - x_2, ..., x_T - x_{T-1}\}, \\ DM &= \{|x_1 - \overline{x}|, |x_2 - \overline{x}|, ..., |x_T - \overline{x}|\}, \end{aligned} \tag{1}$$

where $T = 100$ and $\overline{x}$ is the mean of the original *rr* sequence. To make these three sequences (*RR*, *DI*, and *DM*) have the same length 100, we add the average of the last three numbers of the DI sequence at the end of the DI sequence. All the features we expanded are listed in **Table 2**. Thus, we end up with a total of 32 attribute features and 3 temporal features for the task of predicting disease stages.

3.1.3. Data Pre-processing

There are some missing values in the dataset. It is either due to the network issues when the data is collected or the users choose not to answer some survey questions for any reason. To fill out the missing values, we used MissForest (Stekhoven and Bühlmann, 2012), a non-parametric iterative imputation technique based on the Random Forest algorithm which is proved capable of handling missing values of different data types. Additionally, we normalized the data to avoid scales influencing between features. Let $min\{X_{i,1:N}\}$ and $max\{X_{i,1:N}\}$ are the minimum and maximum values of the attribute feature X_i for all N samples. The min-max normalization values of feature X_i is computed as follows:

$$X'_{i,j} = \frac{X_{i,j} - min\{X_{i,1:N}\}}{max\{X_{i,1:N}\} - min\{X_{i,1:N}\}} \quad , \quad j = 1, 2, ..., N \tag{2}$$

Where $N = 1{,}480$ is the sample size.

Similarly, for the temporal sequence features, we use min-max normalization to normalize the data for all samples at each time point. Let $min\{X_{k,t,1:N}\}$ and $max\{X_{k,t,1:N}\}$ are the minimum and maximum values of the temporal feature X_k for all N samples at time t. The min-max normalization values of feature X_k is computed as:

$$X'_{k,t,j} = \frac{X_{k,t,j} - min\{X_{k,t,1:N}\}}{max\{X_{k,t,1:N}\} - min\{X_{k,t,1:N}\}} \quad , \quad j = 1, 2, ..., N, \quad t = 1, 2, ..., T \tag{3}$$

Where $N = 1{,}480$ is the sample size and $T = 100$ is the length of the temporal sequence.

3.2. Model for Disease Stage Identification

3.2.1. Theoretical Model

We formulate the problem of identifying disease stages as a multi-class classification problem. From a feature matrix X of a patient, we need to build a classifier f that classifies whether the patient is in *Mild, Moderate, Severe,* or *Recovery* stage.

In this task, our classification model utilizes two data streams described in Section 3.1: *temporal stream* and *attribute stream*. A temporal stream has temporal characteristics or sequential order. The temporal streams can be real-time, so if our model is embedded in wearable devices in the future, it will be very helpful for early-stage detection. The attribute stream has no temporal characteristics such as demographic information, patient's background disease, etc. Formally, assume that the dataset $\mathcal{D}$ of size N is defined as $\mathcal{D} = \{(X_i, Y_i), i = 1, ..., N\}$, where Y_i is the class label and $X_i = (X_i^t, X_i^a)$, represents the i-th sample of the combination of the temporal stream (denoted as X^t) and attribute stream (denoted as X^a). The developed classification model f parameterized by θ will classify disease stages based on input streams as the following equation:

$$stages \simeq f(\theta, H_2(\Phi(H_1(X^t), X^a))), \tag{4}$$

where H_1 and H_2 are latent feature extractors, which are two types of neural networks in our model, Φ is an aggregation function that fuses the latent features from $H_1(X^t)$ with attribute stream data X^a.

3.2.2. Network Design and Data Fusion Strategy

As mentioned earlier, the two input streams of the model are the temporal stream and the attribute stream. The LSTM network is suitable for temporal stream since it is a type of recurrent neural network (RNN) and addresses the problems of vanishing and exploding gradient in general RNNs. Hochreiter (1998). Therefore, in Equation (4), we choose H_1 as an LSTM based network to learn latent features from the temporal stream X^t. For the attribute stream X^a, after combining them with the outputs of the LSTM based network, we use H_2, a network of multiple fully-connected layers (DNN), to extract their latent features for the final disease stage classification. The DNN is chosen to force the network to explore all the possible relationships of both attribute streams and temporal streams. This is also an approach to combining DNN with LSTM to obtain a novel end-to-end neural network.

Figure 2 shows the overall model which composes of two subnetworks, LSTM and DNN. The two subnetworks are merged to predict the final disease stages. Suppose each patient has D input sequences with a common time length T. An LSTM passes forward over the entire temporal data sequences. We use the hidden size $H = 1$ in the LSTM, so later we can use an affine layer to map the hidden outputs to one-dimensional data of the same dimensional size as the attribute data. The LSTM unit is composed of a cell state c_t, a so-called memory cell, a hidden state h_t, an input gate i, a forget gate f, an output gate o, and an input modulation gate g. They are called gates because they control the flow through the LSTM. The four gates will be computed at each time step for cell and hidden state updates. The following is the outline formula of LSTM:

$$\begin{aligned} \begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} &= \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \\ c_t &= f \odot c_{t-1} + i \odot g \\ h_t &= o \odot \tanh(c_t) \end{aligned} \tag{5}$$

where σ and *tanh* are the sigmoid function and tanh function, respectively. W is the weight matrix. c_t, h_t, and x_t are the cell state, hidden state, and temporal input at time step t, respectively. $\odot$ represents element-wise multiplication.

After running the forward of the LSTM network, T hidden state outputs, $\{h_1, h_2, ..., h_T\}$, are returned and evenly sampled with a 20% probability to enhance generalization capability and avoid overfitting, that is, we uniformly sample T_1 hidden states from the T hidden states and T_1 = 1/5 T. Next, the combined hidden states are flattened to the temporal latent features thanks to the subsequent Affine layer to concatenate with the attribute stream. The temporal latent features have a final projected size $T_0 = 5$, which is equivalent to putting the temporal latent features into 5 additional latent attribute features. Let's define $h^t = H_1(X^t)$ as the final 5 latent features of the temporal stream and x^a as the sample values for the original attribute stream X^a. The concatenation of these two streams is defined as follows:

$$h^c = \Phi(h^t, x^a) = h^t \oplus x^a \tag{6}$$

where $\oplus$ is the concatenation operator. Then, the concatenated stream h^c is fed into a deep neuron network H_2 which consists of five fully connected layers with number of neurons 1,024, 1,024, 2,048, 1,024, and 1,024, respectively. The output of the model is the predicted probability of being in each disease stage for each sample. Finally, the predicted classification of disease stages y is obtained by the following:

$$y = \arg\max f(\theta, H_2(h^c)) \tag{7}$$

The network uses *Leaky ReLU* activation function and dropout rate of 30% to enhance the robustness of the model and reduce the computational cost. The learning rate is set to 0.001 and the batch size is set to be the same as the sample size. We use the Adam optimizer, gradient descent algorithm, and softmax cross-entropy loss function to optimize the network.

3.2.3. Uncertainty Quantification of the Model

We perform resampling from our existing samples to quantify the built predictive model's uncertainty. This method is also known as Bootstrap, published by Bradley Efron in Efron (1979). We employ the Bootstrap method because 1) it is invariant under re-parametrization; 2) it does not require the population distribution assumption; 3) it is driven by repeated resampling of data and does not depend on theoretical calculation; 4) it can provide the point estimation and assess the accuracy of the estimation when the traditional statistical method fails.

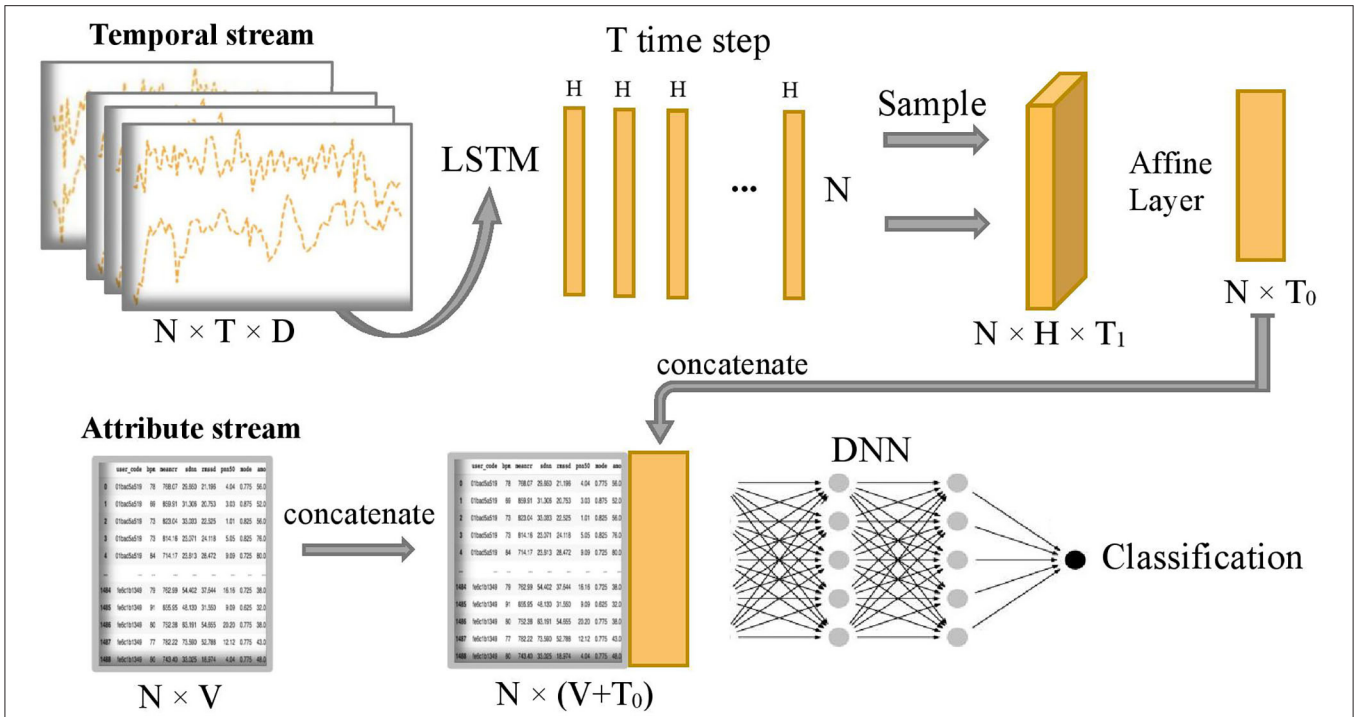

FIGURE 2 | Overview of COVID-19 stage classification model, where N=1,480 represents the sample size, T=100 represents the common length of temporal sequence, D=3 represents the number of temporal sequences, H=1 represents the size of hidden state output by LSTM, T_1=20 represents sampling size of the T hidden states, T_0=5 represents the final projected size of temporal features in the time dimension, and V=32 represents the number of attribute features.

We present details of the uncertainty quantification algorithm in **Algorithm 1**. Overall, the intuition of the algorithm is to create new samples, then obtain the prediction output. This process is repeated many times to result in a distribution of output which helps to quantify the model's uncertainty. In order to generate new samples, bootstrapping technique which was introduced by Efron (1979) is utilized. Here, we summarize its workflow in **Figure 3**:

- Treat the original sample as if it were the population.
- Draw from the sample, at random with replacement, for B times (B is the number of bootstraps).

Given the value of confidence interval (C.I) α%, we will retrain our model from the newly generated samples, perform classification, and obtain a α% confidence interval of the predicted outcomes.

Algorithm 1 Bootstrap method to construct 95% C.I. (Confidence Interval)

function compute_boot_CI ()
Input: Input Train dataset X, label y, Test dataset X^*, model f
Output: 95% C.I. $(l_c, u_c), c = 1, 2, 3, 4$ and c is the class index.

1. For Bootstrap $j = 1, ..., B$
   - Generate bootstrap sample X_j, y_j from dataset X and label y with replacement.
   - Train model f with bootstrap sample X_j, y_j.
   - Feed test dataset X^* to the above trained model and calculate the prediction outputs $p_{jc}, c = 1, 2, 3, 4$.
2. Let l_c and u_c be the 0.025 and 0.975 percentile of $(p_{1c}, ..., p_{Bc})$

return $(l_c, u_c), c = 1, 2, 3, 4$

3.2.4. Baseline Models and Comparison Metrics

To verify the effectiveness and advantages of our proposed approach, we compare the classification results on the test dataset with several classical ML and deep learning models using a five-fold cross-validation approach. The baseline models are as follows:

1. Logistic regression (Logit): a multinomial logistic regression model was used to predict the probabilities of different outcomes for our multi-class problem (Kwak and Clayton-Matthews, 2002).
2. Support vector machine (SVM) (Chang and Lin, 2011): various types of kernels were tried and the kernel with the best result was finally chosen.
3. Attribute-based K-nearest neighbors (KNN) (Peterson, 2009): various number of the k nearest neighbors were tried and the k with the best result was finally chosen.
4. Long short-term memory: a popular extension of artificial recurrent neural network (RNN) architecture. It was first introduced by Hochreiter and Schmidhuber (1997).

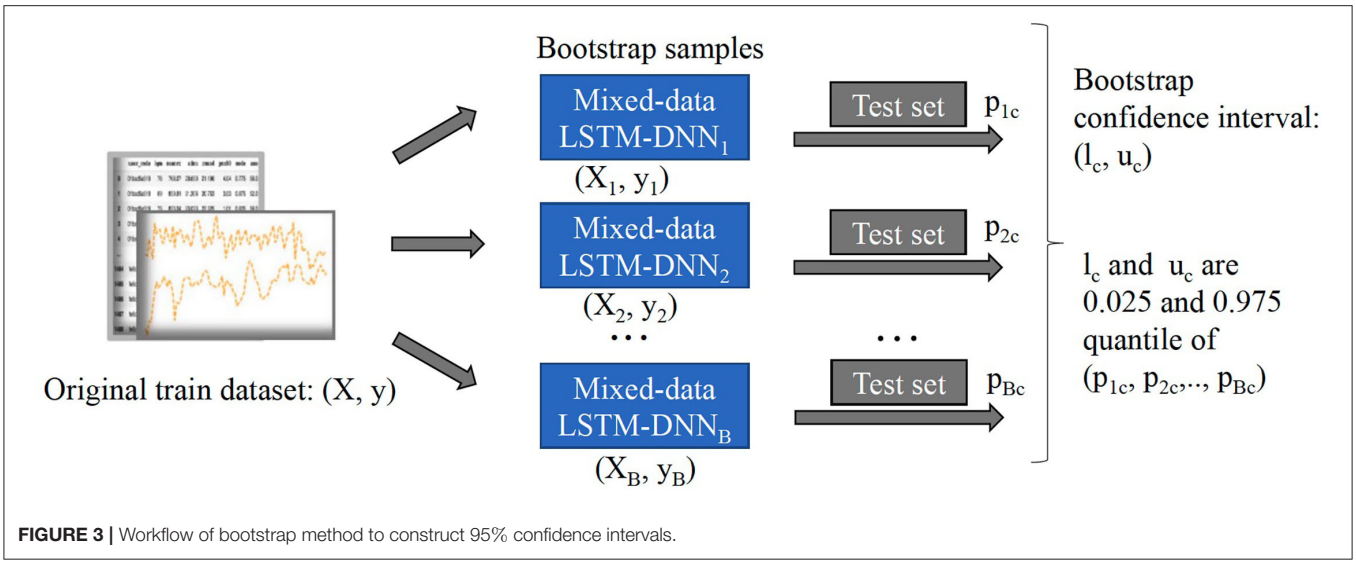

FIGURE 3 | Workflow of bootstrap method to construct 95% confidence intervals.

5. Deep Neural Network: it consists of five fully connected layers with a number of neurons 1,024, 1,024, 2,048, 1,024, and 1,024 respectively, and with the same activation function, dropout rate, learning rate, batch size, optimizer, algorithm, and loss function as our model.

For comparison metrics, we use standard metrics such as *accuracy, precision, recall, f1-score*, and *multi-class AUC* (area under ROC curve) to compare the performance of the models. It is worth noting that the inputs to these traditional models above can only be one of the data types and they cannot directly utilize both temporal data and attribute data jointly, so our model is expected to perform better than these models.

3.3. Feature Importance

To measure the importance of features, we perform the permutation feature importance algorithm on all the temporal and attribute features in turn to break the relationship between the feature and the true outcome. The permutation feature importance algorithm is described in **Algorithm 2**. This algorithm is based on our proposed classification model f. The general idea is that if a feature is essential for a stage, then shuffling or removing its values increases the model error for that stage because in this case, the model relied on the feature for the prediction. On the other hand, a feature is unimportant for a stage if shuffling or removing its values leaves the model error for that stage unchanged because, in this case, the model ignored the feature for the prediction (Fisher et al., 2019). Therefore, we can rank the losses of the built models after removing one variable at a time to select the most influential features. This approach is applied in Section 4.2 to uncover factors associated with different COVID-19 disease stages.

Algorithm 2 Permutation feature importance

function compute_feature_importance ()

Input: Feature X, label y, model f

Output: Output Feature importance FI

1. Estimate the original model error $e^{orig} = L(y, f)$
2. For feature $j = 1, ..., p$
   - Generate feature matrix X^{perm} by removing feature j in the data X. This breaks the association between feature j and true outcome y.
   - Estimate error $e^{perm} = L(y, f(X^{perm}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance $FI_j = e^{perm}/e^{orig}$.
3. Sort features by descending FI.

return FI

3.4. Model in a Case Study: Shape-Based Pattern Extraction Model for Signs of Recovery

In the classification of time series, a subsequence is called Shapelets (Ye and Keogh, 2009) if it maximally represents a class in some sense. Grabocka et al. (2014) introduced an implementable method to learn time-series shapelets. In one of our case studies 4.4, we try to find shapelets from HRV data that can differentiate between unrecovered patients and recovered patients. For signs of recovery, the patterns are two groups of shapelets that can linearly separate the recovered from unrecovered patients. Suppose $x_i, i = 1, 2, \ldots, N$ is the $i - th$ original time series data of length T, and $s_k, k = 1, 2, \ldots, K$ is one of the proposed shapelets with length l. It is easy to know that in a time series, there are exactly $T - l + 1$ segments as long as the starting index of the sliding window is incremented by one. The

distance between x_i and s_k is defined as follows:

$$d(x_i, s_k) = \min_{t \in \{1,2,\ldots,T-l+1\}} ||x_{i,t:t+l} - s_k||_2^2, \tag{8}$$

where $x_{i,t:t+l}$ is the subsequence of x_i from time t to time $t + l$. Since, in our study, the classification task is binary (recovery and unrecovered). Let us define the target variable, i.e., the patient's recovery status $Y_i, i = 1, 2, \ldots, N$:

$$Y_i = \begin{cases} 1 & \text{if the } i-th \text{ patient has recovered} \\ 0 & \text{if the } i-th \text{ patient has not recovered,} \end{cases} \tag{9}$$

Then, the predicted status of the $i-th$ patient is as follows:

$$\hat{Y}_i = W_0 + \sum_{k=1}^{K} d(x_i, s_k) W_k, \tag{10}$$

where W_k, $k = 0, 1, \ldots, K$, are the weights of learning, representing the classification hyperplane. By minimizing the logistic loss function with weight regularization terms, we can learn both the optimal shapelet and the optimal linear hyperplane. The loss function is shown in Equation (11):

$$L = \sum_{i=1}^{N} l(Y_i, \hat{Y}_i) + \lambda ||W||_2^2, \tag{11}$$

where

$$l(Y_i, \hat{Y}_i) = -Y_i log\sigma(\hat{Y}_i) - (1 - Y_i) log(1 - \sigma(\hat{Y}_i)), \tag{12}$$

and σ is the sigmoid function.

In the optimization process, a stochastic gradient descent (SGD) approach is adopted. Note that because SGD needs all the functions to be differentiable, an approximation of the minimum function (8) is used. This function is called the Soft Minimum function (Grabocka et al., 2014) and is shown in Equation (13).

$$\hat{d}(x_i, s_k) = \frac{\sum_{t=1}^{T-l+1} d_{i,k,t} e^{\alpha d_{i,k,t}}}{\sum_{t'=1}^{T-l+1} e^{\alpha d_{i,k,t'}}}, \tag{13}$$

where

$$d_{i,k,t} = (x_{i,t:t+l} - s_k)^2. \tag{14}$$

By applying the above method to the patient's HRV time series data, we aim to find a sequence pattern that can show signs of patient recovery to the greatest extent possible. Our results are shown in Section 4.4.

4. EXPERIMENTAL RESULTS

4.1. Infected Stage Classification Performance Evaluation

We randomly split up the data prior to modeling so that all models can use the same data splits. Each time, the models are trained on 4-folds (80% of the data) and tested on 1-fold (20% of the data). These 5-folds take turns being the test dataset to ensure that each sample can be classified. We perform a comprehensive comparison of model classification results. We add up the confusion matrices of the five experiments to obtain the total confusion matrix, which is therefore based on the result of all samples, as shown in **Figure 4**. For the five evaluation metrics, accuracy, precision, recall, f1-score, and multi-class AUC, we use the average results of the five experiments as the final evaluation results, which are listed in **Table 3**.

On the one hand, we can see the improvement in classifications of our proposed model from the confusion matrix. Our model has less misclassification of disease stages compared to other models. On the other hand, the detailed results in **Table 3** also show the advantages of our model. To be specific, the three models Logit, KNN, and SVM are comparable, having accuracy scores of about 0.66 to 0.79 and AUC of about 0.74 to 0.84. The LSTM model gives poor results due to the fact that it only uses temporal data. DNN model is the second-best model with an accuracy score of 0.903 and AUC of 0.924. Our proposed method has the highest scores under all five metrics, with an accuracy score of 0.914 and AUC of 0.935.

Figure 5 are box plots that present uncertainty quantification of the disease stage predictions of our proposed model for some randomly selected patients (Patient 151, 110, 29, and 182). The narrow box plot indicates the narrow 95% C.I., which presents low uncertainty in the prediction. We observe that for patient 29, all the C.I.s are quite narrow, while for all other patients, the C.I.s for certain stages are wider, which shows high prediction uncertainty. Even though there is high uncertainty in the prediction of certain disease stages, the 95% CI for each stage classification has shown that the probability of the classified stage (final prediction on each patient) always has a higher probability value than other stages. It means that our predictive model successfully identifies the disease stages with the performance results provided in **Table 1**.

4.2. Uncovering Factors Associated With COVID-19 Disease Stages

In this section, we focus our analysis on features from wearable data instead of other factors which have been discussed through news channels such as background diseases or body symptoms. We use a random permutation of values shown in **Algorithm 2** to calculate feature importance values for each feature based on the ratio of the model's errors between permutations. After obtaining the importance values, these values are rescaled to the range [0–1] to make them comparable. The results are shown in **Figure 6**. For each stage, the important features are ranked from high to low. The high importance feature means that prediction performance is highly dependent on this feature.

Figure 6 shows that for mild and moderate stages, the number of days from onset symptoms (*days a.t.*) is the most important since it ranks top among all variables. It means for mild and moderate patients, HRV variables have not yet shown very obvious characteristics, while the number of sick days can best determine the patients at this stage. This phenomenon is more reliable for mild patients since the number of sick days is far

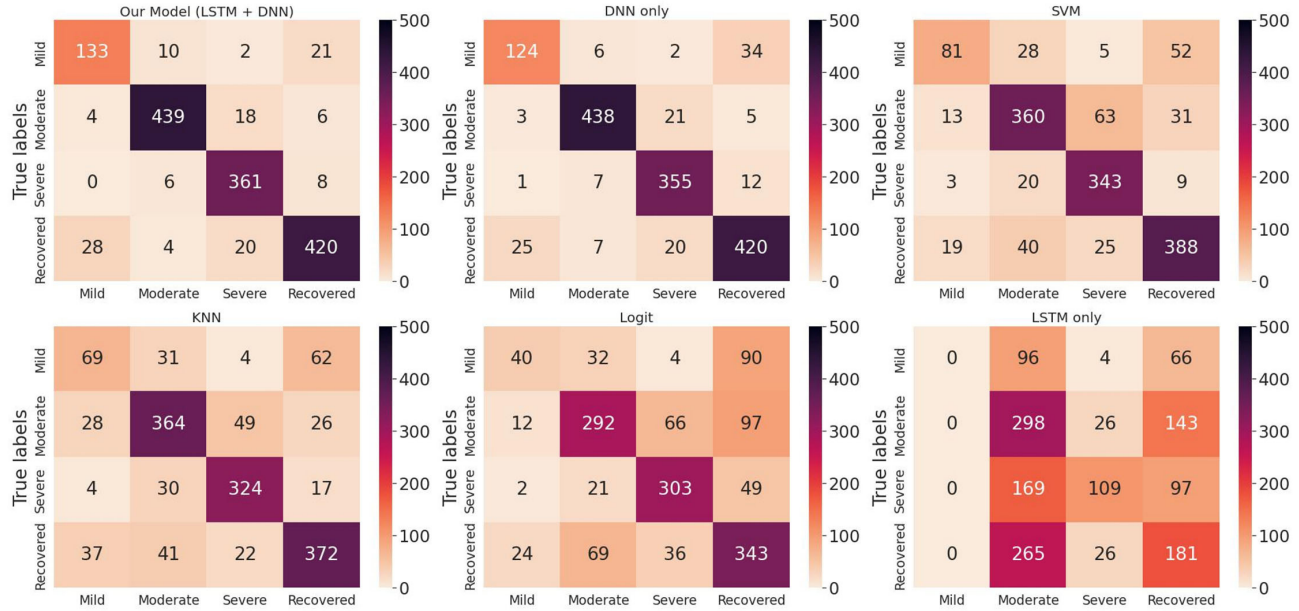

FIGURE 4 | Total confusion matrix for COVID-19 disease stage classification based on 5-fold cross-validation.

TABLE 3 | Infected stage classification results of models based on 5-fold cross-validation.

Model	Accuracy	Precision	Recall	F1-score	AUC
LSTM Only	0.397	0.410	0.397	0.355	0.556
Logit	0.661	0.666	0.661	0.650	0.741
KNN	0.763	0.761	0.763	0.759	0.816
SVM	0.792	0.791	0.792	0.787	0.839
DNN Only	0.903	0.905	0.903	0.903	0.924
Our Model (LSTM+DNN)	**0.914**	**0.917**	**0.914**	**0.914**	**0.935**

The bold values indicate the best result for each metric.

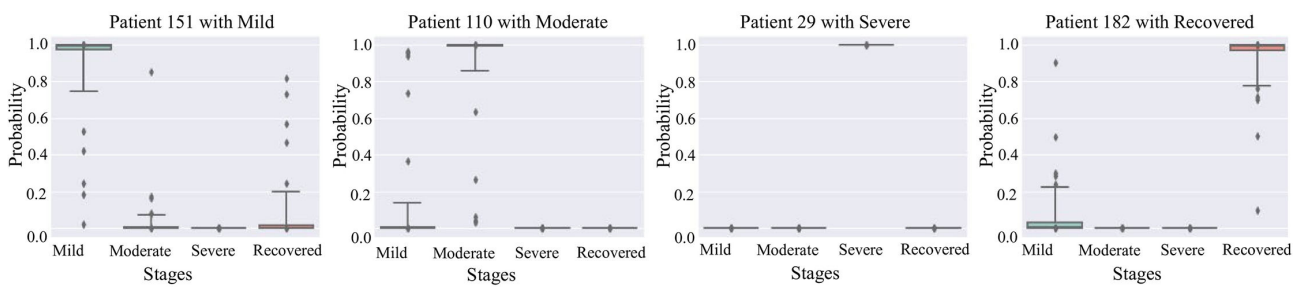

FIGURE 5 | 95% confidence interval of the prediction probabilities for the current stage of COVID-19 patients.

more important than the second-ranked variable. This result can be explained that in the early days of COVID-19 infection, most people have mild symptoms. For severe patients, the number of sick days is no longer important, the average time between each heartbeat, *rr_mean*, occupies the most important position, even though it is very unimportant in other stages. It indicates that the *rr_mean* of severe patients is very different from those patients in other stages. In other words, if the condition of a patient gets worse, it will be most clearly reflected by *rr_mean*. For recovery patients, the total power of waves generated by the heart (*total_power*) and the number of sick days (*days a.t.*) are important variables. This shows that, on the one hand,

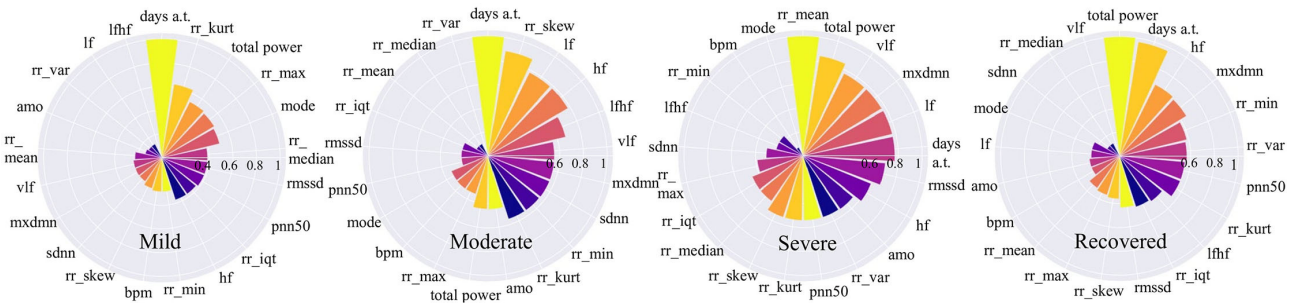

FIGURE 6 | Feature importance for COVID-19 stages. The higher the value is, the more important the feature is.

it takes a certain number of days for patients to recover; on the other hand, a significant change in the total power of the waves generated by the heart is most indicative of the recovery phase.

If we focus on different frequency wave power generated by the heart (high-frequency: *hf*, low-frequency: *lf*, very-low-frequency: *vlf*), we can also find something valuable. In the mild stage, no such variables are important. However, in moderate stage, the importance of all the three along with the ratio of low to high frequency waves (*lfhf*) rank relatively high. Therefore, compared to the patients in the mild stage, the wave power of each frequency of patients in the moderate stage has changed obviously. Besides, for severe patients, the frequency waves that are most different from other stages are low-frequency waves (*vlf*, *lf*). While for recovered patients, the frequency wave that is most different from other stages is a high-frequency wave (*hf*).

4.3. Case Study: Severe Patients vs. Mild Patients

Since heart rate variability (HRV) is popular in many healthcare-related research, we chose to explore it to compare daily patterns of severe patients vs. mild patients. The variables for comparison are the average time between each heartbeat (*rr_mean*), the percent of RR-intervals that fall outside a 50 ms range of the average (*pnn*50), and the total power of high-frequency waves, low-frequency waves, and very-low-frequency waves generated by the heart (*total_power*). All the data is normalized with the min-max technique to make them comparable. In addition, we choose data from 5 days before the onset of symptoms to 16 days after the onset of symptoms to show the difference between different stages in the most critical time. We use polynomial regression to do curve fitting and trending analysis separately. At the same time, 95% confidence intervals of fitted curves are shaded. We can find something interesting in the results shown in **Figure 7**.

We noticed that the highest value of the *total_power* curve and its confidence interval did not exceed 0.3. This range of *total_power* is relatively narrow since we have scaled all the data to the unit interval. It indicates that for people who have COVID-19 symptoms, whether he or she is in the mild stage or the severe stage, the total power of waves generated by the heart is lower approximately a few days before and 2 weeks after the onset. For these three comparative variables, *rr_mean*, *pnn*50, and *total_power*, their curves have a similar pattern. In general, after the symptom onset date, all three variables of severe patients are higher than those of mild patients. The higher value of average time between each heartbeat of severe patients means that their average heart rate is slower than that of mild patients. Furthermore, severe patients usually have higher *pnn*50. In other words, for severe patients, the outlier heartbeats, heartbeats whose intervals are farther apart from the average interval, occupy a larger proportion. It reveals that the heart rhythm of severe patients is more irregular than that of mild patients. Besides, compared to mild patients, heart-generated wave power of severe patients is stronger.

Following the time dimension, we can also find the different development of the above variables during the illness of mild and severe patients. Curves of patients in severe stage show a trend of increasing after decreasing. The curve of patients in mild stage also decreases at the beginning, while gradually stabilized after the curve rose and then again has a decreasing trend at about 12 to 14 days. This may be because the immune regulation of mild patients does not allow them to rise endlessly, which may also be a feature of gradual recovery. We can also see that after about 13 days, the 95% CI of the curves of both severe and mild patients are relatively narrow, which gives us more confidence to believe that severe and mild patients have indeed evolved in two directions.

4.4. Case Study: Signs of Recovery

In this case study, we try to find the most discriminate patterns that classify best the recovered stage and other stages. These patterns will signify the sign of recovery instead of the progressing disease. In addition, HRV data for the evening hours is used for analysis to avoid the influences of daytime activities of patients. We use the HRV sequence variables, which are the interval between consecutive heartbeats(*RR*), its lag difference sequence(*DI*), and its sequence of absolute deviation from the mean(*DM*) to extract the patterns. The methods for creating *DM* and *DI* can be found in Section 3.1.2. All three time series are normalized and combined to explore the discriminate patterns of recovery signs (See Section 3.4).

Figure 8 presents the extracted patterns that best discriminate the sign of recovery (top two subplots) and sample patterns from the patients (bottom four subplots). First, the heartbeat interval

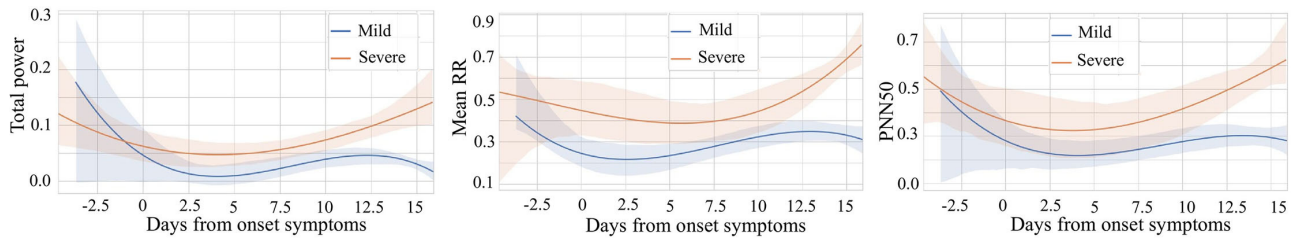

FIGURE 7 | Comparison of mild vs. severe patients based on three variables: *Total power*, *Mean RR*, and *PNN50*.

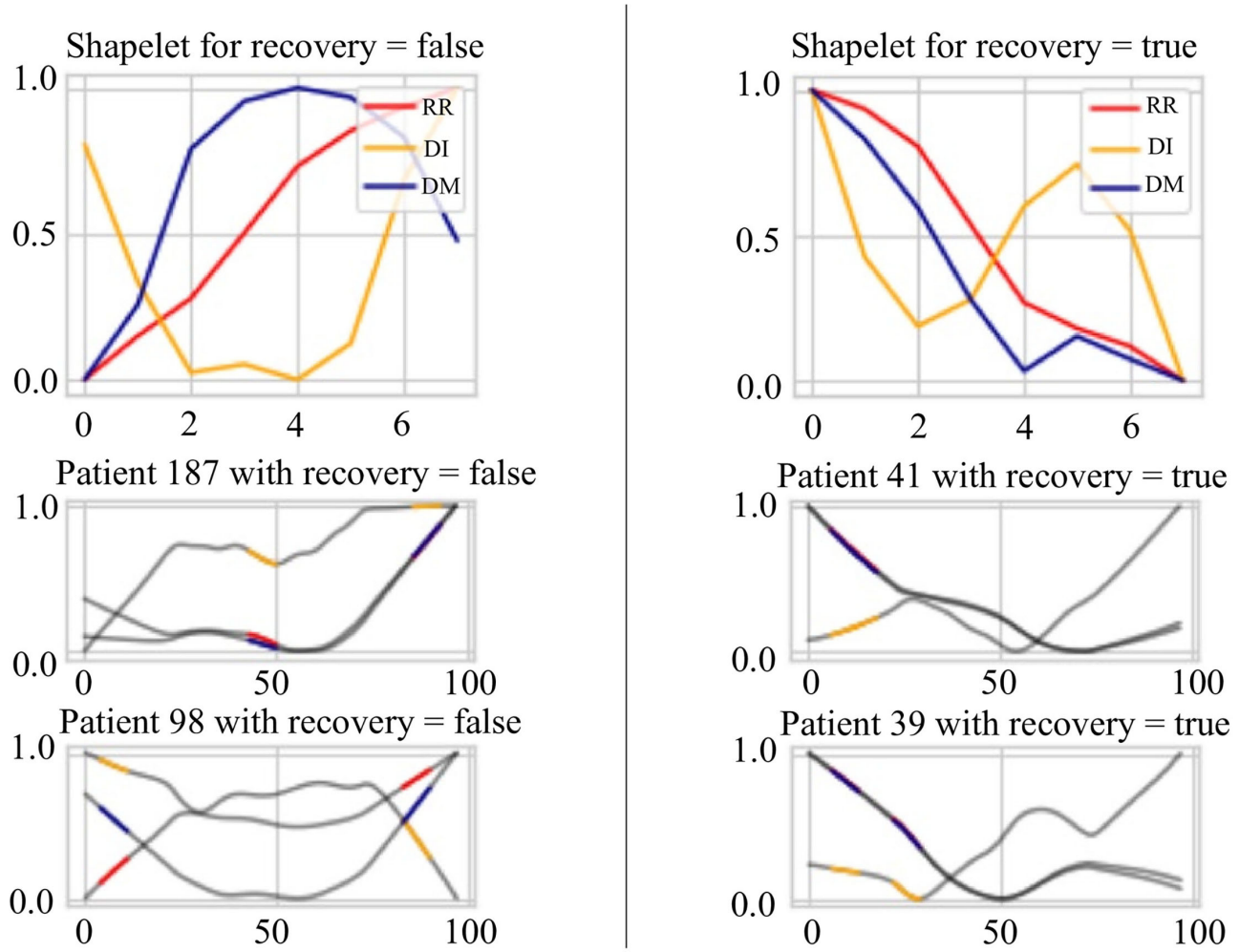

FIGURE 8 | Signs of recovery. The left and right columns present time series shapelets that differentiate between unrecovered patients and recovered patients, respectively. The red shapelet (RR) is the original heartbeat interval. The yellow shapelet (DI) is the differencing transformation of the heartbeat interval. The blue shapelet (DM) is the deviation of the heartbeat interval from the mean value.

data *RR* (in red) shows a decreasing trend for recovery cases than an increasing trend for other stages. Second, the heartbeat interval differencing data *DI* (in yellow) shows a sine-shaped pattern in the recovered group while it is a concave-parabola shape for unrecovered samples. Last, the absolute deviation from the mean data *DM* (in blue) shows a gradually decreasing trend in the recovered stage compared to a convex parabola shape in unrecovered situations. We can conclude a frequent change from these shapelets and an inconsistency of the COVID-19 patients. On the other hand, it shows an overall decreasing trend of the HRV data for the recovered patients in the evening. The subplots of the four patients show portions highlighted by different colors

representing different time series. These portions are the ones that are closely similar (having short Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978) distance in latent space) to the extracted shapelets and contribute to identifying signs of recovery.

5. LIMITATION

There are a few limitations in our study coming from the selected dataset. The number of patients in the study is 186, and they are not randomly selected. So, they are not representative of the entire population. However, this situation usually happens in healthcare data science research since it is time-consuming and expensive to obtain full data from a large population for the initial study. In addition, the uncertainty quantification of the model is down with the assumption that the set of observations is from an independent and identically distributed population. Moreover, some of the recorded data like coughing, having diabetic disease, etc., are self-reported, which have their own limitation. Self-reported information may not be accurate, depending on how honest the patients were when they did the survey.

6. CONCLUSION

In this work, we propose a novel predictive model to categorize COVID-19 patients into multiple stages (mild, moderate, severe, and recovered), using a wearable device dataset. Our predictive model exploits temporal stream data and attribute stream data simultaneously for disease stage classification and is able to identify severe patients in an earlier stage even if the symptoms seem to be "mild" or "moderate." In addition, we apply bootstrap methods to perform uncertainty quantification for the predictive model, and the experimental results demonstrate our predictive model's higher classification accuracy than other existing baseline approaches. Furthermore, we investigate each feature's importance to uncover its association with COVID-19 using a model-agnostic approach. Lastly, we investigate two cases in detail: 1) the first one is used to illustrate the comparisons between mild patients and severe patients. 2) the second one is used to analyze the signs of recovery. We observe that there are fluctuating HRV patterns in severe patients, but a more stable pattern and a clear trend in mild patients or recovering patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MG contributed to the model design and performed experiments. LN contributed to the experimental design and wrote the first draft of the manuscript. HD contributed to the model performance evaluation. FJ was responsible for the overall supervision and experiment design. All authors contributed to manuscript revision and provided critical feedback and helped shape the research.

REFERENCES

Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Li, P., Xu, H., et al. (2020). A machine learning model to identify early stage symptoms of sars-cov-2 infected patients. *Expert Syst. Appl.* 160, 113661. doi: 10.1016/j.eswa.2020.113661

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., et al. (2020). Utilization of machine-learning models to accurately predict the risk for critical covid-19. *Intern. Emerg. Med.* 15, 1435–1443. doi: 10.1007/s11739-020-02475-0

Barstugan, M., Ozkaya, U., and Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint* arXiv:2003.09424.

Benichou, T., Pereira, B., Mermillod, M., Tauveron, I., Pfabigan, D., Maqdasy, S., et al. (2018). Heart rate variability in type 2 diabetes mellitus: a systematic review and meta-analysis. *PLoS ONE* 13, e0195166. doi: 10.1371/journal.pone.0195166

Bolanos, M., Nazeran, H., and Haltiwanger, E. (2006). Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2006, 4289–4294. doi: 10.1109/IEMBS.2006.260607

Bonato, P. (2005). Advances in wearable technology and applications in physical medicine and Rehabilitation. *J. Neuro Eng. Rehabil.* 2, 2. doi: 10.1186/1743-0003-2-2

Campisi, P., and La Rocca, D. (2014). Brain waves for automatic biometric-based user recognition. *IEEE Trans. Inf. Forensics Security* 9, 782–800. doi: 10.1109/TIFS.2014.2308640

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., and Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of covid-19. *PLoS ONE* 15, e0235187. doi: 10.1371/journal.pone.0235187

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81. Available online at: https://arxiv.org/abs/1801.01489

Goessl, V. C., Curtiss, J. E., and Hofmann, S. G. (2017). The effect of heart rate variability biofeedback training on stress and anxiety: a meta-analysis. *Psychol. Med.* 47, 2578. doi: 10.1017/S0033291717001003

Grabocka, J., Schilling, N., Wistuba, M., and Schmidt-Thieme, L. (2014). "Learning time-series shapelets," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Hildesheim: ACM), 392–401.

Heerfordt, C., and Heerfordt, I. (2020). Has there been an increased interest in smoking cessation during the first months of the covid-19 pandemic? a google trends study. *Public Health* 183, 6. doi: 10.1016/j.puhe.2020.04.012

Hochreiter, S. (1998). Recurrent neural net learning and vanishing gradient. *Int. J. Uncertainity Fuzziness Knowl. Based Syst.* 6, 107–116. doi: 10.1142/S0218488598000094

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Islam, M. M., Mahmud, S., Muhammad, L. J., Islam, M. R., Nooruddin, S., Ayonet, S. I., et al. (2020). Wearable technology to assist the patients infected with novel coronavirus (covid-19). *SN Comput. Sci.* 1, 1–9. doi: 10.1007/s42979-020-00335-4

Kwak, C., and Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nurs Res.* 51, 404–410. doi: 10.1097/00006199-200211000-00009

Maghdid, H. S., Ghafoor, K. Z., Sadiq, A. S., Curran, K., Rawat, D. B., and Rabie, K. (2020). A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: design study. *arXiv preprint* arXiv:2003.07434. doi: 10.1109/IRI49571.2020.00033

Mather, M., and Thayer, J. F. (2018). How heart rate variability affects emotion regulation brain networks. *Curr. Opin. Behav. Sci.* 19, 98–104. doi: 10.1016/j.cobeha.2017.12.017

Ortiz-Martínez, Y., Garcia-Robledo, J. E., Vsquez-Castaeda, D. L., Bonilla-Aldana, D. K., and Rodriguez-Morales, A. J. (2020). Can google®trends predict covid-19 incidence and help preparedness? the situation in colombia. *Travel Med. Infect. Dis.* 37:101703. doi: 10.1016/j.tmaid.2020.101703

Pandya, S., Sur, A., and Kotecha, K. (2020). Smart epidemic tunnel: Iot-based sensor-fusion assistive technology for covid-19 disinfection. *Int. J. Pervasive Comput. Commun*. doi: 10.1108/IJPCC-07-2020-0091. [Epub ahead of print].

Park, S., and Jayaraman, S. (2003). Enhancing the quality of life through wearable technology. *IEEE Eng. Med. Biol. Mag.* 22, 41–48. doi: 10.1109/MEMB.2003.1213625

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia* 4, 1883. doi: 10.4249/scholarpedia.1883

Quer, G., Radin, J. M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E.. et al. (2020). Wearable sensor data and self-reported symptoms for covid-19 detection. *Nat. Med.* 27, 73–77. doi: 10.1038/s41591-020-1123-x

Sakoe, H., and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust*. 26, 43–49. doi: 10.1109/TASSP.1978.1163055

Sessa, F., Anna, V., Messina, G., Cibelli, G., Monda, V., Marsala, G., et al. (2018). Heart rate variability as predictive factor for sudden cardiac death. *Aging* 10, 166. doi: 10.18632/aging.101386

Singh, R. P., Javaid, M., Haleem, A., and Suman, R. (2020). Internet of things (iot) applications to fight against covid-19 pandemic. *Diabetes Metab. Syndr.* 14, :521–524. doi: 10.1016/j.dsx.2020.04.041

Stekhoven, D. J., and Bühlmann, P. (2012). Missforest non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Vaira, L. A., Salzano, G., Deiana, G., and De Riu, G. (2020). Anosmia and ageusia: common findings in covid-19 patients. *Laryngoscope* 30, 1787. doi: 10.1002/lary.28692

Walker, A., Hopkins, C., and Surda, P. (2020). "The use of google trends to investigate the loss of smell related searches during covid-19 outbreak," in *International Forum of Allergy and Rhinology*. (London: Wiley Online Library).

Wang, L., Lin, Z. Q., and Wong, A. (2020). Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-76550-z

WHO (2021). *Who Coronavirus Disease (COVID-19) Dashboard*. WHO.

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., et al. (2020). Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results. *medRxiv*. doi: 10.1101/2020.04.02.20051136

Ye, L., and Keogh, E. (2009). "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 947–956.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Medicine

SYSTEMATIC REVIEW
published: 06 May 2022
doi: 10.3389/fmed.2022.828346

Efficacy of Raman Spectroscopy in the Diagnosis of Uterine Cervical Neoplasms: A Meta-Analysis

Zhuo-Wei Shen[1†], Li-Jie Zhang[1†], Zhuo-Yi Shen[2†], Zhi-Feng Zhang[3], Fan Xu[4], Xiao Zhang[1], Rui Li[5] and Zhen Xiao[1]**

[1] Department of Obstetrics and Gynecology, The First Affiliated Hospital of Dalian Medical University, Dalian, China, [2] Department of Information Science and Technology, Wenhua University, Wuhan, China, [3] Department of Gastroenterology, The First Affiliated Hospital of Dalian Medical University, Dalian, China, [4] The Second Affiliated Hospital of North Sichuan Medical College, Nanchong Central Hospital, Nanchong, China, [5] Department of Physics, Dalian University of Technology, Dalian, China

Background: Uterine cervical neoplasms is widely concerned due to its high incidence rate. Early diagnosis is extremely important for prognosis. The purpose of this article is evaluating the efficacy of Raman spectroscopy in the diagnosis of suspected uterine cervical neoplasms.

Methods: We searched PubMed, Embase, Cochrane Central Register of Controlled Trials (CENTRAL), and Web of science up to September 1, 2021. By analyzing the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) of six included study, we evaluated the pooled and grouping sensitivity, specificity, positive, and negative likelihood ratios (LR), and diagnostic odds ratio (DOR), with 95% confidence intervals (CI), based on random effects models. The overall diagnostic accuracy of Raman spectrum was evaluated by SROC curve analysis and AUC.

Results: After screening with inclusion and exclusion criteria, a total of six study were included in the study. The pooled sensitivity and specificity was 0.98 (95% CI, 0.93–0.99) and 0.95 (95% CI, 0.89–0.98). The total PLR and NLR were 21.05 (95% CI, 8.23–53.86) and 0.03 (95% CI, 0.01–0.07), respectively. And the AUC of the SROC curve which show the overall diagnostic accuracy was 0.99 (0.98–1.00).

Conclusion: Through analysis, we confirmed the role of Raman spectroscopy (RS) in the diagnosis of suspected uterine cervical tumors.

Systematic Review Registration: [https://www.crd.york.ac.uk/prospero/], identifier [CRD42021284966].

Keywords: Raman spectroscopy, uterine cervical tumors, diagnostic efficacy, meta-analysis, translational medicine

OPEN ACCESS

Edited by:
Pietro Lio',
University of Cambridge,
United Kingdom

Reviewed by:
Jinwei Qiang,
Jinshan Hospital, China
Sylvain Honore Woromogo,
Cancer Epidemiology Public Health,
Republic of Congo

****Correspondence:***
Rui Li
rli@dlut.edu.cn
Zhen Xiao
seriousdoc@163.com

[†] These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *03 December 2021*
Accepted: *13 April 2022*
Published: *06 May 2022*

Citation:
Shen ZW, Zhang LJ, Shen ZY, Zhang ZF, Xu F, Zhang X, Li R and Xiao Z (2022) Efficacy of Raman Spectroscopy in the Diagnosis of Uterine Cervical Neoplasms: A Meta-Analysis. Front. Med. 9:828346. doi: 10.3389/fmed.2022.828346

INTRODUCTION

The incidence rate of uterine cervical tumors is the fourth of female cancer. According to statistics, there were about 570,000 uterine cervical tumors patients and 310,000 deaths worldwide in 2018. Among them, China and India are the hardest hit areas of uterine cervical tumors, accounting for nearly two-thirds of the cases (1). Early diagnosis of cervical cancer and cervical intraepithelial neoplasia and early treatment are effective means to improve the survival rate of cervical

cancer patients. Although there are screening tools such as cytological smear (TCT) and human papillomavirus (HPV) detection, the average sensitivity and specificity are not satisfactory (2).

More than 10 years ago, TCT was an effective tool for detecting and preventing uterine cervical tumors. However, the European guidelines for quality assurance of uterine cervical tumors screening (Abstract literature of the Second Edition) released in 2010 pointed out that the false positive rate of cytology is high, which will bring excessive medical treatment and additional economic losses (3). Therefore, HPV DNA detection was recommended due to its high sensitivity. But HPV DNA detection also had the problems of time-consuming and high price. Colposcopy had good sensitivity (>90%), but its specificity was poor (<50%), and the false positive rate was higher, which often lead to unnecessary biopsy. Histopathological examination is the gold standard for the evaluation and diagnosis of cancer, but it includes chemical fixation, dehydration, clearance, infiltration, paraffin embedding, sectioning, and hematoxylin eosin (H&E) staining. It takes about 1 week, which is time-consuming and expensive.

Raman spectroscopy is a new and reliable technology, which can analyze the molecular structure of substances and the chemical composition of human tissues (4). In medical research, Raman imaging has been successfully applied to nasopharyngeal carcinoma (5), gastric cancer (6), lung cancer (7), esophageal cancer (8), renal cell carcinoma (9), brain tumor (10) and so on. Raman technology has been used in the study of uterine cervical tumors for decades. The existing literature has proved that the specificity and accuracy of Raman spectroscopy in the diagnosis of uterine cervical tumors can reach more than 90%, which is no less than the traditional hematoxylin-eosin (HE) staining. Compared with HE staining, Raman technology has the advantages of no staining, no fixation, less demand for professionals, faster and so on, which provides another feasibility for the diagnosis of uterine cervical tumors (11). In conclusion, if Raman spectroscopy can be applied to cervical cancer, we have every reason to believe that it can carry out early diagnosis of cervical cancer and improve the screening rate of cervical cancer and the survival rate of patients. This Meta-analysis reviews the application of Raman spectroscopy in cervical cancer.

METHODS

Literature Research

This meta-analysis searched PubMed, Embase, Cochrane Central Register of Controlled Trials (CENTRAL), and Web of science to ensure that all potentially eligible articles are included (last search: September 1, 2021). We combined all the relevant medical subject heading (MeSH) terms of uterine cervical tumors and Raman spectrum: [(Uterine Cervical Neoplasms) OR (Cervical Neoplasm, Uterine) OR (Cervical Neoplasms, Uterine) OR (Neoplasm, Uterine Cervical) OR (Neoplasms, Uterine Cervical) OR (Uterine Cervical Neoplasm) OR (Neoplasms, Cervical) OR (Cervical Neoplasms) OR (Cervical Neoplasm) OR (Neoplasm, Cervical) OR (Neoplasms, Cervix) OR (Cervix Neoplasms) OR (Cervix Neoplasm) OR (Neoplasm, Cervix) OR (Cancer of the Uterine Cervix) OR (Cancer of the Cervix) OR (Cervical Cancer) OR (Uterine Cervical Cancer) OR (Cancer, Uterine Cervical) OR (Cancers, Uterine Cervical) OR (Cervical Cancer, Uterine) OR (Cervical Cancers, Uterine) OR (Uterine Cervical Cancers) OR (Cancer of Cervix) OR (Cervix Cancer) OR (Cancer, Cervix) OR (Cancers, Cervix)] AND [(Spectrum Analysis, Raman) OR (Raman Spectrum Analysis) OR (Raman Spectroscopy) OR (Spectroscopy, Raman) OR (Analysis, Raman Spectrum) OR (Raman Optical Activity Spectroscopy) OR (Raman Scattering) OR (Scattering, Raman)]. All potential studies were included with no other limitation. The meta-analysis has been registered in PROSPERO (CRD42021284966).

Selection Criteria and Exclusion Criteria

Articles like review articles, comments, report, letters will be eliminated from the study. Criteria as follows: (I) without animal tissues in the experiments; (II) reported the use of RS in uterine cervical tumors; (III) used histopathology to confirm the diagnosis; (V) reported the true positive (TP), false positive (FP), true negative (TN) and false negative (FN), based on which the sensitivity and specificity values can be calculated. After screening, a total of six study were included in the study.

Data Extraction

Two independent investigators extracted a range of data from each study using a standardized data-collecting form: article title, first author, publication year, nationality. All relevant data is contained within the 6 included articles (12–17). Then the primary parameters, which mean the diagnostic value, including TP, FP, TN, and FN. And we can use these parameters to calculate the sensitivity and specificity values. The data obtained were summarized in **Table 1**.

Statistical Analysis

We calculated the primary data of TP, FP, TN, FN from articles included, then calculated sensitivity, specificity, positive and negative likelihood ratios (LR), based on random effects models. We used Review Man 5.3 and Stata/SE 15.1 to generate the forest plots in order to show sensitivity and specificity.

Meanwhile, Summary Receiver Operator Characteristics (SROC) curves was generated to assess the combination of sensitivity and specificity by Stata/SE 15.1. To assess publication bias, we generated funnel plot using Stata/SE 15.1. In the meantime, we found that articles in uterine cervical tumors include *in vivo* and *in vitro* studies. Therefor we conducted a subgroup analysis according to these studies.

Risk of Bias (Quality) Assessment

Two independent investigators used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) guidelines by Review Manager 5.3 to evaluate the quality of included studies. And the risk of bias of included studies was shown in **Figures 1A,B**. To assess publication bias, we plotted funnel plots and Egger's regression test using Stata/SE 15.1. The funnel plots and Egger's regression test included in the study are shown in **Figures 1C,D**.

TABLE 1 | Characteristics of the included studies.

References	Country		N1	N2	N3	TP	FP	FN	TN	Sensitivity	Specificity	Diagnostic algorithm	Sample	Spectra
Daniel et al. (12)	India	Vitro	25	36	U	23	9	2	27	92%	75%	LDA	Fresh tissue slices (20 μm)	784.12 nm
Daniel et al. (13)	India	Vitro	145	64	U	143	2	2	62	99%	97%	PC-LDA	Fresh tissue slices (20 μm)	784.12 nm
Lyng et al. (14)	Ireland	Vitro	10	20	398	195	2	3	198	98%	99%	PC-LDA	FFPP(10 μm)	514.5 nm
Shaikh et al. (15)	India	Vivo	31	30	154	80	4	0	70	100%	95%	PC-LDA	Cervix *in vivo*	785 nm
Shaikh et al. (16)	India	Vivo	20	6	146	61	3	6	76	91%	96%	PC-LDA	Cervix *in vivo*	785 nm
Jing et al. (17)	China	Vitro	11	11	22	11	1	0	10	100%	91%	ORR (NADH/FAD)	Fresh tissue slices (4 μm)	430 nm

U, unknown; N1, number of patients; N2, number of healthy; N3, number of tested spectra; FFPP, Formalin-fixed paraffin preserved, PCA, principal component analysis; LDA, linear discriminate analysis; PC-LDA, Principal-component linear discriminant analysis.

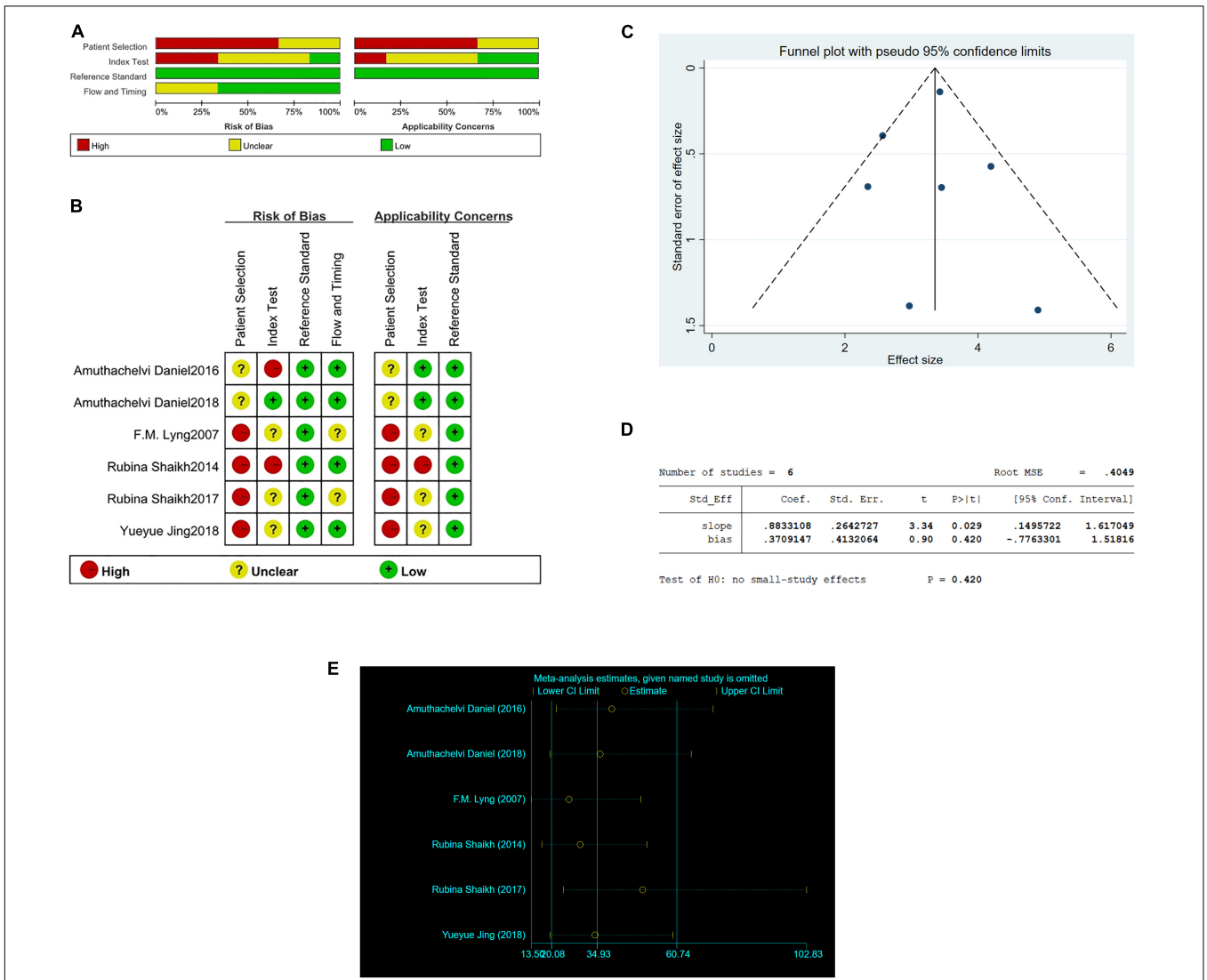

FIGURE 1 | The graphical display of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) of the included studies. **(A)** Risk of bias and applicability concerns evaluation of included studies in pool. **(B)** Risk of bias and applicability concerns evaluation of included studies individually. **(C)** Funnel plot of publication bias in Raman diagnosis of cervical cancer. **(D)** Egger's regression test of publication bias in Raman diagnosis of cervical cancer. **(E)** Sensitivity analysis in Raman diagnosis of cervical cancer.

As shown in the **Figure 1D**, $P = 0.420$, less than 0.05, and Egger's regression test indicates that there is no publication bias.

And we conducted a sensitivity analysis. In **Figure 1E**, the results showed that none of the studies had an impact on this meta-analysis.

RESULTS

Search Results

The process of included articles screening was presented in **Figure 2**. 403 potential articles were searched at first (including PubMed, $n = 106$, Web of science, $n = 186$, Embase, $n = 111$), in which included 198 duplicate records. Among the rest of 205 articles, 38 articles excluded due to: they were review, meeting or letters. Go a step further by browsing the 167 potentially relevant studies, 126 records excluded due to they were cytological study ($n = 79$), serological research ($n = 28$), medicine efficacy study ($n = 11$), animals research ($n = 8$). By reading the rest of 41 articles, 24 reports excluded due to they were biochemical assessment ($n = 12$), failed to give concrete date ($n = 6$) and irrelevant to the subject ($n = 6$). After careful perusing, 5 articles excluded due to failed to mention TN, FN, TP, FP and 6 excluded because of cervical precancer. Ultimately, 6 studies included in this review.

Characteristics of the Included Studies

Table 1 carefully described the particular characteristics of the 6 included articles. Among the 6 articles, 5 were published between 2014 and 2018, the rest of article was published in 2007. There are a total of 242 patients and 167 normal people in the included articles, and the total number of spectra incorporated was 720 (two articles didn't provide the number of spectra). In terms of the nationalities, four studies were from India, other two studies were from China and Ireland, respectively. As for diagnostic algorithm, one article calculated ORR (NADH/FAD), another article used linear discriminate analysis (LDA), and the other four articles utilized Principal-component linear discriminant analysis (PC-LDA). In term of spectra, two studies applied 785 nm, other two studies applied 784.12 nm, and the other two studies applied 430 and 514.5 nm, respectively. All of six studies utilized tissue to research, two studies were *in vivo*, therefor their samples were cervix *in vivo*, and the other four studies were *in vitro*, so their samples were *ex vivo* tissues. Three of four studies *in vitro* obtained fresh tissue slices, the rest of one study obtained Formalin-fixed paraffin preserved tissue.

Pooled Data Analysis

The sensitivity and specificity were calculated to assess diagnostic accuracy of all the six studies. And the forest plot of pooled sensitivity and specificity was shown in **Figure 3**. The sensitivity which meant the detection of uterine cervical tumors by RS, ranged from 0.91 (95% CI, 0.82–0.97) to 1.00 (95% Cl, 0.95–1.00) and the pooled sensitivity was 0.98 (95% Cl, 0.93–0.99). The sensitivity of all the six studies was more than 0.90, which was mean that the missed diagnosis rate of RS for uterine cervical tumors is very low. The specificity ranged from 0.75 (95% CI, 0.58–0.88) to 0.99 (95% Cl, 0.96–1.00), and the pooled specificity was 0.95 (95% Cl, 0.89–0.98). It should be noted that except for one study with sensitivity of 0.75, specificity of the other five studies were more than 0.90. In a word, the ability of RS to distinguish cancer from normal people was worthy of recognition.

The total PLR and NLR were 21.05 (95% CI, 8.23–53.86) and 0.03 (95% CI, 0.01–0.07), respectively. And the AUC of the SROC curve which show the overall diagnostic accuracy was 0.99 (0.98–1.00). The plots were shown in **Figure 3C**.

Subgroup Analysis

Vivo Group

Two studies (15, 16) showed the research of RS to uterine cervical tumors *in vivo* which had a total of 87 samples and 300 tested spectra. The sensitivity of two studies was 1.00 (95% Cl, 0.95–1.00) and 0.91 (95% Cl, 0.82–0.97), respectively, and the specificity was 0.95 (95% Cl, 0.87–0.99) and 0.96 (95% Cl, 0.89–0.99), respectively. Since the number of study included in this group is less than 4, data analysis cannot be done in STATA. All of the data and grouping situation were shown in **Figure 4**.

Vitro Group

Four studies (12–14, 17) showed the research of RS to uterine cervical tumors *in vitro* which had a total of 322 samples and 420 tested spectra (two articles didn't provide the number of spectra). The sensitivity of four studies ranged from 0.92 (95% Cl, 0.74–0.99) to 1.00 (95% Cl, 0.72–1.00), and the pooled sensitivity was 0.98 (95% Cl, 0.89–1.00). The specificity ranged from 0.75 (95% Cl, 0.58–0.88) to 0.99 (95% Cl, 0.96–1.00), and the pooled specificity was 0.97 (95% Cl, 0.94–0.99). Total PLR and NLR were 33.38 (95% Cl, 15.00–74.28) and 0.02 (95% Cl, 0.00–0.12), respectively. The SROC curve was described and the AUC was 0.99 (0.98–1.00). All of the plots of *vitro* group were shown in **Figure 5**.

DISCUSSION

Mahadevan-Jansen et al. first researched uterine cervical tumors *in vivo* and *in vitro* by RS in 1998 (18). That means the research of uterine cervical tumors by RS has had more than 20 years history. Related articles research different substances, such as fresh cervical tissues, cervical cells, blood serum and so on. According to searching, this study is the first meta-analysis attempt to analyze the meaning of RS for uterine cervical tumors by researching fresh cervical tissues, and we intend to confirm its diagnostic accuracy by means of this study.

Meta-analysis showed that RS had high diagnostic accuracy for uterine cervical tumors. The sensitivity of all included articles was more than 90%, and the specificity of most included articles (except for one 75%) were also more than 90%. In the subgroup analysis, the sensitivity and specificity also achieved high standard, that meant whether RS analyze uterine cervical tumors tissues *in vivo* or *in vitro* both showed high diagnostic accuracy. This is strong evidence to explain the diagnostic effect of RS in uterine cervical tumors. Although there are only two

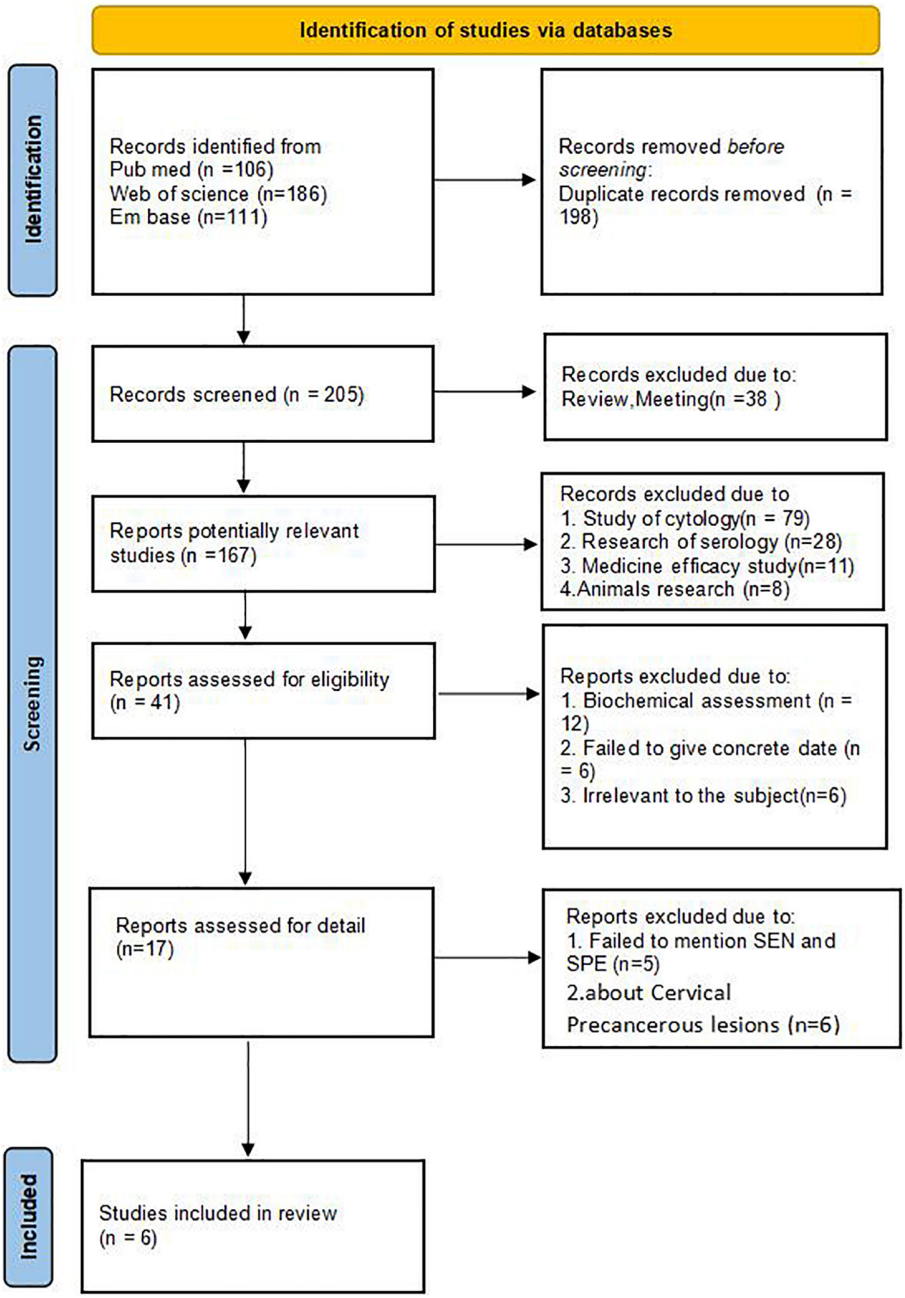

FIGURE 2 | PRISMA 2020 flow diagram.

literatures *in vivo* subgroup analysis, but for new technologies, such high sensitivity and specificity deserve our attention, and we look forward to seeing more research. And from the perspective of the combination of engineering with medicine, such new technologies and new ideas really deserve our attention.

RS also was used in researching uterine cervical tumors by cervical cells and blood serum except fresh cervical tissue. Sitarz et al. (19) studied the cervical cells of 96 women after TCT and HPV testing. They evaluated Glycogen levels in cells of all study groups to prove that RS can also diagnose HPV infected cells. Karunakaran et al. (20) found that the accuracy of RS in diagnosing uterine cervical tumors and normal people using single cells, cell clusters and DNA were 93.84, 74.26, and 92.21%, respectively. Lu et al. (21) studied the serum of 150 women and detected the levels of SCCA and OPN in the serum by RS. This is a convenient and efficient method which maybe a new screening measure for uterine cervical tumors.

With the prevalence of TCT and HPV examination, pathological biopsy is widely used in clinic and is considered as

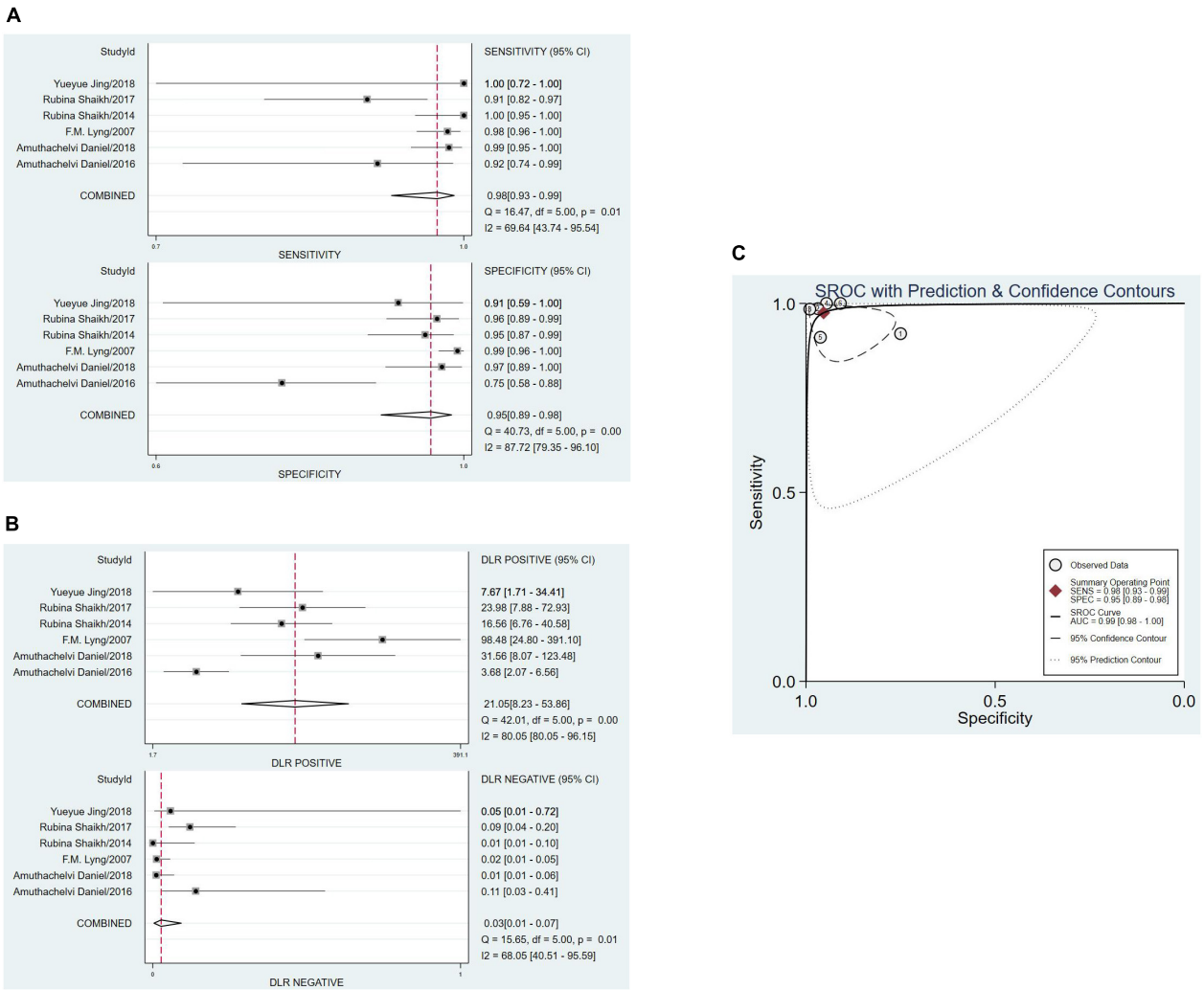

FIGURE 3 | The pooled date analysis of Raman spectroscopy (RS) in uterine cervical tumors. **(A)** The forest plot of pooled sensitivity and specificity of Raman spectroscopy to diagnose uterine cervical tumors of all the six studies. **(B)** The pooled PLR and NLR of Raman spectroscopy in diagnosis of uterine cervical tumors. PLR, positive likelihood ratios; NLR, negative likelihood ratios. **(C)** The SROC curve of Raman spectroscopy in diagnosis of uterine cervical tumors. SROC, summary receiver operator characteristics.

Subgroup analysis-vivo group

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Rubina Shaikh2014	80	4	0	70	1.00 [0.95, 1.00]	0.95 [0.87, 0.99]
Rubina Shaikh2017	61	3	6	76	0.91 [0.82, 0.97]	0.96 [0.89, 0.99]

Subgroup analysis-vitro group

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Amuthachelvi Daniel2016	23	9	2	27	0.92 [0.74, 0.99]	0.75 [0.58, 0.88]
Amuthachelvi Daniel2018	143	2	2	62	0.99 [0.95, 1.00]	0.97 [0.89, 1.00]
F.M. Lyng2007	195	2	3	198	0.98 [0.96, 1.00]	0.99 [0.96, 1.00]
Yueyue Jing2018	11	1	0	10	1.00 [0.72, 1.00]	0.91 [0.59, 1.00]

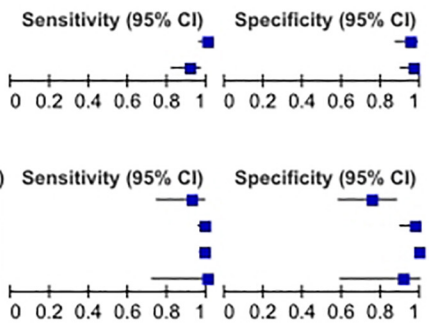

FIGURE 4 | The subgroup analysis of vivo group and *vitro* group.

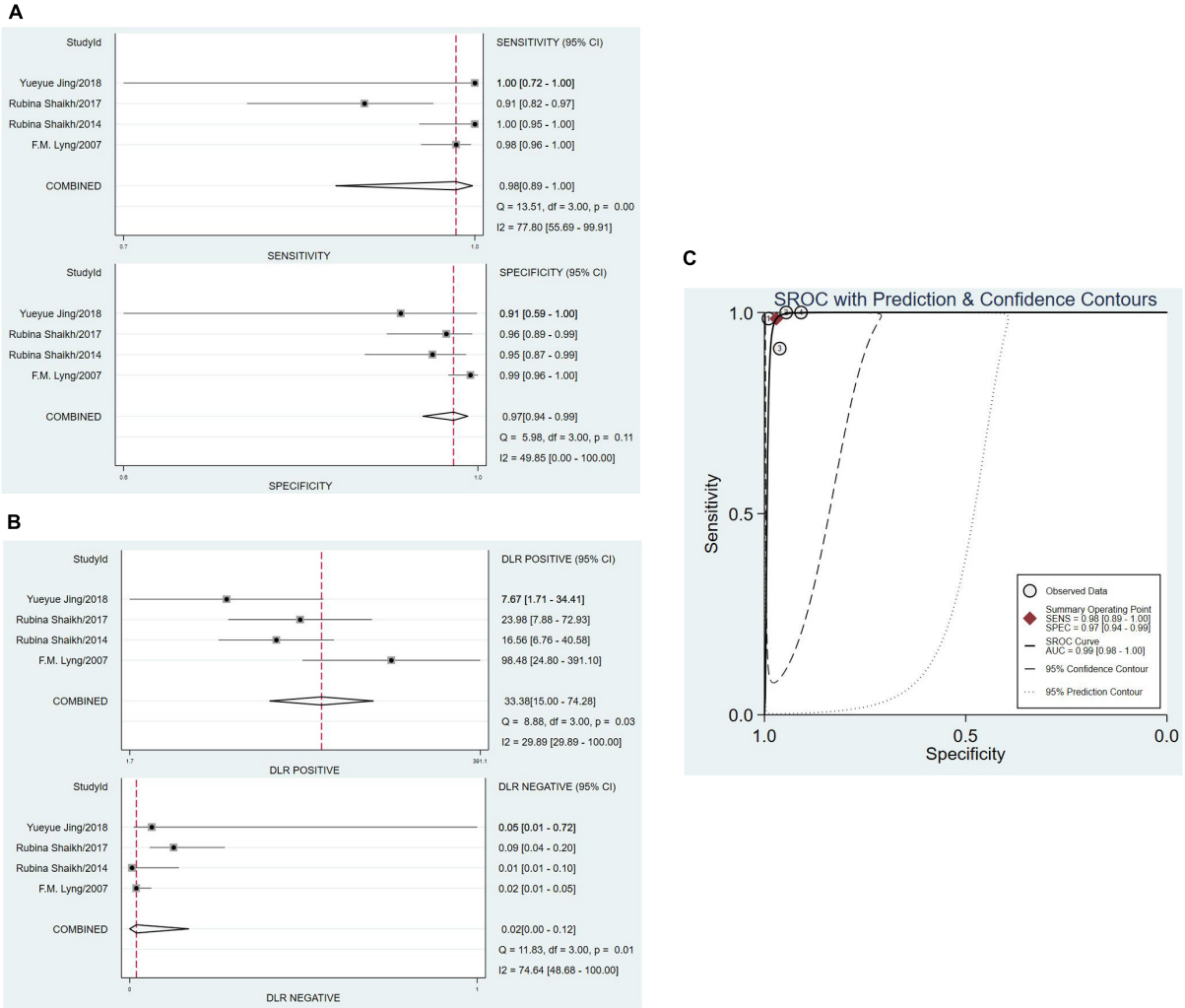

FIGURE 5 | The pooled date analysis of Raman spectroscopy (RS) in uterine cervical tumors *in vitro* group. **(A)** The forest plot of pooled sensitivity and specificity of Raman spectroscopy to diagnose uterine cervical tumors of four studies. **(B)** The pooled PLR and NLR of Raman spectroscopy in diagnosis of uterine cervical tumors. PLR, positive likelihood ratios; NLR, negative likelihood ratios. **(C)** The SROC curve of Raman spectroscopy in diagnosis of uterine cervical tumors. SROC, summary receiver operator characteristics.

```
Meta-regression                                   Number of obs  =       6
REML estimate of between-study variance           tau2           =       0
% residual variation due to heterogeneity         I-squared_res  =   0.00%
Proportion of between-study variance explained    Adj R-squared  =      .%
Joint test for all covariates                     Model F(4,1)   =    0.17
With Knapp-Hartung modification                   Prob > F       =  0.9275
```

logRR	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	-.0647452	.1569677	-0.41	0.751	-2.059209	1.929719
Country	-.3373604	1.635869	-0.21	0.871	-21.12304	20.44832
Diagnostic	.546221	1.104628	0.49	0.708	-13.48941	14.58186
Spectra	-.1595999	.397954	-0.40	0.757	-5.216085	4.896885
_cons	131.2993	317.1446	0.41	0.750	-3898.405	4161.004

FIGURE 6 | Meta-regression analysis on year, country, diagnostic algorithm, spectra.

the gold standard of cervical cancer, what are the outstanding advantages of Raman technology? In other words, how should Raman technology position itself in clinical application?

After reading a lot of literature, people generally believe that the outstanding advantage of Raman microscope lies in its timeliness, such as real-time images, convenience and rapidity, reducing the demand and burden of pathologists and so on. According to the current research progress, Raman technology does not seem to be enough to make us think that it can replace postoperative pathology. However, with the rapid development of modern science and technology, there is an emerging technology called handheld Raman spectrometer, which can quickly and quantitatively detect the anti-cancer drug 5-fluorouracil (5-FU) in serum (22). We have every reason to expect that this technology can be innovated and applied to clinic as soon as possible, such as handheld portable Raman device. This device is smaller, imaging is faster, it is more convenient to determine the scope of lesions, reduce the burden of pathologists, and shorten the time waiting for intraoperative freezing during surgery, so as to realize efficient diagnosis in cost and time.

There are some limitations in this article. First and foremost, the heterogeneity was high. In order to explore the reasons for this result, we conducted a sensitivity analysis, and the results have been analyzed in **Figure 1E**. Excluding the included literature one by one did not have a great impact on heterogeneity. And meta regression, grouped by year, country, analysis tool, and Raman wave number, respectively, *P*-values are greater than 0.05, it means no great significance (**Figure 6**). We believe that the most likely reason is that there is too few research included due to the lack of current research. Second, because the vast majority of studies do not strictly abide by the double-blind test rules when conducting Raman test, there are some errors in the screening of patients, which may affect the analysis results. Third, one of the documents was published in 2007, and the rest were studied in recent 8 years. We don't know whether microscope technology has developed greatly during this period. However, because there are few articles in conformity, we did not rule it out, and we think this meta can better explain the diagnostic effect of Raman technology in cervical cancer in the past 15 years. If someone continues to choose research in the follow-up, they can directly choose the literature from this time. Fourth, there are only two literatures *in vivo* subgroup analysis. Too few may not directly indicate the effectiveness of Raman technology, which needs more sample size and literature research.

CONCLUSION

Due to the high cost and expense of RS, there are not many related studies at present. But in the existing research, it is believed that RS does play an important role in the diagnosis of uterine cervical tumors. This is a satisfactory result which predicts the emergence of a new and efficient diagnostic technology.

Through this meta-analysis, we can confidently believe that Raman spectroscopy has high specificity and sensitivity in the diagnosis of uterine cervical tumors, and we have reason to believe that Raman spectroscopy will become an efficient diagnostic method of uterine cervical tumors in the future. However, more research and evidence are needed to fully demonstrate the role of Raman spectroscopy in the diagnosis of uterine cervical tumors before it is used in clinic. We are also looking forward to more samples and more researches.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

Z-WS, L-JZ, Z-YS, XZ, Z-FZ, and FX wrote the manuscript together and made great contributions to the article. This article has taken place under the guidance of two experienced tutors, ZX and RL. All authors agreed to be responsible for the content of this article and the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China-Liaoning Joint Fund 2019-BS-073 and the Scientific Research Fund of Liaoning Provincial Education Department LZ2019044.

REFERENCES

1. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob Health.* (2020) 8:e191–203. doi: 10.1016/S2214-109X(19)30482-6
2. de Kok IM, van der Aa MA, van Ballegooijen M, Siesling S, Karim-Kos HE, van Kemenade FJ, et al. Trends in cervical cancer in the Netherlands until 2007: has the bottom been reached? *Int J Cancer.* (2011) 128:2174–81. doi: 10.1002/ijc.25553
3. Arbyn M, Anttila A, Jordan J, Ronco G, Schenck U, Segnan N, et al. European guidelines for quality assurance in cervical cancer screening. Second edition–summary document. *Ann Oncol.* (2010) 21:448–58. doi: 10.1093/annonc/mdp471
4. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst.* (2019) 111:923–32. doi: 10.1093/jnci/djy225
5. Shu C, Yan H, Zheng W, Lin K, James A, Selvarajan S, et al. Deep learning-guided fiberoptic Raman spectroscopy enables real-time

in vivo diagnosis and assessment of nasopharyngeal carcinoma and post-treatment efficacy during endoscopy. *Anal Chem.* (2021) 93:10898–906. doi: 10.1021/acs.analchem.1c01559
6. Harmsen S, Rogalla S, Huang R, Spaliviero M, Neuschmelting V, Hayakawa Y, et al. Detection of premalignant gastrointestinal lesions using surface-enhanced resonance Raman scattering-nanoparticle endoscopy. *ACS Nano.* (2019) 13:1354–64. doi: 10.1021/acsnano.8b06808
7. Shin H, Oh S, Hong S, Kang M, Kang D, Ji YG, et al. Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. *ACS Nano.* (2020) 14:5435–44. doi: 10.1021/acsnano.9b09119
8. Almond LM, Hutchings J, Lloyd G, Barr H, Shepherd N, Day J, et al. Endoscopic Raman spectroscopy enables objective diagnosis of dysplasia in Barrett's esophagus. *Gastrointest Endosc.* (2014) 79:37–45. doi: 10.1016/j.gie.2013.05.028
9. He C, Wu X, Zhou J, Chen Y, Ye J. Raman optical identification of renal cell carcinoma via machine learning. *Spectrochim Acta A Mol Biomol Spectrosc.* (2021) 252:119520. doi: 10.1016/j.saa.2021.119520
10. Jermyn M, Mok K, Mercier J, Desroches J, Pichette J, Saint-Arnaud K, et al. Intraoperative brain cancer detection with Raman spectroscopy in humans. *Sci Transl Med.* (2015) 7:274ra19. doi: 10.1126/scitranslmed.aaa2384
11. Ramos IR, Malkin A, Lyng FM. Current advances in the application of Raman spectroscopy for molecular diagnosis of cervical cancer. *Biomed Res Int.* (2015) 2015:561242. doi: 10.1155/2015/561242
12. Daniel A, Prakasarao A, Dornadula K, Ganesan S. Polarized Raman spectroscopy unravels the biomolecular structural changes in cervical cancer. *Spectrochim Acta A Mol Biomol Spectrosc.* (2016) 152:58–63. doi: 10.1016/j.saa.2015.06.053
13. Daniel A, Prakasarao A, Ganesan S. Near-infrared Raman spectroscopy for estimating biochemical changes associated with different pathological conditions of cervix. *Spectrochim Acta A Mol Biomol Spectrosc.* (2018) 190:409–16. doi: 10.1016/j.saa.2017.09.014
14. Lyng FM, Faoláin EO, Conroy J, Meade AD, Knief P, Duffy B, et al. Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool. *Exp Mol Pathol.* (2007) 82:121–9. doi: 10.1016/j.yexmp.2007.01.001 x
15. Shaikh R, Dora TK, Chopra S, Maheshwari A, Kedar KD, Bharat R, et al. In vivo Raman spectroscopy of human uterine cervix: exploring the utility of vagina as an internal control. *J Biomed Opt.* (2014) 19:087001. doi: 10.1117/1.JBO.19.8.087001
16. Shaikh R, Prabitha VG, Dora TK, Chopra S, Maheshwari A, Deodhar K, et al. A comparative evaluation of diffuse reflectance and Raman spectroscopy in the detection of cervical cancer. *J Biophotonics.* (2017) 10:242–52. doi: 10.1002/jbio.201500248
17. Jing Y, Wang Y, Wang X, Song C, Ma J, Xie Y, et al. Label-free imaging and spectroscopy for early detection of cervical cancer. *J Biophotonics.* (2018) 11:e201700245. doi: 10.1002/jbio.201700245
18. Mahadevan-Jansen A, Mitchell MF, Ramanujam N, Malpica A, Thomsen S, Utzinger U, et al. Near-infrared Raman spectroscopy for in vitro detection of cervical precancers. *Photochem Photobiol.* (1998) 68:123–32. doi: 10.1562/0031-865519980682.3.co;2
19. Sitarz K, Czamara K, Bialecka J, Klimek M, Zawilinska B, Szostek S, et al. HPV infection significantly accelerates glycogen metabolism in cervical cells with large nuclei: Raman microscopic study with subcellular resolution. *Int J Mol Sci.* (2020) 21:2667. doi: 10.3390/ijms21082667
20. Karunakaran V, Saritha VN, Joseph MM, Nair JB, Saranya G, Raghu KG, et al. Diagnostic spectro-cytology revealing differential recognition of cervical cancer lesions by label-free surface enhanced Raman fingerprints and chemometrics. *Nanomedicine.* (2020) 29:102276. doi: 10.1016/j.nano.2020.102276
21. Lu D, Ran M, Liu Y, Xia J, Bi L, Cao X. SERS spectroscopy using Au-Ag nanoshuttles and hydrophobic paper-based Au nanoflower substrate for simultaneous detection of dual cervical cancer-associated serum biomarkers. *Anal Bioanal Chem.* (2020) 412:7099–112. doi: 10.1007/s00216-020-02843-x
22. Zhou G, Li P, Ge M, Wang J, Chen S, Nie Y, et al. Controlling the shrinkage of 3D hot spot droplets as a microreactor for quantitative SERS detection of anticancer drugs in serum using a handheld Raman spectrometer. *Anal Chem.* (2022) 94:4831–40. doi: 10.1021/acs.analchem.2c00071

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 09 June 2022
doi: 10.3389/fmed.2022.771219

Screening New Blood Indicators for Non-alcoholic Fatty Liver Disease (NAFLD) Diagnosis of Chinese Based on Machine Learning

Cheng Wang[1†], *Junbin Yan*[2,3†], *Shuo Zhang*[4], *Yiwen Xie*[5], *Yunmeng Nie*[2], *Zhiyun Chen*[2,3*] *and Sumei Xu*[5*]

[1] Applied Math Department, China Jiliang University, Hangzhou, China, [2] The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China, [3] Key Laboratory of Integrative Chinese and Western Medicine for the Diagnosis and Treatment of Circulatory Diseases of Zhejiang Province, Hangzhou, China, [4] Gastroenterology Department, Zhejiang Provincial Hospital of Chinese Medicine, Hangzhou, China, [5] Department of General Practice, Zhejiang Provincial Hospital of Chinese Medicine, Hangzhou, China

OPEN ACCESS

Edited by:
Jingjing You,
The University of Sydney, Australia

Reviewed by:
Qinghua Yao,
Zhejiang Cancer Hospital, China
Jun Liu,
China Jiliang University, China
Damjana Rozman,
University of Ljubljana, Slovenia

***Correspondence:**
Zhiyun Chen
zhiyunchen63@163.com
Sumei Xu
xsmdoctor@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: 06 September 2021
Accepted: 28 April 2022
Published: 09 June 2022

Citation:
Wang C, Yan J, Zhang S, Xie Y, Nie Y, Chen Z and Xu S (2022) Screening New Blood Indicators for Non-alcoholic Fatty Liver Disease (NAFLD) Diagnosis of Chinese Based on Machine Learning. Front. Med. 9:771219. doi: 10.3389/fmed.2022.771219

Background: The prevalence of NAFLD is increasing annually. The early diagnosis and control are crucial for the disease. Currently, metabolic indicators are always used clinically as an auxiliary diagnosis of NAFLD. However, the prevalence of NAFLD is not only increased in obese/metabolic-disordered populations. NAFLD patients with thin body are also increasing. Only using metabolic indicators to assist in the diagnosis of NAFLD may have some deficiencies. Continue to develop more clinical auxiliary diagnostic indicators is pressing.

Methods: Machine learning methods are applied to capture risk factors for NAFLD in 365 adults from Zhejiang Province. Predictive models are constructed for NAFLD using fibrinolytic indicators and metabolic indicators as predictors respectively. Then the predictive effects are compared; ELISA kits were used to detect the blood indicators of non-NAFLD and NAFLD patients and compare the differences.

Results: The prediction accuracy for NAFLD based on fibrinolytic indicators [Tissue Plasminogen Activator (TPA), Plasminogen Activator Inhibitor-1 (PAI-1)] is higher than that based on metabolic indicators. TPA and PAI-1 are more suitable than metabolic indicators to be selected to predict NAFLD.

Conclusions: The fibrinolytic indicators have a stronger association with NAFLD than metabolic indicators. We should attach more importance to TPA and PAI-1, in addition to TC, HDL-C, LDL-C, and ALT/AST, when conducting blood tests to assess NAFLD.

Keywords: non-alcoholic fatty liver disease (NAFLD), TPA, PAI-1, machine learning, support vector machine (SVM), predictive model

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) has become one of the most common liver diseases, affecting about 25% of the general population worldwide, in which Asia (27%) has higher prevalence rates comparing with North America (24%) and Europe (24%) (1, 2). As the largest country in Asia, the prevalence of NAFLD in China is also increasing annually (3). NAFLD is closely related to metabolism (4), so metabolic indicators are often used to assist the diagnosis of NAFLD in the clinic. Total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), alanine transaminase/aspartate transaminase (ALT/AST), and body mass index (BMI), and other indicators of metabolism, are all regarded as important factors related to the risk of NAFLD and contribute to the diagnosis (5–7). However, there are still some shortcomings in using only metabolic indicators as predictors of NAFLD. A paper published in "The Lancet" (8) pointed out that even thin people are not immune to fatty liver disease. Of the total incidence of NAFLD, 40% of patients with NAFLD had normal BMIs (18.5–23.9), and 20% of non-obese people had NAFLD. This inspired us to find more evidence in addition to metabolism for the accurate diagnosis of NAFLD. Our research focused on the fibrinolytic indicators.

The physiological balance of TPA/PAI-1 plays an essential role in regulating blood patency and preventing atherosclerosis (9). Also, plasma TPA and PAI-1 are associated with many metabolic diseases including NAFLD, heart disease, and diabetes mellitus (DM) (10–12). Jin found that plasma PAI-1 levels were significantly increased in children with increased severity of steatosis, lobular inflammation, ballooning, and fibrosis (13). Furthermore, PAI-1 was strongly correlated with plasma lipids and insulin resistance indices (13). By analyzing 210 Taiwanese NAFLD patients and 420 gender- and age-matched control groups, Chang found that based on univariate analysis, TG, BMI, LDL, HDL, ALT, AST, TPA, and PAI-1 are all related to NAFLD (14). However, less research provided good predictive accuracy for NAFLD diagnosis based on fibrinolytic or metabolic indicators. And there was little research on comparing the impacts of these indicators on NAFLD diagnosis quantitatively.

In this study, we applied machine learning (ML) which has been increasingly used in the field of liver disease and liver transplantation (15) to construct the predictive models for NAFLD based on those blood indicators, and obtained good predictive accuracy. It also compares the accuracy of prediction, looking for which indicators are more suitable for NAFLD diagnosis, fibrinolytic indicators, or metabolic indicators? We collected the datasets of 365 patients who had blood tests and NAFLD labels from the Traditional Chinese Medicine hospital of Zhejiang Province. The support vector machine (SVM) method was applied to the dataset to construct a predictive model for NAFLD based on the indicators above. SVM has been used to identify molecular markers of hepatocellular carcinoma (HCC) (16), but no one has yet used it to screen NAFLD auxiliary diagnostic indicators. We compared the prediction accuracy for NAFLD diagnosis based on fibrinolytic indicators (TPA and PAI-1) with the prediction accuracy based on metabolic indicators (TC, HDL-C, LDL-C, ALT/AST), screened the more accurate one.

MATERIALS AND METHODS

Screen and Compare Diagnostic Indicators

Subjects

Ethics Statement

Ethics statement Written informed consent was obtained from each participant, and the study was approved by the Committee for the protection of human subjects of The First Affiliated Hospital, Zhejiang Chinese Medical University. The corresponding ethical approval code (2018-K-061-01).

Inclusion Criteria

This study investigated 365 adult individuals aged 18–65 on whom we had complete data. They are from the health examination center of the Traditional Chinese Medicine hospital of Zhejiang Province. The following subjects were excluded:

(1) pregnant or lactating women;
(2) who has one of the following diseases: heart, brain, blood, lung, kidney, endocrine, mental, viral hepatitis, tuberculosis, AIDS, scarlet fever, drug-induced hepatitis, autoimmune liver disease, Wilson's disease, and liver cancer;
(3) who has taken anticoagulants in the last half month.

365 adult individuals who met the inclusion and exclusion criteria were divided into the Normal group ($n = 99$) and the NAFLD group ($n = 299$) according to the B-ultrasound results for follow-up analysis. Detailed clinical data can be found in **Supplementary Table 1**.

Methods

The following variables are included in our model: gender, age, body mass index (BMI), body height, TPA, PAI-1, TC, HDL-C, LDL-C, ALT/AST. These input variables were linearly scaled to the range [0, 1] and were mapped into a high-dimensional feature space. For details, see **Table 1**.

Comparisons between the two groups (NAFLD vs. non-NAFLD) were conducted using Student t-tests for continuous variables and Pearson tests for categorical variables.

SVM methods were taken to construct predictive models for NAFLD. SVM is a very popular supervised machine learning classifier widely used in classification or discrimination analysis. For non-linear and complicated relationships in high-dimensional variables, SVM is usually more effective than Logistic and other ordinary statistical methods. In this research, the relationship between NAFLD and blood indicators is complicated and no regular mathematical function can precisely describe the mechanisms between NAFLD and blood indicators. So SVM is suitable for our topic.

We introduce briefly the idea of svm. Let X_i denote the input variables such as TPA, PAI-1, BMI and so on in our case, and y_i denote the lable of each sample. The purpose of SVM model is to find a function $\omega^T X_i + b$ to predict the lable as accurate as possible. It implement the following optimal problem to solve

TABLE 1 | The characteristic clinical data between the NAFLD and non-NAFLD patients.

Characteristic	Non-NAFLD ($n = 99$)	NAFLD ($n = 266$)
Gender (*n*, %)	Female (79, 79.8%)	Female (55, 20.7%)
	Male (20, 20.2%)	Male (211, 79.3%)
Age, meidan (IQR)	40 (35, 48)	42 (37, 51)
tpa, meidan (IQR)	5,956.28 (3,923.5, 8,163.93)	9,239.07 (6,383.61, 11,975.68)
pai-1, meidan (IQR)	15,384.16 (13,605.38, 18,530.64)	32,095.67 (23,665.17, 37,275.04)
PAI-1/TPA, meidan (IQR)	2.66 (1.88, 3.88)	3.31 (2.22, 5.25)
BMI, meidan (IQR)	22.08 (20.07, 23.86)	26.37 (24.69, 28.31)
TC, meidan (IQR)	4.19 (3.8, 4.77)	4.8 (4.27, 5.44)
TG, meidan (IQR)	0.9 (0.68, 1.12)	1.68 (1.19, 2.37)
HDL-C, meidan (IQR)	1.49 (1.29, 1.67)	1.09 (0.97, 1.28)
LDL-C, meidan (IQR)	2.08 (1.77, 2.44)	2.65 (2.18, 3.1)
ALT, meidan (IQR)	14 (11, 18)	26 (19, 38)
AST, meidan (IQR)	16 (14, 18)	21 (17, 26)
AST/ALT, meidan (IQR)	1.17 (0.94, 1.34)	0.8 (0.63, 1)

NAFLD, metabolic associated fatty liver disease; IQR, interquartile range.

the function.

$$\min_{\omega,b} \frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{L} \xi_i,$$
$$s.t.\ y_i \left(\omega^T X_i + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0;$$

We attached different weights to the two categories, i.e., the objective function was replaced by

$$\min_{\omega,b} \frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{L} w_i \xi_i.$$

Each ω_i for normal cases (NAFLD label $y_i = 1$) had a common value denoted by $\overline{\omega}$, while each ω_i for NAFLD cases (NAFLD label $y_i = -1$) had another common value denoted by $\underline{\omega}$. We adjusted the value of $\overline{\omega}$ and $\underline{\omega}$ based on particular cases. For practical problems, we take $\overline{\omega} > \underline{\omega}$ if we believe the risk induced by misclassifying a label -1 sample as label 1 is larger than that induced by misclassifying a label 1 sample as label 0. Otherwise, we take $\overline{\omega} < \underline{\omega}$.

We used the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm) to implement the soft margin SVM model. The Gauss kernel function was applied in our study, which gives the highest accuracy for our test. The receiver operating characteristic (ROC) curve was used to assess the predictive performance of our SVM models. We generated the ROC curve by drawing the true-positive rates vs. false-positive rates over a range of thresholds. Each threshold is a cutoff, if an individual's output probability in the SVM is greater than this cutoff, he is judged as NAFLD, otherwise, he is judged as non-NAFLD. For each threshold, we calculated a pair of true-positive rates and false-positive rates. When the thresholds ranged stepwise from 0 to 1 by step size 0.01, we obtained the whole ROC curve. The area under the curve (AUC) was used as a measure of the predictive performance of our SVM models. The following **Figure 1** is the technical line of machine learning.

RESULTS

The Results of Screen and Compare Diagnostic Indicators

Basic Statistical Analysis Results of TPA and PAI-1

In the dataset of 365 cases, the patients' ages ranged from 25 to 65 years old, 266 patients had NAFLD, and 99 patients were normal. We used a *t*-test to compare the TPA, PAI-1, and TPA/PAI-1 between the NAFLD group and the non-NAFLD group. TPA, PAI-1, and TPA/PAI-1 exhibited a significant difference ($P < 0.05$) between the two groups. The mean of TPA and PAI-1 in the NAFLD group was higher than that in the non-NAFLD group. However, the mean of the ratio TPA/PAI-1 in the non-NAFLD group was significantly higher than that in the NAFLD group ($P < 0.05$). The results are summarized in **Table 2**. The results showed that no matter whether it was the plasm level of TPA, PAI-1, or TPA/PAI-1, there were significant differences between the NAFLD and the non-NAFLD patients, which suggests that the plasma levels of TPA, PAI-1, or TPA/PAI-1 have the potential to be regarded as indicators for NAFLD diagnosis.

Predictive Results for NAFLD Using Metabolic Indicators as Predictors

First, we standardized the TC, HDL-C, LDL-C, and BMI data. In order to better assess the performance of the SVM predictive model for NAFLD, we first constructed the Logistic aggression model to predict NAFLD using the standard TC, HDL-C, LDL-C, and BMI data. The Logistic model was implemented in SPSS 25.0 but the predictive accuracy was <30%. Then we used the standard data to construct SVM predictive models for NAFLD. The results of the SVM model were summarized in **Table 3**. Error_1 was used to denote the misclassification rate of predicting normal samples as NAFLD samples and Error_2 was used to denote the misclassification rate of predicting NAFLD samples as normal samples. The results show that in the experiment, the accuracy of the SVM model is much higher than that of the Logistic model, suggesting the SVM model is more suitable for the predictive study.

Predictive Results for NAFLD Using Fibrinolytic Indicators as Predictors

As above, we first constructed the Logistic model using the standardized TPA and PAI-1 as predictors but found that the predictive accuracy was not more than 40%. Next, we constructed an SVM model using the standardized TPA and PAI-1 as input variables. And we found that the predictive accuracy was much higher than that of the Logistic model. The results are shown in **Table 4**. These results suggest that, similar to metabolic indicators, the use of the SVM model to predict fibrinolytic indicators is more accurate.

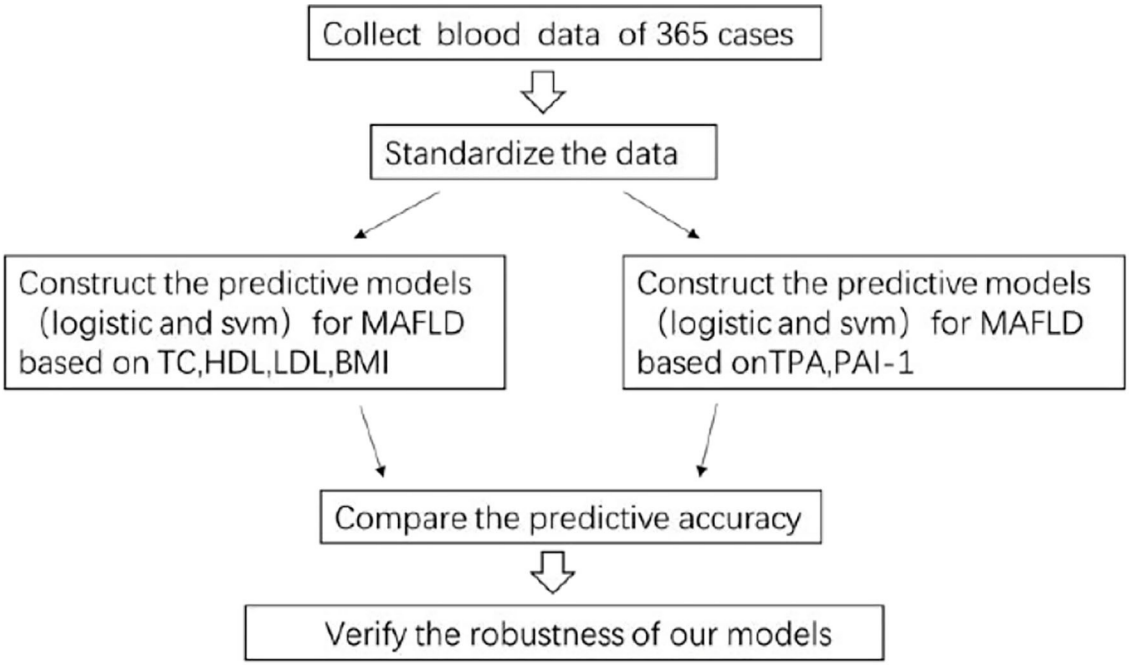

FIGURE 1 | The technical line of our research.

TABLE 2 | Comparison of TPA1, PAI-1 between the NAFLD and non-NAFLD patients.

	Group	Mean ± std	*t*-statistic	*P*
TPA1	NAFLD group	9,596.64 ± 4,190.25	7.76	0.000
	Non-NAFLD group	6,311.64 ± 3,344.76		
PAI-1	Group	Mean ± std	*t*-statistic	*p*
	NAFLD group	31,438.98 ± 8,124.59	22.16	0.000
	Non-NAFLD group	16,589.63 ± 4,461.35		
TPA/PAI-1	Group	Mean ± std	*t*-statistic	*p*
	NAFLD group	0.33 ± 0.18	−2.59	0.01
	Non-NAFLD group	0.40 ± 0.22		

TPA, plasma plasminogen activator; PAI-1, plasminogen activator inhibitor-1; Std, standard deviation.

TABLE 3 | Prediction performance using BMI, TC, HDL-C, and LDL-C as factors.

	Error_1	Error_2	Total accuracy
Training set	37%	5%	85.35%
Testing set	39%	8%	85.87%

BMI, body mass index; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol.

The Comparison of the Prediction of Metabolic and Fibrinolytic Indicators

Interestingly, we found that the predictive accuracy based on TPA and PAI-1 was significantly higher than that based on TC, HDL-C, LDL-C, and BMI. To better see the difference, we drew the two ROC curves (**Figure 2**). The red curve is the ROC curve using TPA and PAI as predictors and the AUC is 0.91; the blue curve is the ROC curve using TC, HDL-C, LDL-C, and BMI as predictors and the AUC is 0.75. The difference was obvious. From the above results, we inferred that TPA and PAI-1 are more suitable than TC, HDL-C, and LDL-C for predicting NAFLD. TPA and PAI-1 have deeper links with NAFLD than TC, HDL-C, and LDL-C do.

TABLE 4 | Prediction performance using TPA and PAI-1 as factors.

	Error_1	Error_2	Total accuracy
Training set	15%	4%	92.58%
Testing set	10%	8%	91.48%

TPA, plasma plasminogen activator.
PAI-1, plasminogen activator inhibitor-1.

We also want to know whether TC, HDL-C, and LDL-C can be complementary to TPA and PAI-1 to achieve better prediction results (in other words, whether TPA and PAI-1 miss some valuable information contained in the TC, HDL-C, and LDL-C data) when predicting NAFLD. Thus, we combined the TPA and PAI-1 data with the TC, HDL-C, and LDL-C data to construct an SVM model. The predictive results are in **Table 5**. The results show that compared with the prediction performance using BMI, TC, HDL-C, and LDL-C as factors, after adding TPA and PAI-1, the prediction accuracy of metabolic indicators is greatly improved. However, the prediction accuracy of the SVM model did not increase significantly compared with TPA and PAI-1 alone as a predictor. These indicate that the blood levels of TPA

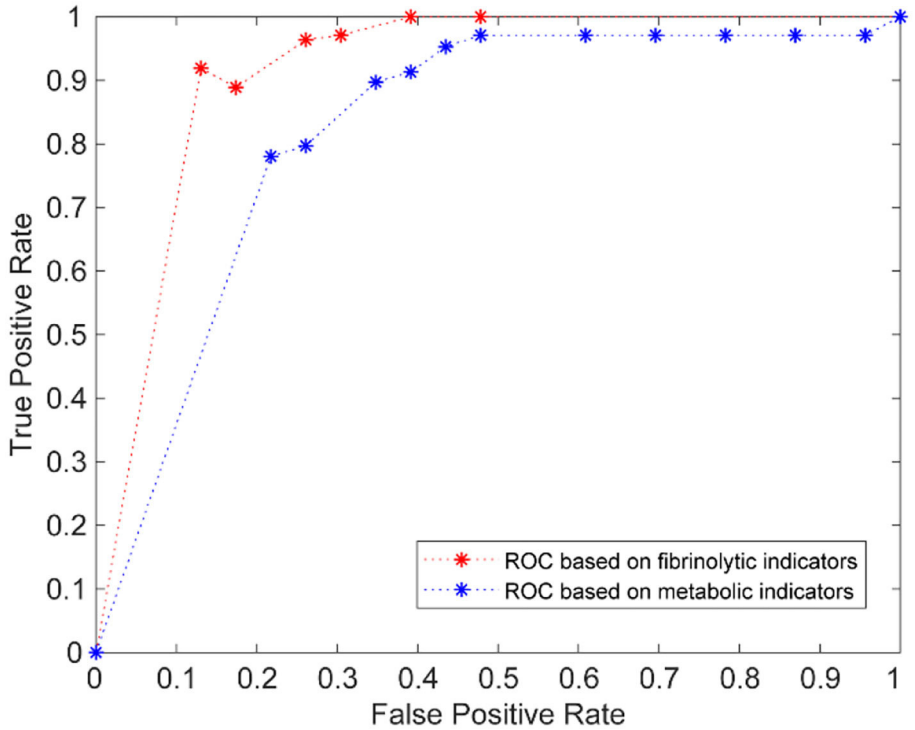

FIGURE 2 | The comparison of ROC curves of SVM based on fibrinolytic indicators and metabolic indicators.

TABLE 5 | Prediction performance using TPA, PAI-1, TC, HDL-C, LDL-C, and BMI as indicators.

	Error_1	Error_2	Total accuracy
Training set	13%	4%	93.57%
Testing set	8%	5%	92.65%

and PAI-1 can be regarded as highly effective indicators to assist the diagnosis of NAFLD, independent of metabolic indicators.

The Robustness of Our SVM Model

To check the robustness of our SVM model based on TPA and PAI-1, we took different percentages of training samples in the total 365 cases. The percentage varied from 25 to 75% and we obtained the corresponding accuracy of predicting NAFLD as in **Figure 3**.

The results show that in different percentages of training samples, the prediction accuracy of training samples and test samples are both high (over 90%), which indicates our SVM model based on TPA and PAI-1 was stable and trustable.

DISCUSSION

NAFLD is characterized by the significant accumulation of lipids, such as TG, TC, HDL- C, LDL-C in hepatocytes and serum, indicating that altered lipid metabolism is crucial in the pathogenesis of NAFLD (17). NAFLD is a broad-spectrum disease, including simple steatosis in the early stage, non-alcoholic steatohepatitis, liver fibrosis, cirrhosis, and even liver cancer in the late stages (18). The pathogenesis of NAFLD has been widely accepted by the "multiple-hit" hypothesis because NAFLD pathogenesis involves many influence factors, such as diet, genetic, environmental, and metabolism that progress through different stages during the occurrence and development of NAFLD (19). Although the number of patients with NAFLD is large and the harm is great, the exact mechanism of NAFLD is still unclear.

TPA and PAI-1 are mainly a pair of biological regulatory factors synthesized and secreted by vascular endothelial cells. Fibrinolytic system balance is affected by many factors, such as blood lipids, blood glucose, stress, gender, and age. And it is associated with obesity, insulin resistance, diabetes, dyslipidemia, and premature aging (13, 20, 21), which all are coexisting conditions of NAFLD. All this suggests that TPA and PAI-1 may be related to the metabolism and hepatic functions of NAFLD patients, but the specific mechanism is currently unknown.

What's more, by reviewing the literature, we found that the imbalance of TPA and PAI-1 activity is of great significance in metabolism, chronic liver disease and has different manifestations in different stages of the disease (22). And Based on our results, the prediction accuracy of NAFLD using TPA and PAI-1 as predictors was higher than that

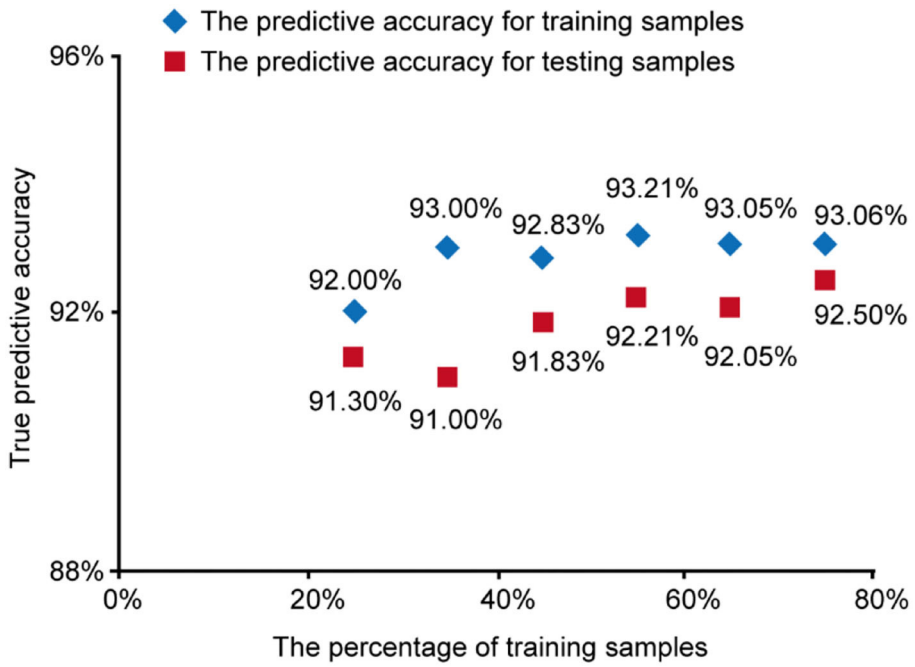

FIGURE 3 | The predictive accuracy for training samples and testing samples vs. the percentage of training samples.

using TC, HDL-C, LDL-C, and ALT/AST as predictors. These discoveries all further suggest that the plasma level of TPA and PAI-1 may be used as new indicators for the diagnosis of NAFLD.

Nowadays, the gold standard for the NAFLD diagnosis is the liver biopsy (23), But liver biopsy cannot be used routinely, since it is an invasive and expensive procedure. In clinical diagnosis, we often use liver B ultrasound combined with clinical symptoms and metabolic indicators to diagnose NAFLD (24). Through the study, we propose that changes to the fatty liver fibrinolytic system are one of the key links in NAFLD progress. The change to the fibrinolytic system was even more significant for NAFLD than the internal metabolic indices such as liver and kidney function. Therefore, we propose that TPA and PAI-1 should be included in normal physical examinations. Further, studies of fibrinolytic activity and drug development may be important for understanding the mechanism and treatment of NAFLD. Based on the perspective of the fibrinolytic system, in-depth discussion on its prediction of NAFLD may play an important role in improving the mechanism of NAFLD.

However, this study also has some shortcomings. In this observation object, our inclusion criteria are B ultrasound diagnosis, so it is difficult to distinguish the stratification of NAFLD disease and Unable to analyze changes in the fibrinolytic system during the disease progression. Therefore, in the following study, we look forward to using H1-MRS, controlled attenuation parameter, through human or animal and cell experiments to analysis of its internal mechanism.

CONCLUSION

In summary, TPA and PAI-1 are also effective indicators for the Chinese to assist in the diagnosis of NAFLD. Its diagnostic accuracy may be higher than metabolic related indicators. We do hope that this study can promote the further development of clinical NAFLD diagnosis and provide valuable guidance for the non-invasive diagnosis of NAFLD.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the committee for the protection of human subjects of the First Affiliated Hospital, Zhejiang Chinese Medical University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

CW, JY, and YN participated in drafting the manuscript. SZ, YX, and YN provided technical assistance. CW and JY revised

the manuscript. All of the authors read and approved the final manuscript.

FUNDING

This work was supported by Natural Science Fundation of Zhe Jiang Province (No. LY21H270009); Chinese Medicine Science and Technology Plan of Zhejiang Province (2021ZA062 and 2020ZB065).

ACKNOWLEDGMENTS

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.771219/full#supplementary-material

REFERENCES

1. Yki-Järvinen H. Diagnosis of non-alcoholic fatty liver disease (NAFLD). *Diabetologia.* (2016) 59:1104–11. doi: 10.1007/s00125-016-3944-1
2. Younossi Z M, Koenig A B, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology.* (2016) 64:73–84. doi: 10.1002/hep.28431
3. Lee HW, Wong VW. Changing NAFLD epidemiology in China. *Hepatology.* (2019) 70:1095–8. doi: 10.1002/hep.30848
4. Eslam M, Sanyal AJ, George J. MAFLD: A consensus-driven proposed nomenclature for metabolic associated fatty liver disease. *Gastroenterology.* (2020) 158:1999–2014.e1. doi: 10.1053/j.gastro.2019.11.312
5. Oral A, Sahin T, Turker F, Kocak E. Relationship between serum uric acid levels and nonalcoholic fatty liver disease in non-obese patients. *Medicina (Kaunas).* (2019) 55:600. doi: 10.3390/medicina55090600
6. Padsalgi G, Chooracken M, Mukkada R, Chettupuzha A, Francis J, Augustine P, et al. Correlation of BMI with Subcutaneous Fat and Total Body Fat in Patients with NAFLD in South India. *J Clin Exper Hepatol.* (2016) 6:S30. doi: 10.1016/j.jceh.2016.06.055
7. Tang Z, Pham M, Hao Y, Wang F, Patel D, Jean-Baptiste L, et al. Sex, age, and BMI modulate the association of physical examinations and blood biochemistry parameters and NAFLD: a retrospective study on 1994 cases observed at shuguang hospital, China. *Biomed Res Int.* (2019) 2019:1246518. doi: 10.1155/2019/1246518
8. Ye Q, Zou B, Yeo Y H, Li J, Huang D Q, Wu Y, et al. Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol.* (2020) 5:739–52. doi: 10.1016/S2468-1253(20)30077-7
9. Zheng Z, Nakamura K, Gershbaum S, Wang X, Thomas S, Bessler M, et al. Interacting hepatic PAI-1/tPA gene regulatory pathways influence impaired fibrinolysis severity in obesity. *J Clin Invest.* (2020) 130:4348–59. doi: 10.1172/JCI135919
10. Peng S, Xue G, Chen S, Chen Z, Yuan C, Li J, et al. tPA point mutation at autolysis loop enhances resistance to PAI-1 inhibition and catalytic activity. *Thromb Haemost.* (2019) 119:77–86. doi: 10.1055/s-0038-1676518
11. Wang J, Yuan Y, Cai R, Huang R, Tian S, Lin H, et al. Association between plasma levels of PAI-1, tPA/PAI-1 molar ratio, and mild cognitive impairment in Chinese patients with type 2 diabetes mellitus. *J Alzheimers Dis.* (2018) 63:835–45. doi: 10.3233/JAD-171038
12. Winter MP, Kleber ME, Koller L, Sulzgruber P, Scharnagl H, Delgado G, et al. Prognostic significance of tPA/PAI-1 complex in patients with heart failure and preserved ejection fraction. *Thromb Haemost.* (2017) 117:471–8. doi: 10.1160/TH16-08-0600
13. Jin R, Krasinskas A, Le NA, Konomi JV, Holzberg J, Romero R, et al. Association between plasminogen activator inhibitor-1 and severity of liver injury and cardiovascular risk in children with non-alcoholic fatty liver disease. *Pediatr Obes.* (2018) 13:23–9. doi: 10.1111/ijpo.12183
14. Chang ML, Hsu CM, Tseng JH, Tsou YK, Chen SC, Shiau SS, et al. Plasminogen activator inhibitor-1 is independently associated with non-alcoholic fatty liver disease whereas leptin and adiponectin vary between genders. *J Gastroenterol Hepatol.* (2015) 30:329–36. doi: 10.1111/jgh.12705
15. Spann A, Yasodhara A, Kang J, Watt K, Wang B, Goldenberg A, et al. Applying machine learning in liver disease and transplantation: a comprehensive review. *Hepatology.* (2020) 71:1093–105. doi: 10.1002/hep.31103
16. Kim J W, Ye Q, Forgues M, Chen Y, Budhu A, Sime J, et al. Cancer-associated molecular signature in the tissue samples of patients with cirrhosis. *Hepatology.* (2004) 39:518–27. doi: 10.1002/hep.20053
17. Jarvis H, Craig D, Barker R, Spiers G, Stow D, Anstee Q M, et al. Metabolic risk factors and incident advanced liver disease in non-alcoholic fatty liver disease (NAFLD): a systematic review and meta-analysis of population-based observational studies. *PLoS Med.* (2020) 17:e1003100. doi: 10.1371/journal.pmed.1003100
18. Sutti S, Albano E. Adaptive immunity: an emerging player in the progression of NAFLD. *Nat Rev Gastroenterol Hepatol.* (2020) 17:81–92. doi: 10.1038/s41575-019-0210-2
19. Marchisello S, Di Pino A, Scicali R, Urbano F, Piro S, Purrello F, et al. Pathophysiological, molecular and therapeutic issues of nonalcoholic fatty liver disease: an overview. *Int J Mol Sci.* (2019) 20:1948. doi: 10.3390/ijms20081948
20. Urano T, Suzuki Y, Iwaki T, Sano H, Honkura N, Castellino FJ. Recognition of plasminogen activator inhibitor type 1 as the primary regulator of fibrinolysis. *Curr Drug Targets.* (2019) 20:1695–701. doi: 10.2174/1389450120666190715102510
21. Kanno Y. The role of fibrinolytic regulators in vascular dysfunction of systemic sclerosis. *Int J Mol Sci.* (2019) 20:619. doi: 10.3390/ijms20030619
22. Beier J I, Arteel G E. Alcoholic liver disease and the potential role of plasminogen activator inhibitor-1 and fibrin metabolism. *Exp Biol Med (Maywood).* (2012) 237:1–9. doi: 10.1258/ebm.2011.011255
23. Machado MV, Cortez-Pinto H. Non-alcoholic fatty liver disease: what the clinician needs to know. *World J Gastroenterol.* (2014) 20:12956–80. doi: 10.3748/wjg.v20.i36.12956
24. Chalasani N, Younossi Z, Lavine J E, Diehl A M, Brunt E M, Cusi K, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology. (2012) 142:1592–609. doi: 10.1053/j.gastro.2012.04.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JL declared a shared affiliation, with one of the author CW to the handling editor at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 29 June 2022
doi: 10.3389/fmed.2022.762091

InterNet: Detection of Active Abdominal Arterial Bleeding Using Emergency Digital Subtraction Angiography Imaging With Two-Stage Deep Learning

Xiangde Min[1†], *Zhaoyan Feng*[1†], *Junfeng Gao*[2†], *Shu Chen*[3], *Peipei Zhang*[1], *Tianyu Fu*[3], *Hong Shen*[3] *and Nan Wang*[1*]

[1] *Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China,* [2] *College of Biomedical Engineering, South-Central of University for Nationalities, Wuhan, China,* [3] *United Imaging Intelligence, Shanghai, China*

OPEN ACCESS

Edited by:
Jingjing You,
The University of Sydney, Australia

Reviewed by:
Jie-Zhi Cheng,
ShanghaiTech University, China
Adriaan Dees,
Ikazia Hospital, Netherlands

***Correspondence:**
Nan Wang
southernwang@sina.com

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *20 August 2021*
Accepted: *25 May 2022*
Published: *29 June 2022*

Citation:
Min X, Feng Z, Gao J, Chen S, Zhang P, Fu T, Shen H and Wang N (2022) InterNet: Detection of Active Abdominal Arterial Bleeding Using Emergency Digital Subtraction Angiography Imaging With Two-Stage Deep Learning. Front. Med. 9:762091. doi: 10.3389/fmed.2022.762091

Objective: Active abdominal arterial bleeding is an emergency medical condition. Herein, we present our use of this two-stage InterNet model for detection of active abdominal arterial bleeding using emergency DSA imaging.

Methods: Firstly, 450 patients who underwent abdominal DSA procedures were randomly selected for development of the region localization stage (RLS). Secondly, 160 consecutive patients with active abdominal arterial bleeding were included for development of the bleeding site detection stage (BSDS) and InterNet (cascade network of RLS and BSDS). Another 50 patients that ruled out active abdominal arterial bleeding were used as negative samples to evaluate InterNet performance. We evaluated the mode's efficacy using the precision-recall (PR) curve. The classification performance of a doctor with and without InterNet was evaluated using a receiver operating characteristic (ROC) curve analysis.

Results: The AP, precision, and recall of the RLS were 0.99, 0.95, and 0.99 in the validation dataset, respectively. Our InterNet reached a recall of 0.7, the precision for detection of bleeding sites was 53% in the evaluation set. The AUCs of doctors with and without InterNet were 0.803 and 0.759, respectively. In addition, the doctor with InterNet assistant could significantly reduce the elapsed time for the interpretation of each DSA sequence from 84.88 to 43.78 s.

Conclusion: Our InterNet system could assist interventional radiologists in identifying bleeding foci quickly and may improve the workflow of the DSA operation to a more real-time procedure.

Keywords: abdominal arterial bleeding, digital subtraction angiography, deep learning, automatic detection, two-stage model

INTRODUCTION

Active abdominal arterial bleeding is a medical emergency that may lead to haemorrhagic shock or circulatory instability if left untreated (1–5). Clinicians experience difficulty in dealing with this complicated condition (2, 5, 6). Most cases of active abdominal arterial bleeding are medically treated by correcting coagulation abnormalities or through endoscopy (7–9). Nonetheless, these methods can fail in some patients with significant bleeding, in which cases endovascular treatment is desired (3, 10–14). Due to its advantages of reduced morbidity and mortality, endovascular treatment using digital subtraction angiography (DSA) is now preferred over open surgery (5, 11, 15–17).

Rapid and accurate diagnosis of arterial bleeding by an interventional physician *via* DSA remains challenging (1). Human limitations in a crowded tertiary hospital include staff shortage, excess workload, and, especially, a lack of knowledge among radiologists regarding arterial bleeding. Under these circumstances, an automated system is needed to alleviate the tedious task of screening out incidental findings and allowing physicians more time to interact with patients and other health care providers. Further, such a system would help address the lack of expert radiologists in rural and community hospitals. What's more, the bleeding could be subtle in some cases. It is difficult to identify subtle bleeding by human quickly, and it is more difficult for junior doctor. So, one of the important values of our system is to shorten diagnosis time and to reduce the rate of missed bleeding sites. Deep learning approaches have provided exciting solutions medical image in medical image detection. The diagnosis of bleeding involves a typical computer visual task of classification of radiological images into bleeding and non-bleeding categories and detection of bleeding sites. However, computer-assisted automated detection of active abdominal arterial bleeding from DSA images has not been previously reported.

In current practice, a captured video sequence is reviewed offline by the physician to identify bleeding sites before the intervention is performed. A usable AI (artificial intelligence) system should be able to replace this offline review with automated detection of bleeding sites. Thus, our system was designed and evaluated based on this first goal. However, the current workflow must ultimately be improved to a more real-time system ideally. If the automated system detects bleeding sites correctly in most frames and at the video frame rate, there might be no need for an offline review. The physician could directly view the highlighted bleeding sites in real-time and perform the surgery, which would reduce the surgery time. In this work, we proposed a two-stage deep learning model (named *InterNet*) for real-time detection of active abdominal arterial bleeding using emergency DSA imaging. We hypothesized that the InterNet can detect active abdominal arterial bleeding at a faster speed and higher sensitivity.

MATERIALS AND METHODS

Data Acquisition

Firstly, 450 patients who underwent abdominal DSA procedures were randomly selected from our PACS system for development of the region localization stage (RLS). Secondly, 160 consecutive patients with active abdominal arterial bleeding who underwent endovascular treatment between January 2013 and January 2020 were retrospectively included for development of the bleeding site detection stage (BSDS) and InterNet (cascade network of RLS and BSDS). These 160 patients had clinical signs of active abdominal arterial bleeding: blood from a postoperative drainage tube, haematuria, haematochezia, hypotension, tachycardia, or a low hemoglobin level. Another 50 patients who underwent abdominal DSA procedures that ruled out active abdominal arterial bleeding were randomly selected and used as negative samples to evaluate InterNet performance.

A standard transfemoral approach was used in all angiographic procedures. A sheath introducer was placed in the right or left common femoral artery using the Seldinger technique. Selective angiography of the abdominal aortic branches was performed using a 5-Fr catheter in all patients. Super selective angiography of the tiny branches was performed using a microcatheter.

DSA images usually contain multiple sequences, and each sequence consisted of 30–50 video frames at six frames per second. All data were stored in Digital Imaging and Communications in Medicine (DICOM) format. All data were manually annotated using LabelImge software (GitHub, Inc., San Francisco, CA, USA). The bleeding sites and angiographic regions were manually segmented and annotated by two radiologists. The segmented images were then reviewed by another experienced radiologist. Any disagreements in segmentation were resolved through consensus among the three radiologists.

Dataset Splitting

A total of 546 sequences from 450 patients were used for RLS development. These patients were randomly split into a training dataset (80%) and a validation dataset (20%). From the 160 patients with active abdominal arterial bleeding, 182 sequences from 90 patients were classified into the BLDS training dataset; 49 sequences from 20 patients were classified as a validation dataset for stability and generalizability of the RLS and BSDS cascade network (InterNet). Sixty-seven sequences from 50 actively bleeding patients and 80 sequences from 50 patients without active bleeding were classified as an independent testing dataset for InterNet.

Deep Learning Model Development

The entire program was performed with Pytorch version 1.2 (Pytorch, Warsaw, Mazowieckie, Poland) as the backend, on a desktop computer equipped with an Intel (R) Xeon(R) Silver 4110 system (Intel Inc., Santa Clara, CA, USA), 64 GB RAM, and a GeForce RTX 2080Ti GPU (Nvidia, Santa Clara, CA, USA). The

InterNet detection system was developed to automatically detect bleeding sites on DSA images using a two-stage process, first localizing the angiographic region from the original frame image (RLS), followed by bleeding site detection on the cropped image (BSDS). The framework of our two-stage detection system is schematized in **Figure 1**. The RLS was based on the sparseness of bleeding sites in a sequence and within a frame image. ResNet50 was used as the backbone for our two-stage deep learning model framework.

Multi-scale features were extracted to create feature maps, and the region proposal networks (RPNs) were applied to generate region proposals *via* classification and regression (18). The proposed regions underwent non-maximum suppression to filter the highly overlapping regions. Region pooling unified the various-sized regions to the same size. The resulting region candidates were put through the Region Based Convolutional Neural Networks (R-CNN). The targets were classified, and the bounding boxes underwent a second regression to achieve the final target detection. Moreover, we applied the feature pyramid networks (FPN) on the framework of our two-stage detection system (19). The cost and benefit of using the FPN compared to the approach without FPN was also evaluated. The detailed network structure of multi-scale features extraction is schematized in **Figure 2**.

To tune the detection system, we adjusted the size of the input image. According to the detection performance, we chose the optimal value of the key parameter of "resize." To avoid overfitting, we used common techniques to augment the data. Contrast-limited adaptive histogram equalization (CLAHE) was applied to reduce the intensity range, followed by random shift and rotation to augment the orientation and position of the bleeding site samples (20). Perturbation of intensities and

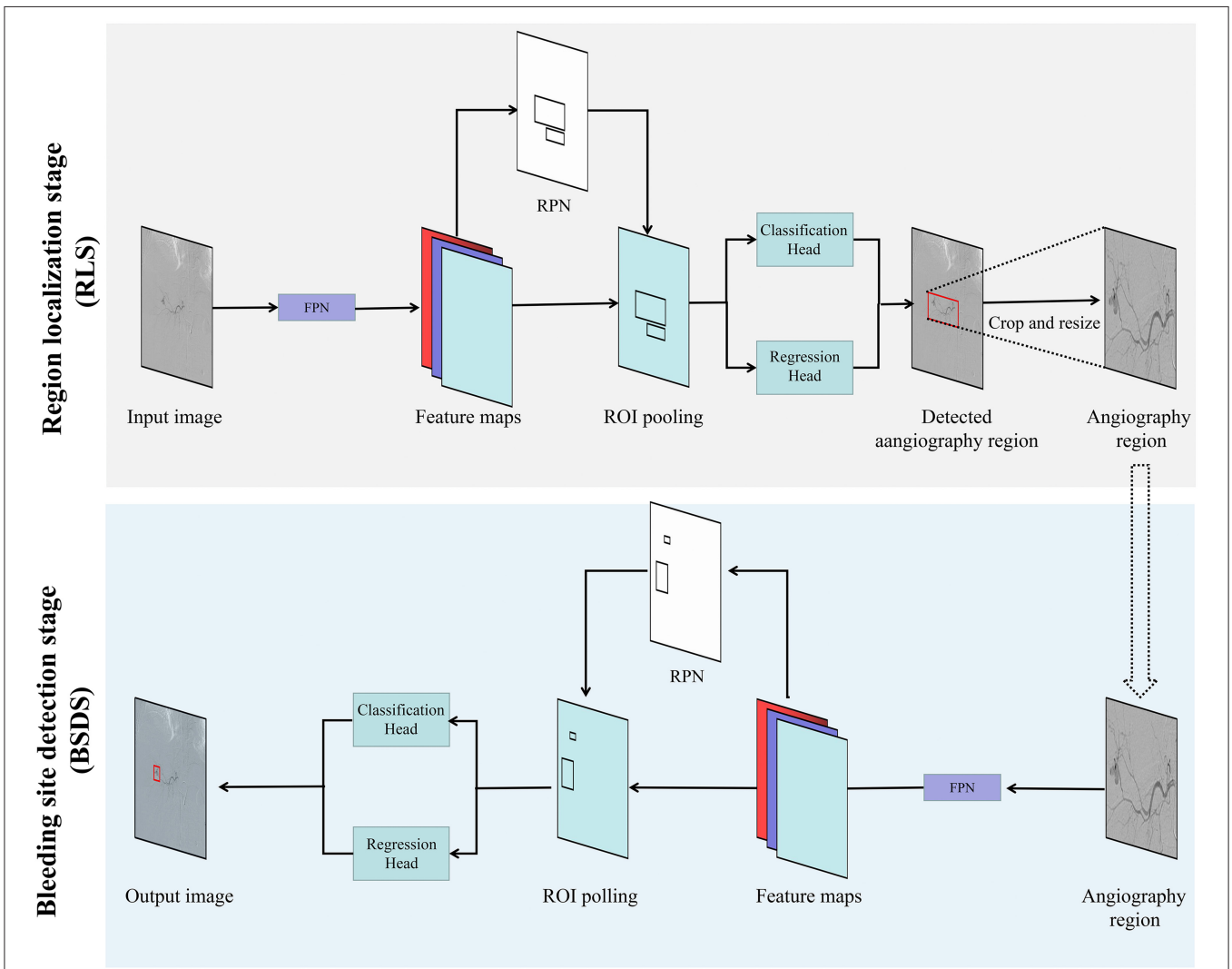

FIGURE 1 | Overview diagram of the proposed two-stage deep learning approach (InterNet). The system first detected the angiographic region from the original frame image. The output of the RLS was used as input in the next stage of redundancy reduction. RPN, region proposal network; ROI, region of interest; FC, fully connected layer; Bbox, bounding box.

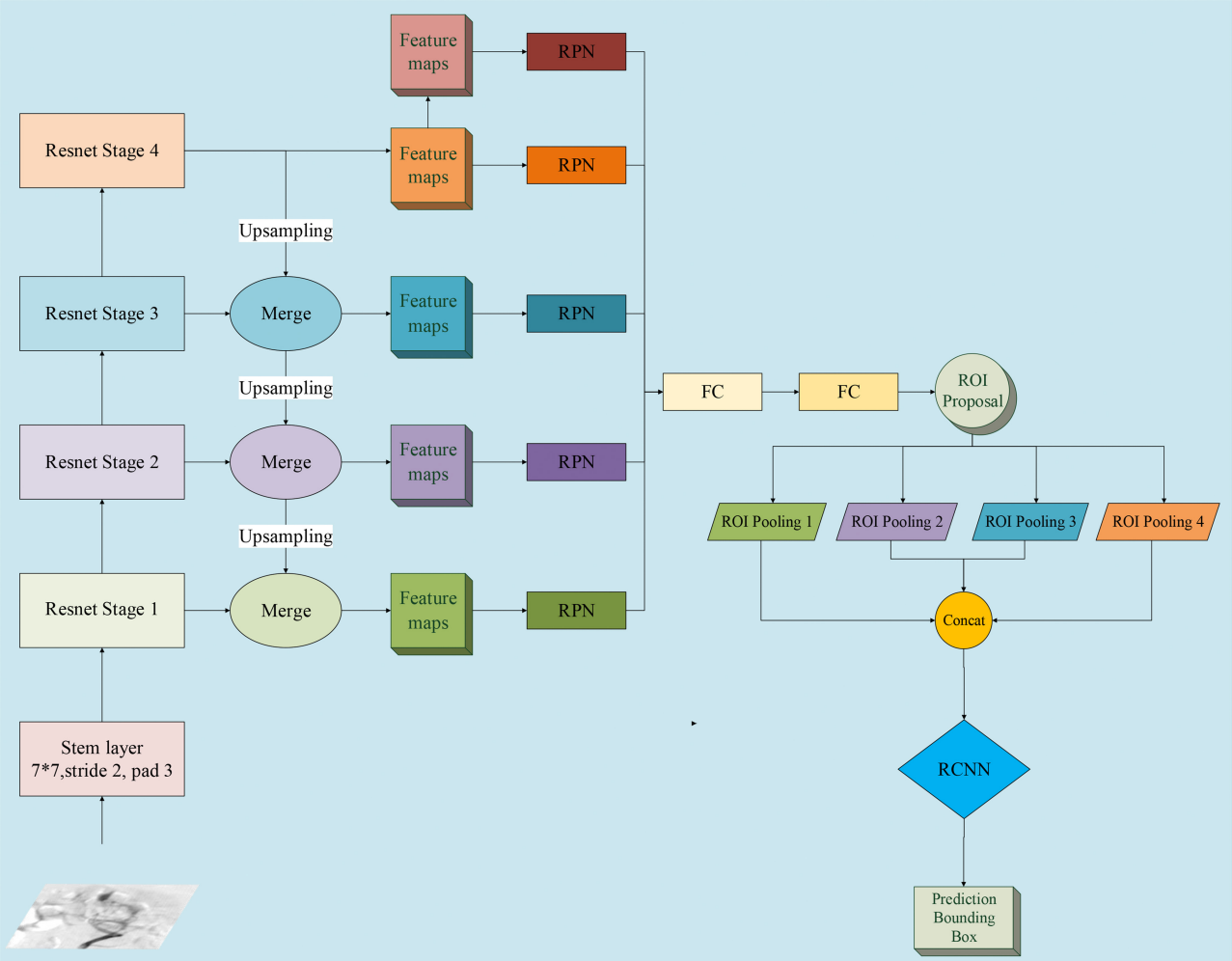

FIGURE 2 | Detailed network structure diagram of feature pyramid network (FPN). We adopted FPN with a ResNet-50 backbone for InterNet. FPN is an outstanding detector for many visual tasks and can identify objects at multi-scales. RPN, region proposal network; ROI, region of interest; FC, fully connected layer; bbox, bounding box; RCNN, Region Based Convolutional Neural Networks.

contrast and a random median filter were applied to improve the distribution of the samples.

Performance Assessment Between Doctors With and Without InterNet Assistant

To evaluate the benefit of InterNet, we compared the classification efficiency between doctors in terms of patients with InterNet assistant using the independent testing dataset. The classification performance and elapsed time were recorded.

Statistical Analyzes

We evaluated the model's efficacy using the precision-recall (PR) curve, which is commonly used to show the compromise between precision and recall. By moving along the curve, various compromises between precision and recall can be acquired, enabling us to choose between the two. A high recall indicates a higher rate of detection (fewer false negatives), and a high precision indicates a lower rate of false positives. The average precision (AP) was used to evaluate the detection precision of the deep learning algorithms. A prediction is considered to be true positive if Intersection over Union (IoU) > 0.5, and false positive if IoU < 0.5. The frame-per-second (FPS) rate of each test was calculated to evaluate whether the bleeding sites could be tracked in real-time. The classification performance of a doctor with and without InterNet was evaluated using a receiver operating characteristic (ROC) curve analysis. The area under the curve (AUC), sensitivity, and specificity were calculated. The differences in elapsed time for a doctor with and without InterNet were compared using the Mann-Whitney *U*-test. Statistical analyses were performed using R (version 3.3.4, http://www.

TABLE 1 | Effect of feature pyramid networks of the detection system.

	Baseline	Baseline + FPN
AP (%)	58.1	60.3
FPS	4.2	5.0

AP, Average Precision; FPS, frames per second.

TABLE 2 | Effects of changing resize parameters of the detection system.

	Baseline + FPN (Resize 1,024 × 512)	Baseline + FPN (Resize 1,333 × 800)
AP (%)	60.3	64.5
Precision	0.514	0.531
Recall	0.683	0.703
FPS	5.0	3.9

FPN, Feature Pyramid Network; AP, Average Precision; FPS, frames per second.

Rproject.org). The threshold for statistical significance was set at a two-sided $p < 0.05$.

RESULTS

We used the area calculated from the segmented mask of each positive DSA image to represents the amount of bleeding. The mean bleeding site area of the 67 sequences from 50 actively bleeding patients in the independent testing dataset is 938.4 ± 1,707.1 square millimeter. Among the 50 patients, the bleeding locations of 24 cases are in kidney; 19 cases are in digestive tract; three cases are in spleen; two cases are in uterus; and two cases are in other organs.

The AP, precision, and recall of the RLS were 0.99, 0.95, and 0.99, respectively. This means that the angiographic region could be correctly recognized in 99 out of 100 testing images. The P-R curve of the RLS on the validation dataset is shown in **Supplementary Figure 1**.

The baseline system showed an AP of 58.1% and FPS rate of 4.2, while the network with FPN showed improved AP of 60.3% and FPS rate of 5.0. The detection results for the system on the validation dataset with and without FPN are shown in **Table 1**. The key parameter of "resize" was found to be optimal at 1,333 × 800. The AP reached 64.5% with this input image size, while the FRS showed a slight decrease from 5.0 to 3.9. The model including Baseline + FPN and resize 1,333 × 800 was selected as the final structure for our InterNet system. The effect of changing the resize parameters of the detection system is shown in **Table 2**. The InterNet P-R curve for the evaluation dataset is shown in **Figure 3**. For the task of detection, a high recall was more desirable than a high precision. Therefore, we picked a spot with a recall of 0.7 and precision of 0.53.

Table 3 and **Figure 4** summarize the classification performance of a doctor with and without InterNet. The doctor with InterNet showed a superior performance to that of the doctor without InterNet. The AUCs of doctors with and without

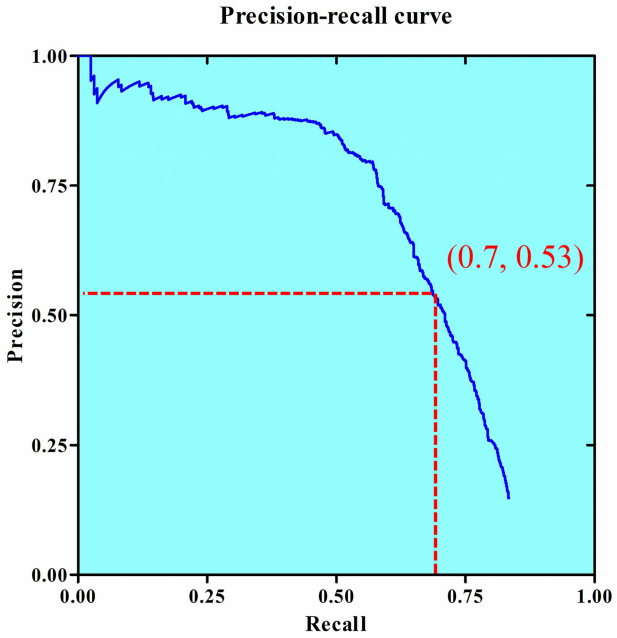

FIGURE 3 | The precision–recall curve of the InterNet. The average precision of the InterNet was 64.5 on the evaluation dataset. We selected a spot with a recall of 0.7 and a precision of 0.53.

TABLE 3 | Classification performance of a doctor with and without InterNet.

	Doctor with InterNet assistant	Doctor without InterNet assistant
AUC	0.803	0.759
Sensitivity (%)	88.06	73.13
Specificity (%)	72.50	78.75
Accuracy (%)	80	76
PPV (%)	73.00	74.00
NPV (%)	88.00	78.00
Time (second/sequence)	43.78	84.88

AUC, area under the receiver operating characteristic curve; PPV, positive predict value; NPV, negative predict value.

InterNet were 0.803 and 0.759, respectively. In particular, the doctor with InterNet assistant showed a substantially increased sensitivity, from 73.17 to 88.06%. In addition, the doctor with InterNet assistant could significantly reduce the elapsed time for the interpretation of each DSA sequence from 84.88 to 43.78 s per sequence ($p < 0.01$; **Figure 5**). Examples of the prediction results obtained by our proposed InterNet are shown in **Figure 6**.

DISCUSSION

Given the efficacy and safety of transcatheter arterial embolization compared with open surgery for the treatment of active abdominal arterial bleeding (5, 11, 15–17), accurate and rapid detection of bleeding sites is the key to success of transcatheter arterial embolization. In this study, we built an

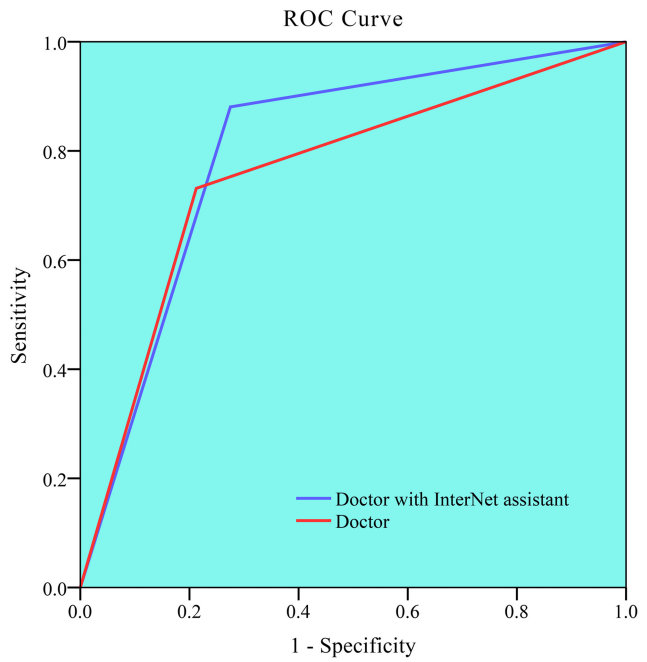

FIGURE 4 | Receiver operating characteristic (ROC) curve for doctor without and doctor with InterNet assistant. Doctor with InterNet assistant showed a superior performance to that of doctor without InterNet assistant. The AUCs of doctor with and doctor without InterNet were 0.803 and 0.759, respectively.

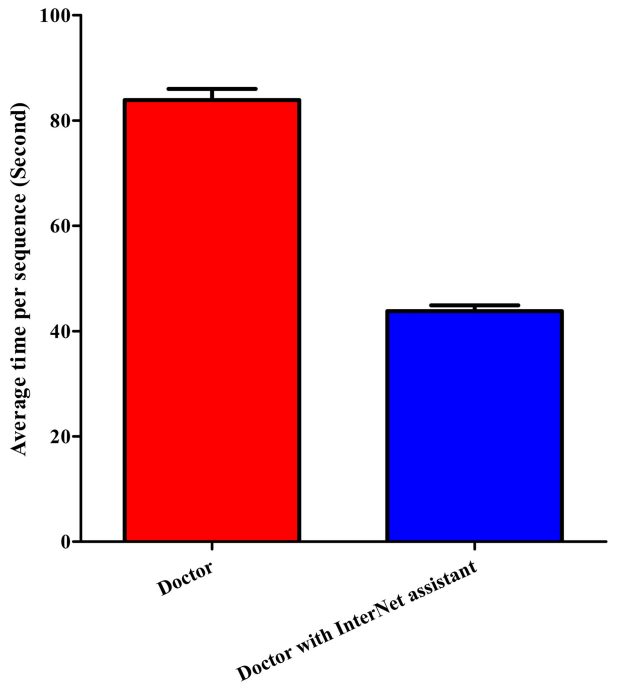

FIGURE 5 | Elapsed time of doctor without and doctor with InterNet assistant. Doctor with InterNet assistant significantly reduced the elapsed time for the interpretation of each DSA sequence from 84.88 to 43.78 s per sequence.

automated system based on a deep neural network model to detect active abdominal arterial bleeding on DSA images. Our InterNet system could help doctors in making a faster and more accurate interpretation. To our knowledge, this is the first system to automatically detect active arterial bleeding sites in DSA images.

In this study, we adopted two-stage deep learning for detection of active abdominal bleeding sites. In RLS, the angiographic region is proposed for detecting potential bleeding sites. In this stage, our detection system located a specific region to reduce the interference from other regions. The output of the RLS was used as input in the next stage of redundancy reduction. A practical benefit of RLS is that any data sequence, whether positive or negative, can be used for training the network for angiographic region extraction. This along with the ease of ROI labeling creates ample data to train a robust algorithm to extract the angiographic regions from the original frame images.

In the current study, we adopted FPN with a ResNet-50 backbone for the InterNet, because FPN is an outstanding detector for many visual tasks. FPN is capable of multi-scale feature extraction, which fits well with the task of detecting bleeding sites that have large variations in their sizes and shapes. FPN has a top-to-bottom pathway in addition to the bottom-to-top pathway of a regular neural network; hence, the semantic information from the top levels helps enhance the detailed information in the lower layers, leading to a powerful multi-scale capacity (19). In our study, the baseline system with FPN showed a relatively higher compared without FPN (60.3 vs. 58.1%).

Embolization requires the localization of bleeding sites, which can be easily missed by a physician. For our abdominal bleeding detection task, a low false negative rate is more desirable than a low false positive rate, since for a physician, it is easy to miss both, a bleeding spot and to rule out one. For this task of detection, a high recall was more desirable than a high precision. Therefore, we picked a spot with a recall of 0.7. At a recall of 0.7, the precision for detection of bleeding sites was 53% in the evaluation set. A recall of 70% means that the bleeding spot will be revealed to the physician in two of the three frames. A precision of 53% means that, on an average, for every correctly detected bleeding site, there will be <1 false detection. This false positive rate should be acceptable and not divert much of the physician's attention. The doctor with InterNet performed superiorly to the doctor without InterNet. In addition, the doctor with InterNet assistant could significantly reduce the elapsed time for the interpretation of each DSA sequence.

In present-day DSA surgery, after the DSA sequences are acquired, the physician reviews the sequence offline to detect the bleeding sites before performing the intervention. This workflow does not require our system to have real-time performance to replace the physician's effort of detecting bleeding sites. A more advanced use of the system would be to improve the workflow of the DSA operation to a more real-time procedure, thus eliminating the need for offline review and discussion. Ideally, the physician would look at the overhead monitor and observe the DSA images with overlaid marks of automatically detected bleeding sites (**Supplementary Figure 2**). To achieve this, it is

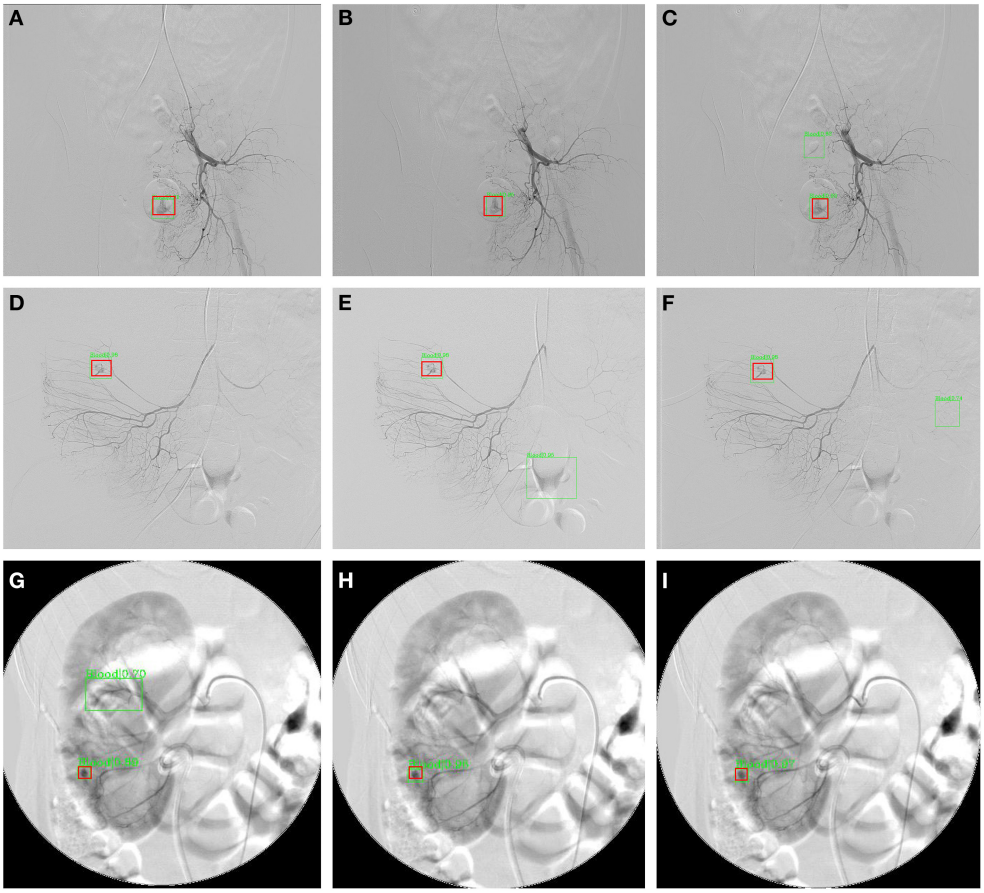

FIGURE 6 | Three sample results from InterNet. **(A–C)** Are images from a patient with subrectal arterial branch bleeding; **(D–F)** are images from a patient with bleeding from a right colonic artery branch; **(G–I)** are images from a patient with a bleeding branch of the inferior artery of the right kidney. The red boxes represent the ground truth bleeding bounding boxes; the green box represents the detected bleeding bounding box.

best for the system to reach six frames per second—the frame rate of the captured imaging sequence. The frame rate achieved with Python in this study is close to 4 frames per second, and it is conceivable that a product based on optimized C++ code should reach six frames per second without much difficulty. In such a system, the physician could watch the video sequence in real-time. The system will produce some false positives, with an average of one false positive every two frames due to a precision of 53%. The bleeding sites in ∼1 of 3 frames will not be marked due to a recall rate of 70%. Despite these imperfections, at an FRS of 4, the physician should be able to mentally make up the gap frames and eliminate the false marks with ease.

In recent years, many deep learning approaches have been developed for medical imaging analysis (21–23). A few studies have applied deep learning in DSA imaging. Alexander et al. trained a CNN system to automatically segment saccular aneurysms (pre- or post-coiling) and surrounding vasculature from DSA images (24). Yufen used residual density to generate a DSA image from a single live image without mask data acquisition, thus avoiding the appearance of motion artifacts in the image (25). To date, no study has applied the deep learning system for the detection of bleeding in DSA. We suspect that the difficulty in obtaining sufficient data is an important factor limiting its application to DSA. In this study, we applied deep learning for the detection of bleeding on DSA for the first time. Further applications of deep learning in DSA should be proposed and evaluated in future work. Deep learning may play an important role in surgery.

The main clinical applications of the proposed method are as followings. First, with the help of the current system, the physician would reduce the rate of missed bleeding sites, especially when the bleeding is subtle. Missed bleeding sites could lead to poor outcomes, and some patients may need a second procedure. Therefore, our system has the potential to improve the prognosis of patients. Second, the deep learning system developed in this study has the potential to shorten operation time, which may also reduce the radiation dosage to doctors and patients during the operation (26). Third, the automated system in our study would help address the lack of expert radiologists in rural and community hospitals. The CT Angiography (CTA) is also a common method for diagnosing active bleeding abdomen bleeding. Compared to CTA, DSA play import roles not only in diagnosis but also in appropriate management of abdomen bleeding. Most of the cases included in current study were

performed emergency DSA surgery, therefor very few patients underwent CTA before DSA due to limited time. For the 50 actively bleeding patients in the independent testing dataset for InterNet, only seven patients underwent CTA before DSA surgery. The two radiologists did not view their CTA results when identify the bleeding sites. Thus, the CTA examination would not influence the research results in this retrospective study.

This study has several limitations. Firstly, this was a retrospective study design at a single institute. The number of images with bleeding sites included in the test set was also not very large. Secondly, only DSA imaging of one manufacture was included. Based on current results, we could not sure whether the method could be generalized to various DSA sequences from various manufactures. Therefore, our system should be validated in multicentre studies of a larger scale. Thirdly, only abdominal bleeding was included. Other types of bleeding (neck or thoracic bleeding) were not included because of their low incidences.

In this study, we presented a two-stage model InterNet for active abdominal bleeding detection using deep learning with DSA data. This work has created a usable system to automatically detect bleeding sites in DSA sequences. Our developed InterNet system could help doctors in achieving a faster and more accurate interpretation. A prospective clinical trial is necessary to determine the effectiveness of this system and whether it will ultimately improve patient care and outcomes.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of Tongji Hospital of Huazhong University of Science and Technology. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XM, ZF, and NW: conception and design. PZ and XM: collection and assembly of data. SC, TF, HS, and JG: data analysis and interpretation. XM, ZF, JG, and NW: manuscript writing. All authors: final approval of manuscript.

FUNDING

This paper was supported by the National Natural Science Foundation of China under Grant No. 81801668 and 61773408.

ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of Wenzhi Lv and Leigang Sun for DICOM data acquisition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.762091/full#supplementary-material

REFERENCES

1. Pech M, Serafin Z, Fischbach F, Damm R, Jargiello T, Seidensticker M, et al. Transarterial embolization of acute iatrogenic hemorrhages: predictive factors for mortality and outcome. *Br J Radiol.* (2020) 93:20190413. doi: 10.1259/bjr.20190413
2. Yang J, Zhang XH, Huang YH, Chen B, Xu JB, Chen CQ, et al. Diagnosis and treatment of abdominal arterial bleeding after radical gastrectomy: a retrospective analysis of 1875 consecutive resections for gastric cancer. *J Gastrointest Surg.* (2016) 20:510–20. doi: 10.1007/s11605-015-3049-z
3. Kaminskis A, Ivanova P, Kratovska A, Ponomarjova S, Ptasnuka M, Demicevs J, et al. Endoscopic hemostasis followed by preventive transarterial embolization in high-risk patients with bleeding peptic ulcer: 5-year experience. *World J Emerg Surg.* (2019) 14:45. doi: 10.1186/s13017-019-0264-z
4. Laing CJ, Tobias T, Rosenblum DI, Banker WL, Tseng L, Tamarkin SW. Acute gastrointestinal bleeding: emerging role of multidetector CT angiography and review of current imaging techniques. *Radiographics.* (2007) 27:1055–70. doi: 10.1148/rg.274065095
5. Venclauskas L, Bratlie SO, Zachrisson K, Maleckas A, Pundzius J, Jonson C. Is transcatheter arterial embolization a safer alternative than surgery when endoscopic therapy fails in bleeding duodenal ulcer? *Scand J Gastroenterol.* (2010) 45:299–304. doi: 10.3109/00365520903486109
6. Zhou TY, Sun JH, Zhang YL, Zhou GH, Nie CH, Zhu TY, et al. Post-pancreaticoduodenectomy hemorrhage: DSA diagnosis and endovascular treatment. *Oncotarget.* (2017) 8:73684–92. doi: 10.18632/oncotarget.17450
7. Celinski K, Cichoz-Lach H, Madro A, Slomka M, Kasztelan-Szczerbinska B, Dworzanski T. Non-variceal upper gastrointestinal bleeding–guidelines on management. *J Physiol Pharmacol.* (2008) 59(Suppl.2):215–29.
8. Pasha SF, Shergill A, Acosta RD, Chandrasekhara V, Chathadi KV, Early D, et al. The role of endoscopy in the patient with lower GI bleeding. *Gastrointest Endosc.* (2014) 79:875–85. doi: 10.1016/j.gie.2013.10.039
9. Gralnek IM, Dumonceau JM, Kuipers EJ, Lanas A, Sanders DS, Kurien M, et al. Diagnosis and management of nonvariceal upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. *Endoscopy.* (2015) 47:a1–46. doi: 10.1055/s-0034-1393172
10. Wong TC, Wong KT, Chiu PW, Teoh AY Yu SC, Au KW, et al. A comparison of angiographic embolization with surgery after failed endoscopic hemostasis to bleeding peptic ulcers. *Gastrointest Endosc.* (2011) 73:900–8. doi: 10.1016/j.gie.2010.11.024
11. Katano T, Mizoshita T, Senoo K, Sobue S, Takada H, Sakamoto T, et al. The efficacy of transcatheter arterial embolization as the first-choice treatment after failure of endoscopic hemostasis and endoscopic treatment resistance factors. *Dig Endosc.* (2012) 24:364–9. doi: 10.1111/j.1443-1661.2012.01285.x
12. Ai M, Lu G, Xu J. Endovascular embolization of arterial bleeding in patients with severe acute pancreatitis. *Wideochir Inne Tech Maloinwazyjne.* (2019) 14:401–7. doi: 10.5114/wiitm.2019.86919
13. Chen YT, Sun HL, Luo JH, Ni JY, Chen D, Jiang XY, et al. Interventional digital subtraction angiography for small bowel gastrointestinal stromal tumors with bleeding. *World J Gastroenterol.* (2014) 20:17955–61. doi: 10.3748/wjg.v20.i47.17955

14. Sirvinskas A, Smolskas E, Mikelis K, Brimiene V, Brimas G. Transcatheter arterial embolization for upper gastrointestinal tract bleeding. *Wideochir Inne Tech Maloinwazyjne.* (2017) 12:385–93. doi: 10.5114/wiitm.2017.72319
15. Defreyne L, Vanlangenhove P, De Vos M, Pattyn P, Van Maele G, Decruyenaere J, et al. Embolization as a first approach with endoscopically unmanageable acute nonvariceal gastrointestinal hemorrhage. *Radiology.* (2001) 218:739–48. doi: 10.1148/radiology.218.3.r01mr05739
16. Delgal A, Cercueil JP, Koutlidis N, Michel F, Kermarrec I, Mourey E, et al. Outcome of transcatheter arterial embolization for bladder and prostate hemorrhage. *J Urol.* (2010) 183:1947–53. doi: 10.1016/j.juro.2010.01.003
17. Tarasconi A, Baiocchi GL, Pattonieri V, Perrone G, Abongwa HK, Molfino S, et al. Transcatheter arterial embolization versus surgery for refractory non-variceal upper gastrointestinal bleeding: a meta-analysis. *World J Emerg Surg.* (2019) 14:3. doi: 10.1186/s13017-019-0223-8
18. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39:1137–49. doi: 10.1109/TPAMI.2016.2577031
19. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu (2017). p. 2117–25. doi: 10.1109/CVPR.2017.106
20. Pisano ED, Zong S, Hemminger BM, DeLuca M, Johnston RE, Muller K, et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J Digit Imaging.* (1998) 11:193–200. doi: 10.1007/BF03178082
21. Zhou Z, Sanders JW, Johnson JM, Gule-Monroe MK, Chen MM, Briere TM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. *Radiology.* (2020) 295:407–15. doi: 10.1148/radiol.2020191479
22. De Vente C, Vos P, Hosseinzadeh M, Pluim J, Veta M. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Trans Biomed Eng.* (2020) 68:374–83. doi: 10.1109/TBME.2020.2993528
23. Zhou LQ, Wu XL, Huang SY, Wu GG, Ye HR, Wei Q, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology.* (2020) 294:19–28. doi: 10.1148/radiol.2019190372
24. Podgorsak AR, Rava RA, Shiraz Bhurwani MM, Chandra AR, Davies JM, Siddiqui AH, et al. Automatic radiomic feature extraction using deep learning for angiographic parametric imaging of intracranial aneurysms. *J Neurointerv Surg.* (2020) 12:417–21. doi: 10.1136/neurintsurg-2019-015214
25. Gao Y, Song Y, Yin X, Wu W, Zhang L, Chen Y, et al. Deep learning-based digital subtraction angiography image generation. *Int J Comput Assist Radiol Surg.* (2019) 14:1775–84. doi: 10.1007/s11548-019-02040-x
26. Lekovic GP, Kim LJ, Gonzalez LF, Bice A, Albuquerque FC, McDougall CG. Radiation exposure during endovascular procedures. *Neurosurgery*. (2008) 63(1Suppl.1):ONS81–5. doi: 10.1227/01.NEU.0000310770.19631.81

Conflict of Interest: SC, TF, and HS were employed by United Imaging Intelligence.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 05 July 2022
doi: 10.3389/fmed.2022.805230

Return-to-Work Predictions for Chinese Patients With Occupational Upper Extremity Injury: A Prospective Cohort Study

*Zhongfei Bai[1], Jiaqi Zhang[2], Chaozheng Tang[3], Lejun Wang[4], Weili Xia[1], Qi Qi[1], Jiani Lu[1], Yuan Fang[1], Kenneth N. K. Fong[2] and Wenxin Niu[1]**

[1] Shanghai YangZhi Rehabilitation Hospital (Shanghai Sunshine Rehabilitation Centre), School of Medicine, Tongji University, Shanghai, China, [2] Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China, [3] Capacity Building and Continuing Education Center, National Health Commission of the People's Republic of China, Beijing, China, [4] Department of Physical Education, Sport and Health Research Center, Tongji University, Shanghai, China

Objective: We created predictive models using machine learning algorithms for return-to-work (RTW) in patients with traumatic upper extremity injuries.

Methods: Data were obtained immediately before patient discharge and patients were followed up for 1 year. K-nearest neighbor, logistic regression, support vector machine, and decision tree algorithms were used to create our predictive models for RTW.

Results: In total, 163 patients with traumatic upper extremity injury were enrolled, and 107/163 (65.6%) had successfully returned to work at 1-year of follow-up. The decision tree model had a lower F1-score than any of the other models (t values: 7.93–8.67, $p < 0.001$), while the others had comparable F1-scores. Furthermore, the logistic regression and support vector machine models were significantly superior to the k-nearest neighbors and decision tree models in the area under the receiver operating characteristic curve (t values: 6.64–13.71, $p < 0.001$). Compared with the support vector machine, logistical regression selected only two essential factors, namely, the patient's expectation of RTW and carrying strength at the waist, suggesting its superior efficiency in the prediction of RTW.

Conclusion: Our study demonstrated that high predictability for RTW can be achieved through use of machine learning models, which is helpful development of individualized vocational rehabilitation strategies and relevant policymaking.

Keywords: upper extremity injury, return-to-work, vocational rehabilitation, support vector machine, machine learning, occupational health

OPEN ACCESS

Edited by:
Jingjing You,
The University of Sydney, Australia

Reviewed by:
William Keith Gray,
Northumbria Healthcare NHS Foundation Trust, United Kingdom
Chich-Haung Richard Yang,
Tzu Chi University, Taiwan

***Correspondence:**
Wenxin Niu
niu@tongji.edu.cn

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *30 October 2021*
Accepted: *06 June 2022*
Published: *05 July 2022*

Citation:
Bai Z, Zhang J, Tang C, Wang L, Xia W, Qi Q, Lu J, Fang Y, Fong KNK and Niu W (2022) Return-to-Work Predictions for Chinese Patients With Occupational Upper Extremity Injury: A Prospective Cohort Study. Front. Med. 9:805230. doi: 10.3389/fmed.2022.805230

INTRODUCTION

Occupational accidents are the most common causes of arm and hand injuries in China. A previous dataset, collected in Chinese cities with concentrated industrial development, showed that 85.4% of patients acquired their injuries in manufacturing industries; severe injuries commonly resulted from working with food, furniture, non-metallic minerals, and wood products (1).

A return-to-work (RTW) is the goal of rehabilitation for patients with work-related injuries. There have been numerous factors for successful RTW in patients with traumatic upper extremity

(UE) injury in other countries (2, 3), including sociodemographic factors (e.g., age, educational level, and income), severity/location of injury (e.g., type of injury, joint injury, amputation), and function of the involved UE (e.g., strength, finger dexterity, and participation in purposeful tasks). Although these factors have enriched our understanding of what may influence patient employment after injury, there are two major limitations that should be addressed. First, it is impractical for rehabilitation service providers to collect extensive data from every patient to predict RTW in clinical settings. Therefore, it is important to create predictive models with higher prediction performance using a smaller number of factors. Second, RTW is not a purely biomedical process; on the contrary, many relevant cultural factors may be involved. Over the past decades, although some epidemic studies have reported the prevalence of hand injury and its prognostic factors in China (1), few authors have investigated which factors may contribute to patients' successful RTW or long-term absence from work after a standard rehabilitation program. It may also limit stakeholders in formation of appropriate policies, such as which patients should be endorsed for sick leave extension.

Conventional statistical methods, such as parametric tests of group means, logistical regression, the Kaplan-Meier method and Cox regression analysis, were used to explore and find predictors for RTW. However, the performance of RTW prediction based on predictor thresholds has not been examined in most studies; this could bring into question how the factors can correctly predict RTW in a specific time frame. Machine learning makes classifications and predictions based on probabilistic modeling and has been widely employed to solve industrial problems, such as prediction of project safety performance at construction sites (4). Recently, this approach has attracted researchers' attention in the biomedical and healthcare fields (5), in hopes of predicting brain disorders using neuroimaging data (6) or classifying the risk of developing a sudden illness, such as stroke (7). Lee and Kim (8) created machine learning models to predict RTW for vocational rehabilitation patients injured in an industrial accident; a high prediction performance was found, as indicated by high areas under the receiver operating characteristic (ROC) curves. Machine learning is still a novel approach for vocational rehabilitation, and more research is warranted in additional patients after an occupational accident.

We conducted a prospective cohort study in Shanghai, enrolling patients after traumatic UE injury due to occupational accidents, and all patients were followed up for 1 year. Four commonly examined algorithms, namely, k-nearest neighbors (kNN), logistic regression, support vector machine (SVM), and decision tree, were used to select the factors of importance for RTW. The predictability of the four models was then evaluated.

MATERIALS AND METHODS

Study Design and Participants

This was a prospective cohort study from January 2016 to December 2017, which enrolled patients after traumatic UE injury, admitted to Shanghai YangZhi Rehabilitation Hospital for treatment.

Patients were enrolled in the cohort if they met the following criteria: patients with traumatic UE injury, such as bone fracture and tendon injury; work-related injury identified by the Shanghai Municipal Human Resources and Social Security Bureau; age $\geq$18 years; first-ever rehabilitation experience after injury. We excluded patients if they met any of the following criteria: comorbid injuries in any other body region or did not complete the rehabilitation. This study was approved by the Research Committee of the Shanghai YangZhi Rehabilitation Hospital (No. YZ2016-097). Written informed consent was obtained from all patients.

Data Description

Patient demographics, injury information, RTW expectation, physical work demands, functional assessments, and a self-rating scale for the severity of post-traumatic stress disorder (PTSD) were assessed by two occupational therapists before patient discharge. These data, with a total of 27 variables, were further used for machine learning modeling.

Patient demographics included age, sex, marital status, and educational level. For injury information, time since injury in number of days, injured hand dominance (i.e., dominant, non-dominant, or bilateral), and injury location (i.e., finger, wrist, forearm, elbow, upper arm, shoulder, or multiple locations) were collected. The intensity of chronic pain due to injuries was measured using a visual analog scale ranging from zero to ten. Zero indicated no pain at all, while 10 signified pain as bad as possible.

Patients were asked about their expectation of RTW using a 5-point Likert scale ranging from zero to four. One and four represented no expectation and complete expectation, respectively. Likewise, patients' family members were asked to rate the extent to which they expected patients to return to work. If the patients' family members were not reachable, the patients answered this question. We also surveyed employers' attitudes toward RTW because they are crucial. However, employers are not usually reachable, and patients were asked to rate the extent to which their employers expected RTW, based on previous communications.

Physical work demands were classified as sedentary, light, medium, heavy, or very heavy, according to work intensity and frequency. Grip and pinch strength were measured using a Jamar hand dynamometer (9). The EvalTech system (BTE, Hanover, Germany) was used to measure the lifting strength of the bilateral UEs and the carrying strength at the waist and shoulder level. Hand dexterity was quantified by the Purdue Pegboard Test, which involved counting the number of objects inserted during the five subtests (10). The capacity of injured UEs to engage in purposeful and skillful tasks was evaluated using the Chinese version of the Disabilities of the Arm, Shoulder, and Hand (DASH) score (11). The DASH is a self-rated questionnaire that measures the severity of disability and symptomology when performing a given task. The DASH score ranges from 0 to 100, with a higher score indicating a more severe UE disability. The severity of PTSD symptoms was evaluated using the Chinese version of the PTSD Checklist–civilian version (PCL-c), with a higher score indicating more severe symptoms of PTSD (12).

TABLE 1 | Univariant logistic regression result comparison between RTW and non-RTW patients.

Variables	RTW ($n = 107$)	Non-RTW ($n = 56$)	t, Mann-Whitney, χ^2		Univariant logistic regression	
			Statistic	p	OR	p
Age (years)	37.4 ± 9.7	39.3 ± 10.9	1.12	0.265	0.982	0.263
Sex						
Male	79	38	0.65	0.421	0.748	0.422
Female	28	18				
Marital status						
Married	90	47	<0.01	1.000	0.986	0.976
Single	17	9				
Educational level						
Illiteracy	1	3	−2.58	0.010	1.713	0.007
Primary school	10	8				
Junior middle school	47	29				
High middle school	35	14				
College diploma or higher	14	2				
Time since injury (days)	142.1 ± 76.4	172.9 ± 91.4	2.28	0.024	0.996	0.029
Injured hand dominance						
Dominant	53	27	1.60	0.512	0.845	0.553
Non-dominant	51	25				
Bilateral	3	4				
Injury location						
Finger	67	21	11.50	0.057	0.813	0.025
Wrist	18	14				
Forearm	5	7				
Elbow	5	4				
Upper arm	2	2				
Shoulder	8	5				
Multi-location	2	3				
Pain intensity	3.0 ± 2.0	3.2 ± 2.2	0.70	0.486	0.946	0.484
Patient's expectation of RTW	2.6 ± 1.0	2.0 ± 1.1	−3.17	0.002	1.661	0.001
Family's expectation of RTW	2.6 ± 1.0	2.0 ± 1.2	−2.84	0.005	1.647	0.002
Employer's expectation of RTW	2.5 ± 0.9	2.0 ± 0.9	−3.26	0.001	1.909	0.001
Physical work demands						
Sedentary	1	0	−0.35	0.724	0.947	0.741
Light	21	12				
Medium	40	17				
Heavy	27	18				
Very heavy	18	9				
Grip strength of the injured UE (kg)	10.2 ± 8.9	17.8 ± 12.0	−4.56	<0.001	1.072	<0.001
Grip strength of the healthy UE (kg)	36.2 ± 10.5	33.2 ± 10.5	−1.72	0.087	1.027	0.088
Pinch strength of the injured UE (kg)	5.7 ± 3.2	3.7 ± 2.8	−3.79	<0.001	1.236	<0.001
Pinch strength of the healthy UE (kg)	10.1 ± 4.7	9.3 ± 4.4	−1.08	0.161	1.047	0.289
Lifting strength of the injured UE (kg)	27.3 ± 16.8	17.0 ± 12.6	−4.06	<0.001	1.055	<0.001
Lifting strength of the healthy UE (kg)	47.8 ± 18.9	42.1 ± 17.6	−1.88	0.062	1.017	0.065
Carrying strength at waist (kg)	27.0 ± 12.7	16.3 ± 12.0	−5.20	<0.001	1.075	<0.001
Carrying strength at shoulder (kg)	21.8 ± 11.3	12.5 ± 9.2	−5.30	<0.001	1.094	<0.001
Purdue pegboard test						
Injured hand	12.2 ± 4.2	9.5 ± 5.4	−3.36	0.001	1.128	0.001
Healthy hand	16.2 ± 1.8	15.7 ± 2.1	−1.57	0.119	1.150	0.120
Both hands	11.2 ± 4.2	8.3 ± 4.7	−3.98	<0.001	1.169	<0.001
Injured + healthy + both	39.6 ± 8.6	33.5 ± 10.9	−3.67	<0.001	1.069	<0.001
Assembly	28.2 ± 10.3	22.5 ± 12.7	−2.90	0.005	1.045	0.003
DASH	34.5 ± 19.3	43.8 ± 17.3	3.00	0.003	0.974	0.004
PCL-c	35.4 ± 12.7	39.8 ± 13.9	2.03	0.044	0.975	0.047

All variables were compared between patients who returned to work and those who did not. Independent sample t-tests (t) were used for continuous data, while Mann-Whitney tests were used for ordinal data. The differences on categorical data were checked by using Chi-square tests (χ^2). In addition, univariant logistic regression tests were employed to investigate whether variables were individually predictive for RTW. Only those variables which showed significant predictability were included for machine learning modeling. In this table, continuous data are expressed as mean ± SD, ordinal and nominal data are expressed as a number. RTW, return to work; OR, odds ratio; UE, upper extremity; DASH, Disability of the Arm, Shoulder and Hand; PCL-c, Post-traumatic Stress Disorder Checklist–civilian version.

All patients were followed-up for 1 year by a social worker via telephone. A successful RTW case was defined as a patient who returned to work for at least one month in the first year after discharge.

Machine Learning Modeling

In this study, kNN, logistic regression, SVM, and decision tree algorithms were used to train predictive models for the dependent outcome (i.e., RTW at 1-year follow-up), which was defined as binary. Univariant logistic regression tests indicated that 17/27 variables (**Table 1**) were significantly predictive of RTW; these were then selected as input variables for model training. In view of the small sample size ($n = 163$), overfitting could be easily induced, regardless of the algorithm, if a large number of variables were input. Therefore, we further selected the best subsets of variables for kNN, logistic regression, and SVM using an exhaustive feature search. Specifically, the variable number of subsets started from one and all possible subsets with one variable were created. Then, the models were trained with all subsets, and the one with the most optimal performance was selected. Finally, the variable number of subsets was increased, and the optimal subset updated. The search was stopped if the performance of the models did not improve, even as more variables were input. The aforementioned search was not applied for decision tree model training because this algorithm can select the most relevant variables automatically, according to their importance, and discard irrelevant variables.

In the validation method, data were separated into two datasets for model training (70%) and validation (30%). Because of the limited sample size, random separation could produce substantially varied and unreliable model performance. Therefore, each model was trained 100 times to obtain its performance distribution, which was then compared among the models. The F1-score, which is the harmonic mean of precision and recall, was used to evaluate the performance of models on validation datasets. This was done even with the imbalance of outcome classes, due to 65.6% of our included patients successfully RTW. Optimal hyperparameter combinations were selected using a grid search method. The scikit-learn toolkit (version 0.24.0) was used for model training and validation (13).

Statistical Analysis

Statistical analysis was performed using SPSS22 (IBM, NY, and USA) with a level of significance of 0.05. Initially, the baseline differences between RTW and non-RTW patients were compared using independent *t*-tests, Mann-Whitney tests, or chi-square tests when appropriate. Second, univariate logistic regression was used to determine whether individual variables were predictive of RTW. Third, to evaluate performance of the four models, F1-scores and areas under the ROC were compared using one-way repeated measures analysis of variance (ANOVA), and *post-hoc* analyses were conducted using paired *t*-tests with the Bonferroni correction (corrected alpha threshold = 0.05/6). One-way ANOVA was used to examine whether the F1 score from 100 training sessions was significantly different from sets with larger numbers of training sessions.

RESULTS

A total of 179 adult inpatients with traumatic UE injury were enrolled. Ultimately, 163 patients were successfully followed up, of which 107 (65.6%) successfully returned to work by 1-year. Comparisons between RTW and non-RTW patients indicated significant differences in many variables that were also predictive of RTW (**Table 1**). A one-way repeated measures ANOVA indicated a significant difference in the F1-score among the four models ($F = 47.61$, $p < 0.001$), as shown in **Figure 1**. *Post-hoc* analysis by paired *t*-tests found that the decision tree model had a lower F1-score than any of the others (t values ranging from 7.93 to 8.67, all $p < 0.001$, survived Bonferroni correction), and the rest of the comparisons were not significant (t values ranging from 0.92 to 1.73, *p*-values ranging from 0.087 to 0.361). In terms of the factors selected for modeling, time since injury, carrying strength at the waist, carrying strength at the shoulder, Purdue pegboard test score (injured hand), and Purdue pegboard test score (both hands) were five optimal variables for kNN, two variables (patient's expectation of RTW and carrying strength at the waist) for logistic regression, and four [injury located at fingers, patient's expectation to RTW, carrying strength to shoulder, and Purdue pegboard test score (both hands)] for SVM.

The ROC analysis results are shown in **Figure 2**. One-way repeated measures ANOVA indicated significant differences among the four models ($F = 95.48$, $p < 0.001$). *Post-hoc* analysis indicated that the logistic regression and SVM models had comparable areas under the ROC ($t = 0.13$, $p = 0.896$) and were significantly superior to the kNN and decision tree models (t values ranging from 6.64–13.71, all $p < 0.001$, survived Bonferroni correction). In addition, the area under the ROC curve of the kNN model was also significantly larger than that of the decision tree model ($t = 6.70$, $p < 0.001$, surviving Bonferroni correction).

In view of limited computational resources, each model was trained 100 times. To evaluate the effect of the number of training sessions on performance estimation, number of training sessions was manipulated from 5 to 10,000. As shown in **Figure 2B**, the F1-score was relatively precise when larger numbers of training (e.g., 500, 2,000, and 10,000) were applied, regardless of the algorithms. In contrast, small numbers of training sessions (e.g., 5, 10, and 30) yielded substantially variable and much lower F1-socres than larger training sets. One-way ANOVA suggested that F1-scores resulting from 100 training sessions were not significantly different from 500, 2,000 or 10,000 training sessions for the kNN ($F = 0.110$, $p = 0.954$), logistical regression ($F = 1.88$, $p = 0.131$), SVM ($F = 1.95$, $p = 0.119$), and decision tree ($F = 0.285$, $p = 0.836$) models, indicating that 100 times was sufficient for model training (**Figure 2**).

DISCUSSION

We demonstrate that machine learning models can be used for RTW prediction in Chinese patients after traumatic UE injuries, indicating high predictive performance. Although both logistical regression and SVM displayed better performance than the others, logistical regression required a smaller number of

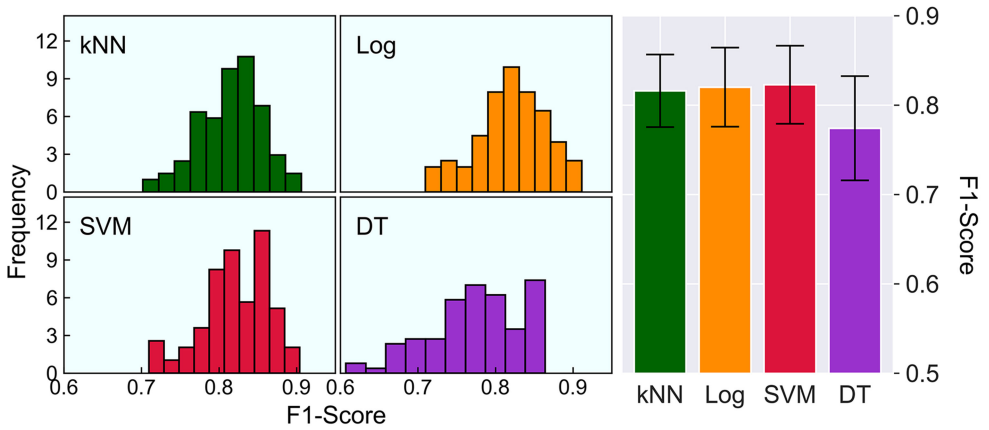

FIGURE 1 | Comparison on F1-scores of the four models. The left histograms show the distribution of the F1-score, and the right bar chart shows a direct comparison on the F1-scores of kNN (0.816 ± 0.041), Log (0.820 ± 0.044), SVM (0.823 ± 0.044) and DT (0.774 ± 0.059). The error bars represent one standard deviation of uncertainty. kNN, k-nearest neighbors; Log, logistic regression; SVM, support vector machine; DT, decision tree.

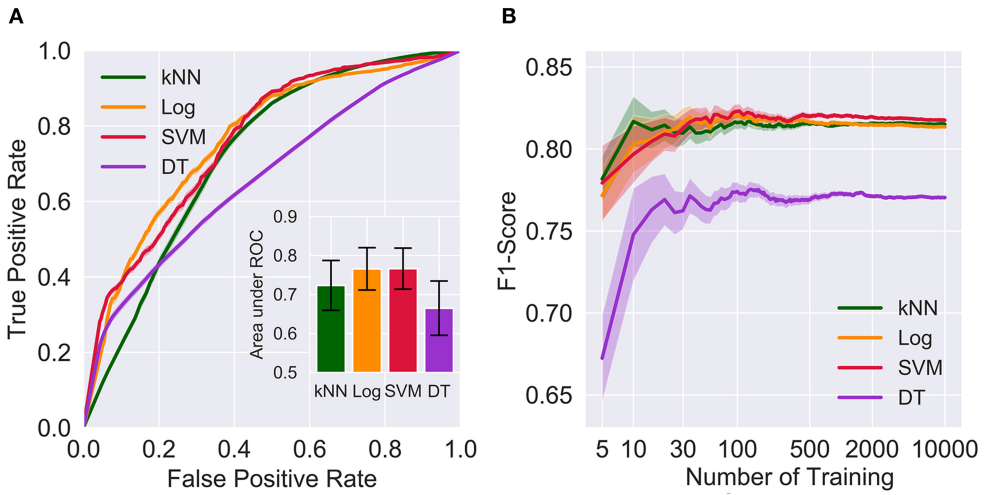

FIGURE 2 | Comparison on the areas under the ROC of the kNN (0.723 ± 0.064), Log (0.766 ± 0.054), SVM (0.766 ± 0.053) and DT (0.665 ± 0.070) and the effects of the number of trainings on F1-scores. The error bars in **(A)** represent one standard deviation of uncertainty. The shaded regions in **(A,B)** represent one standard error of the mean. kNN, k-nearest neighbors; Log, logistic regression; SVM, support vector machine; DT, decision tree; ROC, receiver operating characteristic curve.

factors, suggesting its high efficiency. We also discovered a large number of factors which were in line with previous studies associated with RTW (2, 14). Our machine learning models selected several important factors, such as carrying strength at the waist, patient's expectation of RTW, and Purdue pegboard test score (both hands).

RTW factors following various work-related injuries have been analyzed using traditional statistical methods (2, 3, 15). Our study is the first to use machine learning models to predict RTW in patients after a traumatic UE injury. Logistical regression and SVM were the two best algorithms for predicting RTW. Recently, prediction of risk level classification, differential diagnoses, and prognoses of various diseases have been investigated using machine learning models with excellent performance (6, 7, 16). In particular, SVM has shown superior performance (6, 8), which is in line with our findings. While the black-box problem of SVM is a complex mathematical formulation, it is difficult to interpret the model. Most recently, Rudin (17) argued that when addressing practical problems, designing inherently interpretable models is the way forward, rather than trying to explain black box models. By contrast, logistical regression classifies samples based on probability which is easily interpreted. Although comparable performance was obtained with SVM and logistic regression, logistic regression required only two factors, namely, the patient's expectation of RTW and carrying strength at the waist, suggesting its superior efficiency.

Shi et al. (2) reported that the severity of injury as well as pre-injury income were consistent factors for RTW. Recently, Marom et al. reported additional factors contributing to RTW, such as compensation, educational level, self-efficacy, work demands, pain, and physical capacity (3). In our study, pre-injury income was not included because most patients refused to disclose their financial status. Instead of assessing the severity of the injury, a series of functional assessments were conducted for three reasons. First, the initial severity of hand injury is only partially correlated with functional performance, which is more relevant to the probability of RTW after injury (18). Second, our patients had different single or multiple locations of injury, and it was difficult to evaluate the severity using a uniform score. Third, this study was conducted in a rehabilitation hospital and functional assessments were of practical convenience. Among these functional assessments, carrying strength using both hands was an important factor for RTW. A possible explanation might be that most of our patients were manual workers from manufacturing industries, for whom carrying strength is an essential demand to return to previous work (19). In addition, the patient's expectation of RTW was a critical factor selected by both logical regression and SVM. These findings were in line with those by Heijbel et al. (20) that individuals with expectations of RTW had an approximately eight times higher possibility of RTW than those without that expectation.

The main goal of rehabilitation for occupational injuries is to improve overall functional capacity and ultimately facilitate RTW. Accurate prediction of RTW is helpful for individualized vocational rehabilitation treatment plans. For instance, work-hardening training is crucial for patients who have a high probability of returning to previous work; in contrast, patients who are not likely to return to work, due to severe functional impairments, have to seek supported employment, duty modification, or job transition assistance (21, 22). Most recently, Lee and Kim (8) used similar algorithms to predict whether patients could RTW successfully after an industrial accident. Specific assessments of body function were missing in their study. We focused only on patients with traumatic UE injuries; in particular, a series of functional assessments for UEs were included for modeling, making our findings more specific to the targeted population.

We provide a novel direction for stakeholders when formulating policies relevant to occupational RTW. An RTW policy is designed to help injured workers to return to work in a safe and timely manner, which is beneficial for both employers and the workers themselves. Our machine learning models can obtain a patients' probability of RTW based on this previous dataset. Therefore, stakeholders can assign more individualized policies to workers after an injury. Currently, all occupational injury workers identified by the Shanghai Municipal Human Resources and Social Security Bureau can be approved for one-year sick leave with compensation. However, this policy may not be appropriate without consideration of individual body function. For instance, those with worse body function usually have a lower probability of RTW and should be endorsed for sick leave extensions. However, a shorter period was adequate for those with a higher probability of RTW.

This study has some limitations. First, our sample size was small, which may lead to overfitting, even though some modeling strategies have been employed to compensate for this disadvantage. Second, expectations of RTW were assessed using a 5-point Likert scale, which may not be adequate to represent the full construct of expectation. More standardized assessments with better construct validity are recommended for use in future studies, such as the questionnaire used by Sampere et al. (23). Third, only four commonly used machine learning algorithms were investigated, and higher predictability may have been yielded by others.

CONCLUSION

RTW can be highly predicted by machine learning models, of which both logistic regression and SVM demonstrated high predictability. In particular, logistical regression selected for only two essential factors: a patient's expectation of RTW and carrying strength at the waist. The selected factors can be considered the most relevant factors for prediction of RTW after traumatic UE injury. Predictive models could contribute to the development of tailor-made vocational rehabilitation programs. Furthermore, machine-learning-based predictive models provide a novel direction for stakeholders while formulating policies relevant to occupational RTW.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Committee of Shanghai YangZhi Rehabilitation Hospital. All procedures were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and the Helsinki Declaration. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZB, JZ, KF, and WN contributed to conception and design of the study. WX, JL, and YF collected the data. ZB, LW, and QQ organized the algorithms and database. JZ and QQ performed the statistical analysis. ZB wrote the first draft of the manuscript. JZ, CT, LW, and WX wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (32071308), Shanghai Sailing

Program (20YF1445100), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), and Fundamental Research Funds for the Central Universities.

ACKNOWLEDGMENTS

The authors would like to thank all the participants enrolled in this study.

REFERENCES

1. Jin K, Lombardi DA, Courtney TK, Sorock GS, Li M, Pan R, et al. Patterns of work-related traumatic hand injury among hospitalised workers in the People's Republic of China. *Inj Prev.* (2010) 16:42–9. doi: 10.1136/ip.2008.019737
2. Shi Q, Sinden K, MacDermid JC, Walton D, Grewal R. A systematic review of prognostic factors for return to work following work-related traumatic hand injury. *J Hand Ther.* (2014) 27:55–62. doi: 10.1016/j.jht.2013.10.001
3. Marom BS, Ratzon NZ, Carel RS, Sharabi M. Return-to-work barriers among manual workers after hand injuries: 1-year follow-up cohort study. *Arch Phys Med Rehabil.* (2019) 100:422–32. doi: 10.1016/j.apmr.2018.07.429
4. Poh CQX, Ubeynarayana CU, Goh YM. Safty leading indicators for construction sites: a machine learning appraoch. *Automat Constr.* (2018) 93:375–86. doi: 10.1016/j.autcon.2018.03.022
5. Luo W, Phung D, Tran T, Gupta S, Ranna S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* (2016) 18:e323. doi: 10.2196/jmir.5870
6. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage.* (2017) 145:137–65. doi: 10.1016/j.neuroimage.2016.02.079
7. Li X, Bian D, Yu J, Li M, Zhao D. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC Med Inform Decis Mak.* (2019) 19:261. doi: 10.1186/s12911-019-0998-2
8. Lee J, Kim HR. Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques. *J Korean Med Sci.* (2018) 33:e144. doi: 10.3346/jkms.2018.33.e144
9. Bai Z, Shu T, Niu W. Test-retest reliability and measurement errors of grip strength test in patients with traumatic injuries in the upper extremity: a cross-sectional study. *BMC Musculoskelet Disord.* (2019) 20:256. doi: 10.1186/s12891-019-2623-z
10. Buddenberg LA, Davis C. Test-retest reliability of the purdue pegboard test. *Am J Occup Ther.* (2000) 54:555–8. doi: 10.5014/ajot.54.5.555
11. Chen H, Ji X, Zhang W, Zhang Y, Zhang L, Tang P. Validation of the simplified Chinese (Mainland) version of the disability of the arm, shoulder, and hand questionnaire (DASH-CHNPLAGH). *J Orthop Surg Res.* (2015) 10:1–6. doi: 10.1186/s13018-015-0216-6
12. Wu K, Chan S, Yiu VF. Psychometric properties and confirmatory factor analysis of the posttraumatic stress disorder checklist for Chinese survivors of road traffic accidents. *Hong Kong J Psychiatry.* (2008) 18:144–51.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* (2011) 12:2825–30.
14. Bai Z, Song D, Deng H, Li-Tsang CWP. Predictors for return to work after physical injury in China: a one-year review. *Work.* (2018) 60:319–27. doi: 10.3233/WOR-182735
15. Khorshidi HA, Marembo M, Aickelin U. Predictors of return to work for occupational rehabilitation users in work-related injury insurance claims: insights from mental health. *J Occup Rehabil.* (2019) 29:740–53. doi: 10.1007/s10926-019-09835-4
16. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke.* (2019) 50:1263–5. doi: 10.1161/STROKEAHA.118.024293
17. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x
18. Saxena P, Cutler L, Feldberg L. Assessment of the severity of hand injuries using 'hand injury severity score', and its correlation with the functional outcome. *Injury.* (2004) 35:511–6. doi: 10.1016/S0020-1383(03)00211-0
19. Michener SK, Olson AL, Humphrey BA, Reed JE, Stepp DR, Sutton AM, et al. Relationship among grip strength, functional outcomes, and work performance following hand trauma. *Work.* (2001) 16:209–17.
20. Heijbel B, Josephson M, Jensen I, Stark S, Vingård E. Return to work expectation predicts work in chronic musculoskeletal and behavioral health disorders: prospective study with clinical implications. *J Occup Rehabil.* (2006) 16:169–80. doi: 10.1007/s10926-006-9016-5
21. Krause N, Dasinger LK, Neuhauser F. Modified work and return to work: a review of the literature. *J Occup Reh.* (1998) 8:113–39. doi: 10.1023/A:1023015622987
22. Seing I, MacEachen E, Ekberg K, Ståhl C. Return to work or job transition? Employer dilemmas in taking social responsibility for return to work in local workplace practice. *Disabil Rehabil.* (2015) 37:1760–9. doi: 10.3109/09638288.2014.978509
23. Sampere M, Gimeno D, Serra C, Plana M, López JC, Martínez JM, et al. Return to work expectations of workers on long-term non-work-related sick leave. *J Occup Rehabil.* (2012) 22:15–26. doi: 10.1007/s10926-011-9313-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Medicine

REVIEW
published: 07 July 2022
doi: 10.3389/fmed.2022.864882

Learning Causal Effects From Observational Data in Healthcare: A Review and Summary

*Jingpu Shi and Beau Norgeot**

Anthem, Inc., Point of Care AI, Palo Alto, CA, United States

Causal inference is a broad field that seeks to build and apply models that learn the effect of interventions on outcomes using many data types. While the field has existed for decades, its potential to impact healthcare outcomes has increased dramatically recently due to both advancements in machine learning and the unprecedented amounts of observational data resulting from electronic capture of patient claims data by medical insurance companies and widespread adoption of electronic health records (EHR) worldwide. However, there are many different schools of learning causality coming from different fields of statistics, some of them strongly conflicting. While the recent advances in machine learning greatly enhanced causal inference from a modeling perspective, it further exacerbated the fractured state in this field. This fractured state has limited research at the intersection of causal inference, modern machine learning, and EHRs that could potentially transform healthcare. In this paper we unify the classical causal inference approaches with new machine learning developments into a straightforward framework based on whether the researcher is most interested in finding the best intervention for an individual, a group of similar people, or an entire population. Through this lens, we then provide a timely review of the applications of causal inference in healthcare from the literature. As expected, we found that applications of causal inference in medicine were mostly limited to just a few technique types and lag behind other domains. In light of this gap, we offer a helpful schematic to guide data scientists and healthcare stakeholders in selecting appropriate causal methods and reviewing the findings generated by them.

Keywords: electronic health record, causal inference, machine learning, healthcare, treatment effects, review, potential outcome framework, patient population

OPEN ACCESS

Edited by:
Enrico Capobianco,
University of Miami, United States

Reviewed by:
Juan M. Banda,
Georgia State University, United States
Tomiko Oskotsky,
University of California, San Francisco, United States

***Correspondence:**
Beau Norgeot
beau.norgeot@anthem.com

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *28 January 2022*
Accepted: *17 June 2022*
Published: *07 July 2022*

Citation:
Shi J and Norgeot B (2022) Learning Causal Effects From Observational Data in Healthcare: A Review and Summary. Front. Med. 9:864882. doi: 10.3389/fmed.2022.864882

INTRODUCTION

In healthcare, it is important to distinguish between association and causation when we study treatment effects on patient outcomes. Association between two variables is non-directional and implies that the two variables are correlated. In contrast, causation is directional and indicates that one variable causes the other. In clinical studies, we are more interested in causal analysis to reveal whether a treatment causes a desired outcome.

Using observational data to infer causal treatment effects has become popular in the past decade due to two pivotal advances: the increasingly available patient data captured in Electronic Health Records (EHRs) and machine learning techniques that can efficiently and intelligently analyze large-scale data. On the data side, health care providers worldwide have widely adopted

EHRs (1, 2), which capture patients' clinical and demographic information during interactions with health systems. In addition to EHRs, patient claims data are increasingly available to improve models in the healthcare domain (3). On the algorithm side, machine learning models such as artificial neural networks are powering online search engines, shopping websites, and recommender systems (4). These machine learning models are increasingly used to improve causal inference algorithms.

In the past, many different schools of learning causality coming from different fields of statistics resulted a fractured state of causal inference, creating confusion about which algorithm to use in a study. Recently, the intersection of causal inference, machine learning, and patient data has formed a new front in clinical research. Accordingly, many traditional causal inference models have been improved and many new models have been proposed. While this has enhanced the number of model options to select from in causal inference studies, it has also led to even greater confusion about which type of algorithm is appropriate for a given application. Lack of systematic knowledge of which approaches are promising in theory vs. the approaches that have been validated through real world applications further complicates the debate.

There are different stakeholders in healthcare, including healthcare providers, administrators, clinical researchers, data scientists, and many others. While data scientists, computer engineers, and biomedical statisticians may be less prone to such confusion, the fractured state in this field makes it difficult for other participants to understand the many different types of models and intuitively interpret the model results. We believe it is imperative to address this confusion for all healthcare participants to unlock the massive potential to improve patient outcomes that could be obtained by studying the causal effects of interventions from large-scale, representative, observational patient data that is now available.

In this review, we start by explaining the broad and heterogenous fields of causal inference. We then distill all of these techniques down into a simple unified framework of three algorithm families, based on size of the target patient population that the causal effect estimation will be applied to. This simple unified frame based on the size of the target patient population is important: while statisticians in medical informatics may not necessarily group the algorithms this way, it is beneficial for frontline healthcare professionals such as doctors and nurses to understand the drug effect in the context of its target population, and the effect's variance and bias characteristics when the drug is applied to the treated patient. From the perspective of this unified framework, we then review all existing applications of causal inference in healthcare in the literature, and identify key components of causal inference that are, as of now, lacking in the healthcare domain. Finally, we use these insights to create an intuitive schematic to guide researchers and stakeholders through the process of selecting an appropriate causal inference technique based on their study objectives.

This review is an extension of several works in previous literature on observational causal inference. For example, the authors in Yao et al. (5), Guo et al. (6), and Ding and Li (7) reviewed causal inference in general but without a focus on clinical settings. The authors in Landsittel et al. (8) offered a narrative review of basic concepts of causal inference but did not consider new developments in this field. Prior reviews (9–11) have narrowly focused on the matching method of causal inference, while in this paper we expand to include a much broader algorithm types.

We conclude this section by providing below a summary of all the approaches we review, with respect to their variance-bias trade-off, advantages, disadvantages, and how widely they are applied in clinical studies.

CAUSAL INFERENCE ASSUMPTIONS, FRAMEWORKS, AND TARGET-POPULATION INTERVENTION SIZES

Confounding Variables

Causal inference differs from associative studies due to the modeling of confounding variables (covariates), defined as variables that affect both the treatment and the outcome. In associative studies which focus on patient outcome estimates, confounding variables are modeled in an inclusive manner because the inclusion of these variables in the model improves estimate accuracy. In contrast, causal inference which reveals the causal relationship between treatments and patient outcomes models the confounding variables in an exclusive manner in that their effects are removed through various approaches we review in this paper.

Assumptions

In the literature, several assumptions are widely adopted in causal inference (12). The unconfoundedness assumption, also known as ignorability, states that all confounding variables are observed in the data. In practice, domain experts often examine as many patient variables as possible, including their demographic and clinical characteristics, so that this assumption can be met. The common support or positivity assumption states that any patient has a non-zero probability of being present in any of the treatment groups. The validity of this assumption can be checked by calculating the patients' propensity scores (12). The Stable Unit Treatment Value assumption (SUTVA) states that a patient's outcome only depends on the treatment this patient receives, and not affected by the outcome or treatment of any other patients. The consistency assumption links the potential outcomes to the observed data and implies that the potential outcome under an observed exposure is precisely the outcome that is observed (13).

Bias-Variance Tradeoffs Based on Target-Population Intervention Sizes

Researchers, clinicians, and other healthcare stakeholders may wish to know the treatment effects at different population levels for different purposes. For example, they may want to evaluate the overall effectiveness of the treatment on the whole population. They may want to understand treatment effect differences in different subpopulations to identify the subpopulation where the treatment is the most effective or least

effective. When they treat an individual patient, they may want to know the individual-level treatment effects considering the patient's unique medical benefits and risks.

Driven by such needs, researchers conduct causal inference at different target-population intervention sizes: at one end of the spectrum is the Average Treatment Effect (ATE) that captures the treatment effect for a population at large; at the other end is the Individual Treatment Effect (ITE) that captures the treatment effect heterogeneity across individuals; in between is the conditional average treatment effect (CATE) that captures the treatment effect for subpopulations.

In clinical practices, at the receiving end of any treatment are individual patients. Correspondingly, different treatment effects (ATE, CATE, and ITE) are eventually applied to individual patients. Therefore, it is important to understand the variance-bias tradeoff of the estimate at different target-population intervention sizes: if we use ATE as the treatment effect for an individual patient, the bias will be high due to effect heterogeneity across patients in the population, but the variance will be low due to more data being used in the inference; in contrast, if we use ITE for a patient, the bias will be low, but the variance will be high.

As the rest of the paper shows, ATE provides the best option and fosters estimate efficiency for the whole population, but may not provide the most accurate estimate for any individual patient. ITE maximally leverages the data, but risks being uninterpretable to clinical practitioners. CATE represents a balance between bias and variance and tracks the clinical definition of patient subgroups.

Two Frameworks

There are two widely accepted frameworks in the literature for causal inference: the structural causal model (SCM) (14–16) and the potential outcome framework (POF) (12, 17, 18). SCM consists of two components, the causal graph and the structural equations. A causal graph is a directed acyclic graph (DAG) where the edges represent causal relationships, and the nodes represent variables including treatments, outcomes, and covariates that may or may not be observed. Causal effects can be quantitatively specified through a set of structural equations.

The DAG and structural equations together provide a comprehensive theory of causality and seamlessly tie essential concepts and methodologies in causal inference (14, 19, 20). In addition, it can possibly deal with cases where confounders cannot be measured. For example, in Barter (21), the author used the blood type as an instrument variable—defined as a variable that affects the outcome only through the treatment variable—to estimate the average survival benefit from receiving a liver transplant.

The other framework, called the potential outcome framework, centers on the concept of potential outcomes. In the simplest term, potential outcomes are all the possible outcomes for a patient under all possible treatments, with each outcome corresponding to a treatment. Note that only one potential outcome can be observed for a given patient at a given time. We call the potential outcome that would have been observed had the treatment been different the counterfactual or the missing outcome. In the simplest case, there is only one treatment to consider. A patient can be either given the treatment, i.e., assigned to the treated group, or given no treatment, i.e., assigned to the control group. Under the potential outcome framework, the treatment effect is the difference between the potential outcome if the patient is treated and that if the patient is not treated.

CSM and POF are not competing frameworks but can be unified (22). Despite this fact, the two frameworks have differences in what causal questions they are best suited to handle. Given its strong theoretical grounding, CSM is ideally suited to identifying unknown causal and confounding variables, as well as facilitating explanation. While it is useful to identify all the variables in the causal graph and their causal connections, the primary objective in healthcare is often to estimate the actual effect of a given treatment. POF is best suited for generating these estimates, because comparing potential outcomes eases the removal of confounding effects and enables a natural connection to traditional statistical analyses. For this reason, POF is more widely adopted for healthcare research and will be the focus of this review.

CAUSAL INFERENCE METHODS BY TARGET-POPULATION INTERVENTION SIZES

In this section we review causal inference approaches in the literature under the potential outcome framework and the assumptions stated in Section Causal Inference Assumptions, Frameworks, and Target-Population Intervention Sizes. We organize our review by the approaches' target-population intervention size: from ATE for the whole population to CATE for subpopulations and ITE for individual patients.

We first explain some key notations. Suppose we are interested in the causal effect of a treatment A on outcome Y. The potential outcome denoted by Y^a is the outcome that we would observe under a possible treatment $A = a$. In a binary treatment case, a can possibly take on two values $a \in \{0, 1\}$, where 0 indicates the patient is not treated and 1 indicates the patient is treated. We denote the confounding variables by X. For simplicity, we only focus on the binary treatment case in this paper.

Estimate ATE for the Whole Population

In the binary treatment case, the ATE estimate for the population can be calculated as

$$\tau = E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right) \tag{1}$$

It is the difference between the expected potential outcomes of the population if everyone is treated ($A = 1$) and if no one is treated ($A = 0$).

Note that ATE cannot be directly calculated from equation (1) because only one of the potential outcomes, either Y_i^1 or Y_i^0, can be directly observed for patient i, nor can it be directly calculated from the expected outcomes of the treated and control groups,

$$E\left(Y^1 - Y^0\right) \neq E\left(Y|A = 1\right) - E\left(Y|A = 0\right) \tag{2}$$

due to the existence of confounding variables X. In general, the distribution of confounding variables is different in the treated and control group. If their expected outcomes are directly compared to calculate treatment effects without adjusting for confounding variables, the calculated treatment effects would be biased.

Propensity Score-Based Approaches

Propensity score of a patient is the conditional probability that this patient with $X = x$ is assigned to the treated group. It is expressed as

$$\pi(x) = P_r(A = 1 | X = x),$$

and can be estimated using models such as logistic regression (12). We can use the propensity score in three different ways to balance the covariate distribution between the treated and control group and thus make the two groups comparable.

The first way is to create new control and treated groups using propensity score matching (12, 23). The most straightforward approach is greedy one-to-one matching: one patient from the control group is matched to one patient from the treated group based on their propensity scores. Data of unmatched patients gets thrown away. The covariate distribution of the matched control and treated group is balanced. Then we can calculate the difference of the expected outcomes of the two new groups as the average treatment effect (ATE). In contrast to equation (2), the equation below is now correct due to balanced covariate distributions,

$$E\left(Y^1 - Y^0\right)_{balanced} = E(Y|A = 1)_{balanced} - E(Y|A = 0)_{balanced}$$

In addition to one-to-one matching, propensity score is used in other similar algorithms to create matched groups. These algorithms differ from each other in whether patients are chosen with or without replacement (24), whether matching is optimal, greedy (24), one-to-one, or one-to-many (25), and what metric is used to measure similarity between two patients (11, 23, 26, 27).

The second way of using propensity scores, known as Inverse Probability of Treatment Weighting (IPTW) (28), is to assign different patients with different weights in the calculation of ATE. For patient i, the weight is calculated as

$$w_i = \frac{A_i}{P(A_i = 1|X_i)} + \frac{1 - A_i}{1 - P(A_i = 1|X_i)}.$$

From this equation, we can see that if patient i is in the treated group ($A_i = 1$), the weight assigned to this patient is $w_i = \frac{1}{P(A_i=1|X_i)} = \frac{1}{\pi(x_i)}$. If the patient i is in the control group ($A_i = 0$), the weight then becomes $w_i = \frac{1}{1-P(A_i=1|X_i)} = \frac{1}{1-\pi(x_i)}$. The weight of a patient in a group is just the inverse probability of this patient being assigned to this group. The ATE of the population can then be calculated as

$$\hat{\tau} = \frac{1}{n_1}\sum\nolimits_i w_i y_i^1 - \frac{1}{n_0}\sum\nolimits_i w_i y_i^0$$

where y_i^1 (y_i^0) is the observed outcome for patient i if this patient is treated (untreated), n_1 and n_0 are the number of patients in the treated and control group, respectively. Intuitively, the IPTW approach balances covariate distributions between the two groups by giving the patients underrepresented (overrepresented) in a group higher weight (lower weight).

The third way of using propensity score in ATE estimate is to stratify the population into subpopulations based on the propensity scores of the patients (29). The treatment effect from each subpopulation is then calculated and combined to estimate the ATE of the whole population.

Propensity score-based approaches are intuitive, easy to understand, and capable of producing an unbiased ATE estimates if the propensity score is correctly estimated. If the propensity models are misspecified (for example, the function form in the logistic regression is wrong), the propensity score estimates and subsequent ATE estimates would be biased.

Outcome Regression-Based Approaches

One fundamental challenge in causal inference is the missing data problem: only one of the potential outcomes is observable for a given treatment and patient. Regression models can be used to estimate the missing outcomes, thus solve the missing data problem (17, 30).

Here we outline how outcome regression models are used in ATE estimates but leave the detailed review of these models to Section Estimate ITE for Individual Patients. Suppose the outcome regression function for the control and treated group is $m_0(X)$ and $m_1(X)$, respectively. Once the two functions are fitted, the missing potential outcomes can be predicted as $\hat{Y^0} = m_0(X)$ and $\hat{Y^1} = m_1(X)$. The average treatment effect for the population can be estimated as,

$$\hat{\tau} = E\left(Y^1 - Y^0\right) = \frac{1}{n_0 + n_1}\sum_{k=0}^{n_0+n_1-1}\left(\hat{Y_k^1} - \hat{Y_k^0}\right) \quad (3)$$

which first calculates the difference between the two predicted outcomes of each patient, then averages these differences over all the patients in both groups. Note that $m_0(X)$ and $m_1(X)$ can either take on the same function form, in which case the treatment assignment variable A must be explicitly included in the model as one of the independent variables, or take on different function forms, in which case A is excluded in the model.

Outcome regression models do not require an estimate of propensity scores. However, misspecification of the regression model (for example, the regression function form is wrong) can lead to biased treatment effect estimates.

Doubly Robust Estimator

Both the outcome regression and the propensity model can be misspecified. A combination of the two models, known as a Doubly Robust Estimator (DRE), is proposed in Robins et al. (31) and Funk et al. (32). It calculates the expected outcome for the treated and control group as

$$E\left(Y^1\right) = \frac{1}{n_0 + n_1}\sum_{i=0}^{n_0+n_1-1}\left\{\frac{A_i Y_i}{\pi_i(X_i)} - \frac{A_i - \pi_i(X_i)}{\pi_i(X_i)} m_1(X_i)\right\} \quad (4)$$

and

$$E\left(Y^0\right) = \frac{1}{n_0+n_1}\sum_{i=0}^{n_0+n_1-1}\left\{\frac{(1-A_i)Y_i}{1-\pi_i(X_i)} - \frac{A_i-\pi_i(X_i)}{1-\pi_i(X_i)}\, m_0(X_i)\right\} \quad (5)$$

respectively. Then the ATE can be estimated as $E\left(Y^1\right) - E\left(Y^0\right)$. Essentially, this DRE is an IPTW estimator augmented by term $\frac{A_i-\pi_i(X_i)}{\pi_i(X_i)}\, m_1(X_i)$ in Equation (4) and term $\frac{A_i-\pi_i(X_i)}{1-\pi_i(X_i)}\, m_0(X_i)$ in equation (5). For this reason, it is also called an augmented IPTW estimator.

Another type of DRE is the Targeted Maximum Likelihood Estimator (TMLE), initially proposed in Laan and Rubin (33) and further studied in Schuler and Rose (34). In this approach, an outcome regression model is first used to estimate $E(Y|A,X)$, which is then updated using estimated propensity score $\pi(X)$ in the so called "targeting" step, yielding a better estimate $E^*(Y|A,X)$. Average treatment effect can be calculated as $E^*\left(Y^1\right) - E^*\left(Y^0\right)$.

As implied in the name, DREs have a nice doubly robust property that ensures the ATE estimate is unbiased if only the outcome regression model or only the propensity model is correct. These models also tend to be more efficient than just the IPTW estimators.

Estimate CATE for Subpopulations

In some cases, researchers may be interested in treatment effects for subpopulations, which can be calculated through CATE estimates. These subpopulations can be learned directly from the data or defined by several criteria, ranging from demographic strata or existing clinical heuristics with the goal of creating groups for which the treatment effect and goals are expected to be similar.

Direct and Indirect Stratification

CATE can be calculated *via* population stratification. The idea is to first stratify the population on $f(X)$, i.e., a function of patient covariates X, into subpopulations. Then CATE for each subpopulation is calculated as the difference between the two expected potential outcomes within that subpopulation. As in Morgan and Winship (35), it is mathematically expressed as

$$\tau_{\mathrm{CATE}} = \mathrm{E}\left(\mathrm{Y}\middle|\mathrm{A}=1,\ \mathrm{f}(\mathrm{X})\right) - \mathrm{E}\left(\mathrm{Y}\middle|\mathrm{A}=0,\mathrm{f}(\mathrm{X})\right)$$

Function $f(X)$ can take on different forms. In the basic form $f(X) = X$, the population is stratified directly on covariate X as described in Imbens and Rubin (36), which we call direct stratification. With this approach, the covariates within each stratum (subpopulation) are similar in values across different patients. Suited for scenarios where subpopulations are predefined, this approach provides simple and transparent interpretation of the subpopulation but may lead to data sparsity in some stratum or violation of the positivity assumption. Function $f(X)$ can take on a more complex function form, which we call indirect stratification. If $f(X) = \pi(X)$, the population is stratified on propensity scores (12, 29). This approach alleviates the data sparsity problem, but the interpretation of subpopulations is less intuitive.

Data Driven Determination of Subpopulations

A subpopulation can be viewed as a subspace in the multi-dimensional covariate space. A data driven approach to calculate CATE partitions the covariate space into subspaces in a way that the treatment effect heterogeneity across subspaces is maximized. The resulting subspaces (or subpopulations) reflect the heterogeneity of the underlying data. Some subspaces may be wider or narrower in certain dimensions than others depending on how quickly the treatment effect changes along these dimensions, which is a desired property.

Machine learning models, due to their flexibility, are well-suited for this approach. One of such estimators is proposed in Athey and Imbens (37) based on the classification and regression tree (CART) (38). While a CART model minimizes a predefined loss function in associative studies, it maximizes heterogeneous treatment effect across leaves when used in causal inference. Different sets of samples are used for constructing the tree and for estimating the treatment effect for each subpopulation. Because of this, the approach is called an honest estimation.

In contrast to the approach in Athey and Imbens (37) where only one decision tree is used, the approach proposed by Breiman (39) estimates treatment effects based on the random forest model consisting of multiple decision trees (40).

These machine learning-based models are non-parametric and thus robust to model misspecification. They can capture the heterogeneity structure in the underlying data and reduce the variance of effect estimates in a subpopulation. However, the complexity of such models makes the results less explainable compared to simpler ones, creating obstacles for the medical community to widely adopt these models in clinical applications.

Estimate ITE for Individual Patients

Treatment effects can be different not only across subpopulations, but across different patients as well. Due to the existence of such heterogeneity at individual patient level, ITE estimates are important for personalized medicine and have been increasingly gaining attention in healthcare (41). In the strictest sense, the ITE estimate is conditioning on an individual's characteristics so can be regarded as CATE. However, in this work, we review ITE as a distinct algorithm category separated from CATE. This decision emphasizes the fact that ITE targets individual patients, while CATE targets subgroups of patients.

Intuitively, ITE can be calculated as the difference between the two potential outcomes for a patient. One of the potential outcomes is missing but can be estimated with an outcome regression model, where the potential outcome is the dependent variable and the covariates are the independent variables. In essence, such an outcome regression model fits a function to estimate the regression surface (or outcome surface) in the covariate space using observed patient outcome samples. Note that the function used in outcome regression can be linear, non-linear, or even non-parametric, depending on the underlying data structure. There are two approaches to fit the model, based

on whether the samples from the treated and control group are pooled together in the training step.

One Regression Function

To estimate ITE, we can fit one regression function using pooled samples from both the treated and the control group and regard the treatment assignment A as one of the independent variables, as shown in the equation below,

$$E(Y|X,A) = m(X,A) \tag{6}$$

where $m(X,A)$ estimates the potential outcome conditioned on X and A. Then the ITE estimate for patient i is calculated as $m(X_i,1) - \mathrm{m}(X_i,0)$. One example of such a model is the Bayesian Additive Regression Trees (BART) introduced in Hill (42), Chipman et al. (43), and Chipman et al. (44), where the authors constructed a set of trees using ensemble learning, and imposed a prior regularization to constrain each tree to be a weak learner. Another example is proposed in Foster et al. (45), where the authors used a random forest to fit $m(X,A)$ to estimate ITE. The approach proposed in Nie and Wager (46) fits a single outcome surface first to isolate the impact of the treatment on the outcome, then fits a regression model where the ITE is the only independent variable.

The models fitting one outcome surface are well-suited for scenarios where the treatment effect is small. The analysis in Wendling et al. (47) validates the performance of the BART model using synthetic data based on two major healthcare databases in the United States and concludes that the smaller the ITE is (i.e., the closer the outcome surfaces are between the two treatment groups), the better such models perform. These models perform poorly if there are complex interactions between the treatment assignment and covariates, which makes the outcome surface $f(\cdot)$ very different for the treated and control groups. Such model drawbacks are studied in detail in Alaa and Schaar (48) and Hahn et al. (49).

Two Regression Functions

Instead of fitting one regression function, one can fit two separate functions for the treated and control groups to calculate ITE. In this case, the treatment variable does not need to be included as one of the independent variables in the model because the outcome difference between the two groups is captured with different model parameters. The two regression functions can be expressed as

$$E\left(Y^1|X\right) = m_1(X) \tag{7}$$

and

$$E\left(Y^0|X\right) = m_0(X) \tag{8}$$

for the treated ($A = 1$) and control ($A = 0$) group, respectively. The ITE estimate for patient i is then calculated as $m_1(X_i) - m_0(X_i)$. Different base learners can be used for $m_0(X)$ and $m_1(X)$, as proposed in Athey and Imbens (37), Lu et al. (50), Powers et al. (51), and Künzel et al. (52).

The approach fitting two outcome surfaces separately is suited for the scenarios where the outcome surface is very different for different treatment groups. The downside of this approach is that some common patterns between the two groups get lost during model fitting. A multitask-learning estimator introduced in Alaa and Schaar (48) and Alaa and Schaar (53) fits two outcome surfaces separately but attempts to recover common underlying patterns between the treated and control group through a joint optimization for the two groups.

Estimate Error Bound

Several theories proposed in the literature study the error of the ITE estimate. The authors in Shalit et al. (54) derived a theoretical upper bound for the error, which is a sum of the standard generalization-error in the representation space and the error resulted from the distance between the two treatment group covariate distributions induced by the representation. An extension of this work (named context-aware importance sampling re-weighing) is proposed in Hassanpour and Greiner (55) to theoretically address the selection bias in observational datasets, leading to a solution that weights the samples in such a way that the covariate distribution imbalance between the treated and control group is reduced. Related to the theoretical works above, practical solutions based on deep learning were proposed to incorporate in the loss function the dissimilarity of the learned representations for the treated and control groups so that the error induced by such dissimilarity can be reduced (56–58).

CLINICAL APPLICATIONS OF CAUSAL INFERENCE

Although there are a large number of causal inference techniques in the literature as we reviewed above, these techniques are not applied equally to solve real-world clinical problems. In this section, we review the patterns of how the various causal inference approaches are used in published clinical studies.

Reporting Methods

In searching for published application papers of causal inference models, we follow the applicable guidelines in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (59). The modified PRISMA flow charts for each category of causal inference models are in the **Supplementary Material**. Note that although we follow the PRISMA guidelines whenever deemed applicable to make our search systematic, the review in this section is not a systematic review in the strictest sense, as our goal is not to answer a well-defined and narrowly focused clinical question, but to gain general understanding of the application landscape of causal inference.

Results

Below we list the most relevant published clinical applications for each of the causal models we have identified. If the application list is too long (more than 15 publications), we just list below the top 15 most cited ones according to Google Scholar due to space limitations. The total number of applications

identified with the inclusion and exclusion criteria is given in the **Supplementary Material**.

Applications of ATE Estimators for the Whole Population

Propensity score-based models have been applied to study the effect of interruption of sedation on the death of the patient in Requena et al. (60), the effect of corticosteroids on mortality for patients with influenza A (H1N1pdm09) in Delaney et al. (61), the cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with certain drugs in Graham et al. (62), the association of animal and plant protein intake with all-cause and cause-specific mortality in Song et al. (63), the effect of nasal cannula therapy failure on mortality in Kang et al. (64), the prevalence of sarcopenia in COPD and its impact on health in Jones et al. (65), the safety and efficacy of digoxin in Ziff et al. (66), clinical outcomes after transapical or transfemoral transcatheter aortic valve replacement in Blackstone et al. (67) and many other health related issues in Chang et al. (68), Bangalore et al. (69), Kost and Lindberg (70), Grool et al. (71), Snowden et al. (72), Han et al. (73), and Prati et al. (74).

Applications of outcome regression-based models in clinical studies have been rare. In fact, we did not find any applications of this approach that meet our search criteria.

Doubly robust estimators have been widely applied in real-world clinical studies to determine the effect of sepsis on late mortality in Prescott et al. (75), the effect of proton pump inhibitors use on risk of death in Xie et al. (76), cardiovascular risks of testosterone replacement therapy in men with androgen deficiency in Cheetham et al. (77), the effectiveness of influenza vaccines among elderly people in Izurieta et al. (78), whether antifungal de-escalation leads to adverse outcome in Bailly et al. (79), the association of the use of transthoracic echocardiography with 28-day mortality in Feng et al. (80), the effect of risk assessment on clinical outcomes in Chaffee et al. (81), comparison of children currently and previously diagnosed with autism in Blumberg et al. (82), whether there is a causal link between the Magnet status of a hospital and the central-line-associated bloodstream infections in Barnes et al. (83), as well as a range of health-related issues from association of aspirin with hepatocellular carcinoma and liver-related mortality to effect of angiotensin on hemoglobin levels in Breslau et al. (84), Simon et al. (85), Ajmal et al. (86), Millett et al. (87), Reed et al. (88), and Kawasaki et al. (89).

Application of CATE Estimators

CATE estimators using stratification have been widely applied in clinical studies, for example, to analyze the adverse outcomes of underuse of β-Blockers in elderly patients in Soumerai et al. (90), the rate of mortality in patients receiving drug-eluting stents and undergoing coronary-artery bypass grafting in Hannan et al. (91), the effect of Hydroxychloroquine and tocilizumab therapy on mortality in COVID-19 patients in Ip et al. (92), medical therapy on long-term outcome in patients with myocardial infarction (93), the impact of female sex on clinical outcomes for Atrial Fibrillation in Kuck et al. (94), and a range of other clinical issues (95–104).

There are very few applications of the data driven approach in clinical studies. The recursive partitioning approach (37) is used to study the effect of fluoxetine in patients with a recent stroke in Graham et al. (105), the effect modification in a study of surgical mortality in Lee et al. (106).

Application of ITE Estimators

The applications of ITE estimators are very rare in the literature. The BART model is used to predict the papillary thyroid carcinoma in Guo et al. (107) and to study the consequences of contact with the criminal justice system for health in Esposito et al. (108).

Methods

Search Strategy

Here we describe the search strategy we use to find the published clinical applications of a causal approach. First, we identify the paper in which the model is proposed. If multiple models hence multiple papers exist—there might be model variations, extensions, or improvements—we pick a paper that generated the most citations in Google scholar. We then search in Google Scholar for all the publications citing the identified paper, which we call the anchoring paper, and apply the inclusion and exclusion criteria described below to determine what papers should be included in the application list of the causal approach.

Note that this search strategy is not exhaustive and is not intended to be a scoping review. Using the anchoring paper, we can only identify a subset of the application papers in a causal inference category. Our goal is not to precisely count the number of all applications, but to understand the extent to which different causal models are applied clinically. Accordingly, our strategy is to sample a limited number of publications, but in a systematic way, so that our search is manageable but still reflective of the application landscape in this field.

Inclusion and Exclusion Criteria

For each category of the causal inference approach, we search for publications that cite the anchoring paper in Google Scholar. In the returned result, we exclude any records not in the healthcare domain, which are those that do not contain any of these keywords: medicine, hospital, patient, clinics, healthcare, physician, and disease. We then screen the titles and abstracts of the remaining papers and exclude those not pertaining to applications. Most of the papers eliminated in this step are about models and algorithms related to the causal inference model described in the anchoring paper. The papers remaining after this step are clinical applications that cite the anchoring paper. However, the anchoring paper can be cited in many ways: it can be mentioned in the related work section; it can be cited in the discussion section; or it can be used to derive findings and insights. We proceed to read the papers that are cited more than 10 times, focusing on the section where the anchoring paper is cited. We include the paper in the final application list if the model in the anchoring paper is used as the method (or one of the methods) to draw conclusions, derive findings, or gain insights.

TABLE 1 | Summary of causal inference approaches in healthcare.

Target-Population intervention sizes	Estimator types	Models and algorithms	Advantages	Disadvantages	Variance	Bias	Clinical application patterns and references
Whole population	ATE	Propensity scores-based, propensity score matching and IPTW	Simple, transparent, mimic clinical trials	Model can be misspecified	Low	High	Widely used (60, 68)
		Outcome regression, variations of G-computation	No need to estimate propensity score	Model can be misspecified			Few applications
		Doubly robust estimator, targeted maximum likelihood estimator	Efficient, doubly robust property	Yield biased estimate if both models are misspecified			Widely used (75, 84)
Sub population	CATE	Direct stratification	Easy to interpret	Data sparsity proble	Medium	Medium	Widely used (90, 95)
		Indirect stratification, propensity score-based approach	Robust, easy to satisfy positivity assumption	Subpopulation hard to interpret			
		Data driven, tree based algorithms	Low variance within subpopulation	Subpopulation hard to interpret			Few applications (105, 106)
Individuals	ITE	Fit one outcome surface, BART model etc	Capture common underlying data structure	Not flexible, especially when the outcome surfaces are very different in distinct groups	High	Low	Few applications (107, 108)
		Fit two outcome surfaces	Flexible, allow for different data structure in groups	Does not capture common data pattern in two groups			

Observations

A pattern emerged from surveying and analyzing the applications of causal models in healthcare: although state-of-the-art machine learning-based approaches have been consistently used to improve causal inference techniques algorithmically and generated excitement in the medical research community, these approaches have not been widely adopted in clinical studies. In contrast, simpler approaches based on propensity scores have been widely applied to solve real-world clinical problems. This conclusion is evident from the citation numbers in the **Supplementary Material**: while the number of machine learning applications, such as those based on models in Rubin (30) and Athey and Imbens (37), is in single digit at most, the number of applications based on propensity scores (12) is in hundreds.

We suggest several potential explanations for the wider adoption of propensity score-based approaches. First, the gold standard for causal inference in healthcare has long been the Randomized Controlled Trial (RCT). Propensity score-based approaches provide methods that mimic RCTs while using large-scale, observational data. Secondly, as we mapped out in **Table 1**, propensity score-based approaches offer relatively low variance at the risk of higher bias, which is consistent with medical applications where the goal to minimize patient harm outweighs the potential to increase benefits for a few. Third, there is an issue of timing, newer methods have simply been in existence for a shorter period of time and therefore have had less chance for adoption. However, this answer is least satisfying because many of the newer machine learning approaches have been successfully applied in many other fields such as gaming, online shopping, and advertising (4). Additionally, many machine learning-based causal models have been around for a long time. For example, as of the time this paper is written, the BART model (44) has existed for over a decade, and yet we have not seen many clinical applications of it. A fourth potential reason for lower adoption of purely machine learning based approaches is method explainability. In healthcare, where lives are frequently at stake, the requirement for methods that are explainable to a wide audience are significantly higher than other fields, where effectiveness alone may be sufficient.

We believe that lower historical adoption of more modern observational causal inference approaches is sensible, but that it also represents a gap in the field, especially given the potential promise of more personalized medicine using ITE-type estimators. This gap could potentially be closed in the near future by collaborative pairing of biostatisticians and machine learning scientists with clinicians.

FLOWCHART FOR ALGORITHM SELECTION

In this section we provide a guide in **Figure 1** to help the healthcare community choose which algorithm to use in estimating treatment effects based on the target-population intervention sizes, domain knowledge about the treatment, and track record of healthcare applications of the algorithm. While every problem is unique, and individual judgement must always

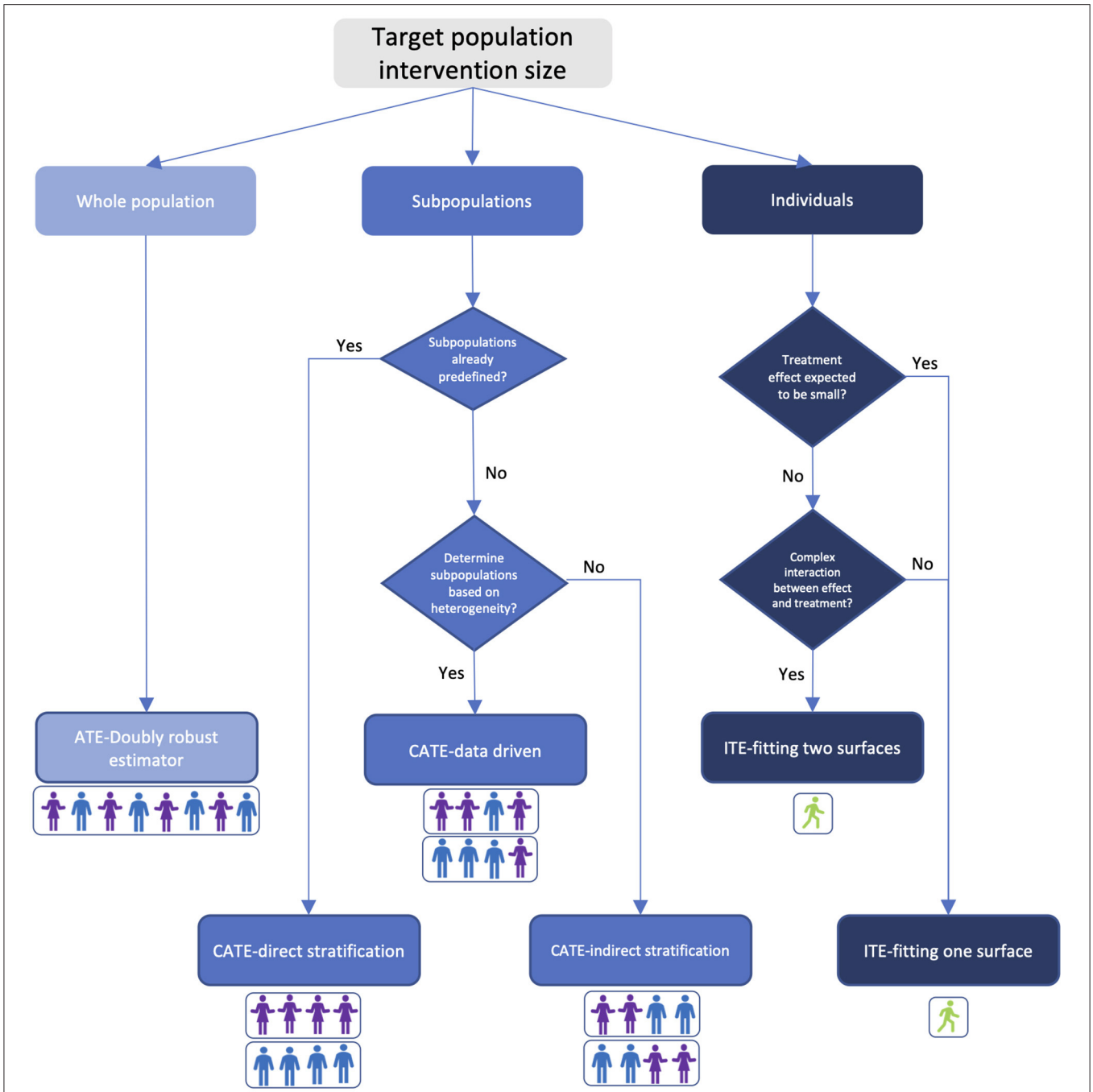

FIGURE 1 | Treatment effect estimator selection guide based on target-population intervention size and prior knowledge. Colors in the figure indicate bias-variance tradeoff. Light blue: high bias and low variance; blue: medium bias and variance; dark blue: low bias and high variance. Person icons under each estimator illustrate the composition of the targeted population.

be exercised, this flowchart can act as a starting point to determine which algorithmic approach may be most appropriate.

DISCUSSION

In this paper we reviewed the literature on causal inference with a focus on clinical settings, in light of recent advances in machine learning and large scale EHR adoption. With this review, the algorithm selection guide, and the summary table, we hope to help researchers and healthcare stakeholders gain better understanding of causal inference and make informed decisions on what estimator to use in their daily practices when many choices are on the table.

We have observed that sophisticated causal models based on state-of-the-art machine learning have not been widely applied in clinical studies for a myriad of reasons such as lack of similarity to RCTs and explainability (Section Clinical Applications of Causal Inference), computational intractability of these models, and the healthcare participants being highly conservative when adopting new models. To address the same issue and improve model transparency, a MI-CLAIM check list in Norgeot et al. (109) was proposed regarding the study design of projects, preparation and usage of data, model selection, performance evaluation, model validation, and data pipelines. Our review stresses the importance to follow these guidelines to promote trust on sophisticated models among clinical practitioners.

There are some limitations of the review. First, it may not be exhaustive and include every approach. Causal inference is a very broad topic. While we can limit our review to a specific topic to be exhaustive, it is also important to survey the entire field of causal inference, thus sacrificing the completeness to some degree. Second, causal inference approaches are grouped into ATE, CATE, and ITE categories in this review. These categories might not be mutually exclusive. Such classification, however, does provide an intuitive way for medical professionals to understand causal inference from patient perspectives. Third, there are certain limitations of using citations to rank the applications. For instance, an algorithm applied in clinics might not have been published. Additionally, for a recent work, the citation number might be low, and might not accurately reflect the application potential of the work. Fourth, **Table 1** and **Figure 1** do not cover all the details of choosing an algorithm, nor do they lead a user to a specific algorithm. They were designed to provide all healthcare participants with an initial but intuitive guide on what family of algorithms to choose for their studies. Finally, our search to find published applications of causal models may not be exhaustive. The search results show that the application disparity of different models is so huge that a different (and potentially more comprehensive) search strategy will unlikely change our conclusions and insights in any significant way.

There is a view in the literature that causal inference is just plain statistical inference, especially after the causal assumptions and parameters are identified (110). The role of causal inference with respect to statistical analysis remains a debate. This debate is out of scope for this paper. We refer to the reviewed models as causal inference models without endorsing any particular view on this matter, but simply use this name to refer to the statistical inference models that reveal causal relationships.

In summary, we reviewed a diverse and complex field of causal inference applied in health care. We distilled the many approaches into three algorithmic families based on the target-population intervention size. We explained the approach type, population size, and bias-variance tradeoff. We then investigated the clinical application of each of the approaches. We finally consolidate all the information into an algorithm selection guide for both researchers and other healthcare stakeholders to decide on which algorithm is applicable to their studies.

AUTHOR CONTRIBUTIONS

JS conducted the research and developed the figures. BN conceived of the research topic and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

The authors JS and BN are employed by Anthem, Inc. The funder had no other involvement in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Chris Jensen for his assistance submitting this work, Paula Alves for her helpful discussions, Abhishaike Mahajan, Daniel Brown, and Dong Wang for proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.864882/full#supplementary-material

REFERENCES

1. *Health IT Dashboard*. Available online at: https://dashboard.healthit.gov/quickstats/pages/physician- ehr-adoption-trends.php
2. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *J Am Med Inform Assoc*. (2017) 24:1142–8. doi: 10.1093/jamia/ocx080
3. Goodman KE, Pineles L, Magder LS, Anderson DJ, Ashley ED, Polk RE, et al. Electronically available patient claims data improve models for comparing antibiotic use across hospitals: results from 576 US facilities. *Clin Infect Dis*. (2021) 73:e4484–92. doi: 10.1093/cid/cia a1127
4. Das S, Dey A, Pal A, Roy N. Applications of artificial intelligence in machine learning: review and prospect. *Int J Comput Applic*. (2015) 115:31–41. doi: 10.5120/20182-2402
5. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A survey on causal inference. *ACM Trans Knowl Discov Data*. (2021) 15:1–46. doi: 10.1145/3444944
6. Guo R, Cheng L, Li J, Hahn PR, Liu H. A survey of learning causality with data: problems and methods. *ACM Comput Surv*. (2020) 53:1–37. doi: 10.1145/3397269
7. Ding P, Li F. Causal inference: a missing data perspective. *Stat Sci*. (2018) 33:214–37. doi: 10.1214/18-STS645
8. Landsittel D, Srivastava A, Kropf K. A narrative review of methods for causal inference and associated educational resources. *Qual Manag Health Care*. (2020) 29:260–9. doi: 10.1097/QMH.0000000000000276
9. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. (2010) 25:1–21. doi: 10.1214/09-STS313
10. Shah RB, Laupacis A, Hux EJ, Austin CP. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. (2005) 58:550–9. doi: 10.1016/j.jclinepi.2004.10.016

11. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *Surg Acqu Cardiovasc Dis.* (2007) 134:1128–35. doi: 10.1016/j.jtcvs.2007.07.021
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrilca.* (1983) 70:41–55. doi: 10.1093/biomet/70.1.41
13. Robins JM, Rotnitzky A, Zhao LP. Marginal structural models and causal inference in epidemiology. *Epidemiology.* (2000) 11:550–60. doi: 10.1097/00001648-200009000-00011
14. Pearl J. Causal diagrams for empirical research. *Biometrika.* (1995) 82:669–88. doi: 10.1093/biomet/82.4.669
15. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann (1988). doi: 10.1016/B978-0-08-051489-5.50008-4
16. Lauritzen SL. *Graphical Models*. Oxford: Clarendon Press (1996).
17. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period–applications to control of the healthy workers survivor effect. *Math Model.* (1986) 7:1393–512. doi: 10.1016/0270-0255(86)90088-6
18. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. *Stat Sci.* (1999) 14:29–46. doi: 10.1214/ss/1009211805
19. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc.* (1996) 91:444–55. doi: 10.1080/01621459.1996.10476902
20. Pearl J. Comment: graphical models, causality and intervention. *Stat Sci.* (1993) 8:266–9. doi: 10.1214/ss/1177010894
21. Barter RL. *Visualization, Prediction, and Causal Inference: Applications in Healthcare*. UC Berkeley Electronic Theses and Dissertations, University of California, Berkeley, CA, United States (2019).
22. Thomas S, Richardson JMR. Single world intervention graphs: a primer. In: *Second UAI Workshop on Causal Structure Learning*. Bellevue, WA (2013).
23. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* (1985) 39:33–8. doi: 10.1080/00031305.1985.10479383
24. Rosenbaum PR. *Observational Studies*. New York, NY: Springer-Verlag (2002). doi: 10.1007/978-1-4757-3692-2
25. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat.* (1993) 2:405–20. doi: 10.1080/10618600.1993.10474623
26. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med.* (2008) 27:2037–49. doi: 10.1002/sim.3150
27. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Indian J Stat Ser A*. (1973) 35:417–46.
28. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica.* (2003) 71:1161–89. doi: 10.1111/1468-0262.00442
29. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* (1984) 79:516–24. doi: 10.1080/01621459.1984.10478078
30. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J Am Stat Assoc.* (1979) 74:318–28. doi: 10.1080/01621459.1979.10482513
31. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* (1994) 89:846–66. doi: 10.1080/01621459.1994.10476818
32. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol.* (2011) 173:761–7. doi: 10.1093/aje/kwq439
33. Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat.* (2006) 2:11. doi: 10.2202/1557-4679.1043
34. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol.* (2016) 185:65–73. doi: 10.1093/aje/kww165
35. Morgan SL, Winship C. *Counterfactuals and Causal Inference*. Cambridge University Press (2014). doi: 10.1017/CBO9781107587991
36. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press (2015). doi: 10.1017/CBO9781139025751
37. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Nat Acad Sci USA*. (2016) 113:7353–60. doi: 10.1073/pnas.1510489113
38. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Chapman and Hall/CRC (1984).
39. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
40. Wang P, Sun W, Yin D, Yang J, Chang Y. Robust tree-based causal inference for complex ad effectiveness analysis. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai). (2015). p. 67–76. doi: 10.1145/2684822.2685294
41. Meid AD, Ruff C, Wirbka L, Stoll F, Seidling HM, Groll A, et al. Using the causal inference framework to support individualized drug treatment decisions based on observational healthcare data. *Clin Epidemiol.* (2020) 12:1223–34. doi: 10.2147/CLEP.S274466
42. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat.* (2011) 20:217–40. doi: 10.1198/jcgs.2010.08162
43. Chipman HA, George EI, McCulloch RE. Bayesian ensemble learning. In: *NIPS'06: Proceedings of the 19th International Conference on Neural Information Processing Systems* (Vancouver). (2006). p. 265–72.
44. Chipman HA, George EI, McCulloch RE. BART: bayesian additive regression trees. *Ann Appl Stat.* (2010) 4:266–98. doi: 10.1214/09-AOAS285
45. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med.* (2011) 30:2867–80. doi: 10.1002/sim.4322
46. Nie X, Wager S. Quasi-Oracle estimation of heterogeneous treatment effects. *Biometrika*. (2020) 108:299–319. doi: 10.1093/biomet/asaa076
47. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat Med*. (2018) 37:3309–24. doi: 10.1002/sim.7820
48. Alaa AM, Schaar M. Limits of estimating heterogeneous treatment effects: guidelines for practical algorithm design. In *International Conference on Machine Learning*. Stockholm (2018).
49. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Anal.* (2020) 15:965–1056. doi: 10.1214/19-BA1195
50. Lu M, Sadiq S, Feaster DJ, Ishwarana H. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat.* (2018) 27:209–19. doi: 10.1080/10618600.2017.1356325
51. Powers S, Qian J, Jung K, Schuler A. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med.* (2018) 37:1767–87. doi: 10.1002/sim.7623
52. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci USA*. (2019) 116:4156–65. doi: 10.1073/pnas.1804597116
53. Alaa AM, Schaar M. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *31st International Conference on Neural Information Processing Systems*. Long Beach, CA (2017).
54. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, NSW (2017).
55. Hassanpour N, Greiner R. CounterFactual regression with importance sampling weights. In*: Twenty-Eighth International Joint Conference on Artificial Intelligence* (Macao). (2019). doi: 10.24963/ijcai.2019/815
56. Belthangady C, Stedden W, Norgeot B. Minimizing bias in massive multi-arm observational studies with BCAUS: balancing covariates automatically using supervision. *BMC Med Res Methodol.* (2021) 21:190. doi: 10.1186/s12874-021-01383-x
57. Bengio Y, Courville AC, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* (2013) 35:1798–828. doi: 10.1109/TPAMI.2013.50
58. Shi C, Blei DM, Veitch V. Adapting neural networks for the estimation of treatment effects. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, BC (2019).

59. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLos Med*. (2009) 6:e1000097. doi: 10.1371/journal.pmed.1000097
60. Requena CC, Muriel A, Peñuelas Ó. Analysis of causality from observational studies and its application in clinical research in intensive care medicine. *Med Intens*. (2018) 42:292–300. doi: 10.1016/j.medine.2018.01.010
61. Delaney JW, Pinto R, Long J, Lamontagne F, Adhikari NK, Kumar A, et al. The influence of corticosteroid treatment on the outcome of influenza A(H1N1pdm09)-related critical illness. *Crit Care*. (2016) 20:75. doi: 10.1186/s13054-016-1230-8
62. Graham DJ, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, et al. Cardiovascular, bleeding, and mortality risks in elderly medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*. (2015) 131:157–64. doi: 10.1161/CIRCULATIONAHA.114.012061
63. Song M, Fung TT, Hu FB, Willett WC, Longo VD, Chan AT, et al. Association of animal and plant protein intake with all-Cause and cause-specific mortality. *JAMA Intern Med*. (2016) 176:1453–63. doi: 10.1001/jamainternmed.2016.4182
64. Kang BJ, Koh Y, Lim CM, Huh JW, Baek S, Han M, et al. Failure of high-flow nasal cannula therapy may delay intubation and increase mortality. *Intensive Care Med*. (2015) 41:623–32. doi: 10.1007/s00134-015-3693-5
65. Jones SE, Maddocks M, Kon SSC, Canavan JL, Nolan CM, Clark AL, et al. Sarcopenia in COPD: prevalence, clinical correlates and response to pulmonary rehabilitation. *Thorax*. (2015) 70:213–8. doi: 10.1136/thoraxjnl-2014-206440
66. Ziff OJ, Samra M, Kirchhof P, Steeds RP, Kotecha D. Safety and efficacy of digoxin: systematic review and meta-analysis of observational and controlled trial data. *BMJ*. (2015) 351:h4451. doi: 10.1136/bmj.h4451
67. Blackstone EH, Suri RM, Rajeswaran J, Babaliaros V, Douglas PS, Fearon WF, et al. Propensity-Matched comparisons of clinical outcomes after transapical or transfemoral transcatheter aortic valve replacement. *Circulation*. (2015) 131:1989–2000. doi: 10.1161/CIRCULATIONAHA.114.012525
68. Chang SH, Chou IJ, Yeh YH, Chiou MJ, Wen MS, Kuo CT, et al. Association between use of non–vitamin k oral anticoagulants with and without concurrent medications and risk of major bleeding in nonvalvular atrial fibrillation. *JAMA*. (2017) 318:1250–9. doi: 10.1001/jama.2017.13883
69. Bangalore S, Guo Y, Zaza Samadashvili, Blecker S, Xu J, Hannan EL. Everolimus-eluting stents or bypass surgery for multivessel coronary disease list of authors. *N Engl J Med*. (2015) 372:1213–22. doi: 10.1056/NEJMoa1412168
70. Kost K, Lindberg L. Pregnancy intentions, maternal behaviors, and infant health: investigating relationships with new measures and propensity score analysis. *Demography*. (2015) 52:83–111. doi: 10.1007/s13524-014-0359-9
71. Grool AM, Aglipay M, Momoli F, Meehan WP. Association between early participation in physical activity following acute concussion and persistent postconcussive symptoms in children and adolescents. *JAMA*. (2016) 316:2504–14. doi: 10.1001/jama.2016.17396
72. Snowden JM, Caughey AB, Cheng YW. Planned out-of-hospital birth and birth outcomes. *N Engl J Med*. (2015) 373:2642–53. doi: 10.1056/NEJMsa1501738
73. Han HS, Shehta A, Ahn S, Yoon YS, Cho JY, Choi Y. Laparoscopic versus open liver resection for hepatocellular carcinoma: case-matched study with propensity score matching. *J Hepatol*. (2015) 63:643–50. doi: 10.1016/j.jhep.2015.04.005
74. Prati F, Romagnoli E, Burzotta F, Limbruno U, Gatto L, Manna AL, et al. Clinical impact of OCT findings during PCI: the CLI-OPCI II study. *J Am Coll Cardiol Img. Nov*. (2015) 8:1297–305. doi: 10.1016/j.jcmg.2015.08.013
75. Prescott HC, Osterholzer JJ, Langa KM, Angus DC, Iwashyna TJ. Late mortality after sepsis: propensity matched cohort study. *BMJ*. (2016) 353:i2357. doi: 10.1136/bmj.i2375
76. Xie Y, Bowe B, Li T, Xian H, Yan Y, Al-Aly Z. Risk of death among users of proton pump inhibitors: a longitudinal observational cohort study of United States veterans. *BMJ Open*. (2017) 7:e015735. doi: 10.1136/bmjopen-2016-015735
77. Cheetham TC, An J, Jacobsen SJ. Association of testosterone replacement with cardiovascular outcomes among men with androgen deficiency. *JAMA Intern Med*. (2017) 177:491–9. doi: 10.1001/jamainternmed.2016.9546
78. Izurieta HS, Chillarige Y, Kelman J, Wei Y, Lu Y, Xu W, et al. Relative effectiveness of cell-cultured and egg-based influenza vaccines among elderly persons in the United States, 2017–2018. *J Infect Dis*. (2019) 220:1255–64. doi: 10.1093/infdis/jiy716
79. Bailly S, Leroy O, Montravers P, Constantin JM, Dupont H, Guillemot D, et al. Antifungal de-escalation was not associated with adverse outcome in critically ill patients treated for invasive candidiasis: post hoc analyses of the AmarCAND2 study data. *Intensive Care Med*. (2015) 41:1931–40. doi: 10.1007/s00134-015-4053-1
80. Feng M, McSparron JI, Kien DT, Stone DJ, Roberts DH, Schwartzstein RM, et al. Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database. *Intensive Care Med*. (2018) 44:884–92. doi: 10.1007/s00134-018-5208-7
81. Chaffee BW, Cheng J, Featherstone JD. Baseline caries risk assessment as a predictor of caries incidence. *J Dent*. (2015) 43:518–24. doi: 10.1016/j.jdent.2015.02.013
82. Blumberg SJ, Zablotsky B, Avila RM, Colpe LJ, Pringle BA, Kogan MD. Diagnosis lost: differences between children who had and who currently have an autism spectrum disorder diagnosis. *Autism*. (2015) 20:783–95. doi: 10.1177/1362361315607724
83. Barnes H, Rearden J, McHugh MD. Magnet hospital recognition linked to lower central line-associated bloodstream infection rates. *Res Nurs Health*. (2016) 39:96–104. doi: 10.1002/nur.21709
84. Breslau J, Leckman-Westin E, Yu H, Han B, Pritam R, Guarasi D, et al. Impact of a mental health based primary care program on quality of physical health care. *Admin Policy Ment Health Ment Health Serv Res*. (2018) 45:276–85. doi: 10.1007/s10488-017-0822-1
85. Simon TG, Duberg AS, Aleman S, Chung RT, Chan AT, Ludvigsson JF. Association of aspirin with hepatocellular carcinoma and liver-related mortality. *N Engl J Med*. (2020) 382:1018–28. doi: 10.1056/NEJMoa1912035
86. Ajmal A, Gessert CE, Johnson BP, Renier CM, Palcher JA. Effect of angiotensin converting enzyme inhibitors and angiotensin receptor blockers on hemoglobin levels. *BMC Res Notes*. (2013) 6:443. doi: 10.1186/1756-0500-6-443
87. Millett PJ, Espinoza C, Horan MP, Ho CP, Warth RJ, Dornan GJ, et al. Predictors of outcomes after arthroscopic transosseous equivalent rotator cuff repair in 155 cases: a propensity score weighted analysis of knotted and knotless self-reinforcing repair techniques at a minimum of 2 years. *Arch Orthop Trauma Surg*. (2017) 137:1399–408. doi: 10.1007/s00402-017-2750-7
88. Reed GW, Abdallah MS, Shao M, Wolski K, Wisniewski L, Yeomans N, et al. Effect of aspirin coadministration on the safety of celecoxib, naproxen, or ibuprofen. *J Am Coll Cardiol*. (2018) 71:1741–51. doi: 10.1016/j.jacc.2018.02.036
89. Kawasaki R, Konta T, Nishida K. Lipid-lowering medication is associated with decreased risk of diabetic retinopathy and the need for treatment in patients with type 2 diabetes: a real-world observational analysis of a health claims database. *Diabetes Obes Metab J Pharmacol Ther*. (2018) 20:2351–60. doi: 10.1111/dom.13372
90. Soumerai SB, McLaughlin TJ, Spiegelman D, Hertzmark E, Thibault G, Goldman L. Adverse outcomes of underuse of β-blockers in elderly survivors of acute myocardial infarction. *JAMA*. (1997) 277:115–21. doi: 10.1001/jama.277.2.115
91. Hannan EL, Wu C, Walford G, Culliford AT, Gold JP, Smith CR, et al. Drug-Eluting stents vs. coronary-artery bypass grafting in multivessel coronary disease. *N Engl J Med*. (2008) 358:331–41. doi: 10.1056/NEJMoa071804
92. Ip A, Berry DA, Hansen E, Goy AH, Pecora AL, Sinclaire BA, et al. Hydroxychloroquine and tocilizumab therapy in COVID-19 patients—An observational study. *PLoS ONE*. (2020) 15:e0237693. doi: 10.1371/journal.pone.0237693
93. Lindahl B, Baron T, Erlinge D, Hadziosmanovic N, Nordenskjöld A, Gard A, et al. Medical therapy for secondary prevention and long-term outcome in patients with myocardial infarction with nonobstructive coronary artery disease. *Circulation*. (2017) 135:1481–9. doi: 10.1161/CIRCULATIONAHA.116.026336
94. Kuck KH, Brugada J, Fürnkranz A, Chun KRJ, Metzner A, Ouyang F, et al. Impact of female sex on clinical outcomes in the FIRE AND ICE trial of catheter ablation for atrial fibrillation. *Circulation Arrhythm Electrophysiol*. (2018) 11:e006204. doi: 10.1161/CIRCEP.118.006204

95. Kushida CA, Nichols DA, Holmes TH, Quan SF, Walsh JK, Gottlieb DJ, et al. Effects of continuous positive airway pressure on neurocognitive function in obstructive sleep apnea patients: the apnea positive pressure long-term efficacy study (APPLES). *Sleep*. (2012) 35:1593–602. doi: 10.5665/sleep.2226
96. Conway PH, Cnaan A, Zaoutis T, Henry BV, Grundmeier RW, Keren R. Recurrent urinary tract infections in children risk factors and association with prophylactic antimicrobials. *JAMA*. (2007) 298:179–86. doi: 10.1001/jama.298.2.179
97. Hackam GD, Pharm MM, Li P, Redelmeier DA. Statins and sepsis in patients with cardiovascular disease: a population-based cohort analysis. *Lancet*. (2006) 367:413–8. doi: 10.1016/S0140-6736(06)68041-0
98. Vikram HR, Buenconsejo J, Hasbun R, Quagliarello VJ. Impact of valve surgery on 6-month mortality in adults with complicated, left-sided native valve endocarditis a propensity analysis. *JAMA*. (2003) 290:3207–14. doi: 10.1001/jama.290.24.3207
99. Martin D, Glass TA, Bandeen-Roche K, Todd AC, Shi W, Schwartz BS. Association of blood lead and tibia lead with blood pressure and hypertension in a community sample of older adults. *Am J Epidemiol*. (2006) 163:467–78. doi: 10.1093/aje/kwj060
100. Hannan EL, Racz M, Holmes DR, King SB III, Walford G, Ambrose JA, et al. Impact of completeness of percutaneous coronary intervention revascularization on long-term outcomes in the stent era. *Circulation*. (2006) 113:2406–12. doi: 10.1161/CIRCULATIONAHA.106.612267
101. Wong YN, Mitra N, Hudes G, Localio R, Schwartz JS, Wan F, et al. Survival associated with treatment vs observation of localized prostate cancer in elderly men. *JAMA*. (2006) 296:2683–93 doi: 10.1001/jama.296.22.2683
102. Ferguson TB, Coombs LP, Peterson ED. Preoperative β-blocker use and mortality and morbidity following CABG surgery in north america. *JAMA*. (2002) 287:2221–7. doi: 10.1001/jama.287.17.2221
103. Potosky AL, Harlan LC, Kaplan RS, Johnson KA, Lynch CF. Age, sex, and racial differences in the use of standard adjuvant therapy for colorectal cancer. *J Clin Oncol*. (2002) 20:1192–202. doi: 10.1200/JCO.2002.20.5.1192
104. Ahmed A, Rich MW, Sanders PW, Perry GJ, Bakris GL, Zile MR, et al. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *Am J Cardiol*. (2007) 99:393–8. doi: 10.1016/j.amjcard.2006.08.042
105. Graham C, Lewis S, Forbes J, Mead G, Hackett ML, Hankey GJ, et al. The FOCUS, AFFINITY and EFFECTS trials studying the effect(s) of fluoxetine in patients with a recent stroke: statistical and health economic analysis plan for the trials and for the individual patient data meta-analysis. *Trials*. (2017) 18:627. doi: 10.1186/s13063-017-2385-6
106. Lee K, Small DS, Hsu JY, Silber JH, Rosenbaum PR. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *J Am Stat Assoc*. (2018) 181:535–46. doi: 10.1111/rssa.12298
107. Guo S, Wang YL, Li Y, Jin L, Xiong M, Ji QH, et al. Significant SNPs have limited prediction ability for thyroid cancer. *Cancer Med*. (2014) 3:731–5. doi: 10.1002/cam4.211
108. Esposito MH, Lee H, Hicken MT, Porter LC, Herting JR. The consequences of contact with the criminal justice system for health in the transition to adulthood. *Longit Life Course Stud*. (2017) 8:57–74. doi: 10.14301/llcs.v8i1.405
109. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. (2020) 26:1320–4. doi: 10.1038/s41591-020-1041-y
110. Maya L, Petersen MJL. Causal models and learning from data. *Epidemiology*. (2014) 25:418–26. doi: 10.1097/EDE.0000000000000078

Conflict of Interest: JS and BN are employed by Anthem, Inc.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/frai.2022.727486

A Novel Bayesian General Medical Diagnostic Assistant Achieves Superior Accuracy With Sparse History

A Performance Comparison of 7 Online Diagnostic Aids and Physicians

Alicia M. Jones and Daniel R. Jones*

Eureka Clinical Computing, Eureka Springs, AR, United States

Online AI symptom checkers and diagnostic assistants (DAs) have tremendous potential to reduce misdiagnosis and cost, while increasing the quality, convenience, and availability of healthcare, but only if they can perform with high accuracy. We introduce a novel Bayesian DA designed to improve diagnostic accuracy by addressing key weaknesses of Bayesian Network implementations for clinical diagnosis. We compare the performance of our prototype DA (MidasMed) to that of physicians and six other publicly accessible DAs (Ada, Babylon, Buoy, Isabel, Symptomate, and WebMD) using a set of 30 publicly available case vignettes, and using only sparse history (no exam findings or tests). Our results demonstrate superior performance of the MidasMed DA, with the correct diagnosis being the top ranked disorder in 93% of cases, and in the top 3 in 96% of cases.

Keywords: Bayesian medical diagnosis, symptom checkers, general medical diagnostic assistant, diagnostic performance, Bayesian network, comparison of physicians with AI decision support, AI medical diagnosis, diagnostic decision support system

OPEN ACCESS

Edited by:
Holger Fröhlich, Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany

Reviewed by:
Ibrahem Kandel, Universidade NOVA de Lisboa, Portugal
Lisa Gandy, Central Michigan University, United States

***Correspondence:**
Daniel R. Jones
jones.ecc@gmail.com

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *18 June 2021*
Accepted: *21 June 2022*
Published: *22 July 2022*

Citation:
Jones AM and Jones DR (2022) A Novel Bayesian General Medical Diagnostic Assistant Achieves Superior Accuracy With Sparse History. Front. Artif. Intell. 5:727486. doi: 10.3389/frai.2022.727486

INTRODUCTION

Online AI symptom checkers and diagnostic assistants (DAs) have tremendous potential to reduce misdiagnosis and cost, while increasing the quality, convenience, and availability of healthcare, but only if they can perform with high accuracy (Millenson et al., 2018; Van Veen et al., 2019; Rowland et al., 2020). Machine Learning (ML) and Bayesian Networks (BNs) are promising technologies in healthcare, but both have limitations for general medical diagnosis. Despite major advances in the application of ML to narrow biomedical applications (Beede et al., 2020; Liu et al., 2020; McKinney et al., 2020), challenges remain for its application to *general* medical diagnosis, including the inability to model causal inference (Velikova et al., 2014; Richens et al., 2020), semantic relationships including subtypes ("is-a" and "part-of"), logic, and heuristics; and lack of interpretability. Furthermore, challenges remain in training or educating DAs with electronic medical record (EMR) data, including proper interpretation of incomplete or missing data (Nikovski, 2000), unreliable labels and label leakage, bias (Ghassemi et al., 2020), and the fact that EMRs are designed to document and support care and reimbursement and to minimize legal risks, rather than to describe disorders.

The use of Bayesian approaches for medical diagnosis is well-documented, from early expert systems (Yu et al., 1988; Shwe et al., 1990; Barnett et al., 1998) to today's chatbot triage and symptom checkers (Zagorecki et al., 2013; Baker et al., 2020). But thus far they have fallen short of the desired accuracy despite incremental improvements (Lemmer and Gossink, 2004; Antonucci, 2011; Richens et al., 2020). In previous studies such DAs have underperformed physicians in diagnostic accuracy (Semigran et al., 2015, 2016; Millenson et al., 2018; Chambers et al., 2019; Yu et al., 2019). For example, Semigran et al. (2015) evaluated the performance of 23 symptom checkers using case vignettes, and found they ranked the correct diagnosis first 34% of the time, and in the top 3 in 51% of cases. In a subsequent paper (Semigran et al., 2016) compared symptom checkers to physicians, and showed much better performance for physicians, who ranked the target diagnosis #1 in 72.1% of cases, vs. only 34% for the symptom checkers. A more recent paper (Baker et al., 2020), using 30 of the case vignettes tested in Semigran et al. (2015) and Semigran et al. (2016), reported performance comparable to physicians: the Babylon system ranked the target diagnosis #1 for 70% of the vignettes and in the top 3 for 96.7%, compared to 75.3 and 90.3%, respectively, for physicians. But even the benchmark of obtaining physician diagnostic accuracy leaves much to be desired, with reported physician diagnostic error rates of 10–24% or greater (Graber, 2012; Meyer et al., 2013; Baker et al., 2020). Diagnostic errors are the leading cause of paid malpractice claims (28.6%), and are responsible for the highest proportion of total payments (35.2%) (Tehrani et al., 2013). Diagnostic errors were almost twice as likely to be associated with patient death as other types of errors (e.g., treatment, surgery, medication, or obstetrics errors). Almost 70% of diagnostic errors occurred in the outpatient setting (Tehrani et al., 2013).

BNs model causal inference using Bayes' theorem. They offer a formal method for representing an evolving process of refining the posterior probabilities of outcomes based on the likelihood of relevant data. This approach is particularly suitable for diagnosis, where clinicians formulate an initial differential diagnosis based on the patient chief complaint, and then proceed to refine the diagnosis based on additional data obtained from the patient interview, exams, tests, and treatment outcomes. In this iterative process, each differential diagnosis ranks the likelihood of each contending disorder, and provides priorities for the next data items to ascertain.

Given a joint random variable $\boldsymbol{X} = X_1, \ldots, X_N$, a Bayesian Network (BN) offers a compact representation of its local conditional probability distributions (Koller and Friedman, 2009). Formally, a Bayesian Network is defined as a pair BN $= (G, P)$, where G is a directed acyclic graph (DAG) and P is the joint probability distribution of $\boldsymbol{X}$ as specified by the conditional probability tables (CPTs) of the graph nodes. The graph $G = (V, E)$, is comprised of nodes or vertices V and directed arcs or edges $E \subseteq V \times V$. Each node in V represents a distinct random variable in $\boldsymbol{X}$, and each arc in E represents the conditional probability of the child node given its parent. Every node is conditionally independent of its non-parent non-descendants, given its parents. It follows that the joint probability distribution $P(\boldsymbol{X})$ reduces to the product of the conditional probability distributions at each node (local Markov property), and can be written as:

$$P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i|\pi_i) \tag{1}$$

where π_i is the state of the joint variable defined by the elements of $\boldsymbol{X}$ that are the parents of X_i (Fagiuoli and Zaffalon, 1998; Antonucci, 2011).

The size of the CPT describing the joint probability distribution at a node grows exponentially with the number of inputs (parents). For problems involving a large number of variables and/or dense graphs, computational complexity and/or lack of sufficient data can make this approach impractical. The leaky noisy-OR function (Henrion, 1987; Antonucci, 2011) is a popular technique for reducing the input parameter requirements from exponential to linear (for binary variables). It does so by assuming the parent nodes are conditionally independent given their joint child. With this assumption, the joint probability distribution of the child node simplifies to:

$$P(x_i|\pi_i) = 1 - (1 - n_i) \prod_{x_j \in \pi_i} \left(1 - P(x_i|x_j)\right)^{\delta_j} \tag{2}$$

where $P(x_i|x_j)$ is the conditional probability of the child node given parent X_j, and $\delta_j = 1$ if $x_j = true$ and 0 if it is *false*. Equation (2) can be interpreted as meaning that X_j only affects change when it is present. Ignoring the $(1 - n_i)$ term for a moment, we see this is simply the probability formula for the union of independent events, i.e., $P\left(\bigcup_i A_i\right) = 1 - \prod_i \left(\left(1 - P(A_i)\right)\right)$. The variable n_i is a noise term, which is optionally a function of X_i, and represents unmodeled causes of X_i assumed to be present.

A classifier can be defined in conjunction with a BN by assigning each node to 1 of 3 types: (1) input, data, features, or evidence; (2) outputs or class labels; and optionally (3) intermediate or hidden nodes. Given K possible outputs, $y_1, \ldots, y_K$, and L inputs, $x_1, \ldots, x_L$, the classifier selects the output node $\hat{y}$ such that

$$\hat{y} = argmax_{i \in \{1,\ldots,K\}} P(y_i|x_1, \ldots, x_L) \tag{3}$$

where *argmax* selects the maximum argument, i.e., the output node that maximizes $P(y_i|x_1, \ldots, x_L)$. Using Bayes Theorem and assuming the output nodes are mutually independent, Equation (3) reduces to

$$\hat{y} = argmax_{i \in \{1,\ldots,K\}} P(y_i) \cdot P\left(x_1, \ldots, x_L|y_i\right) \tag{4}$$

where $P\left(y_i\right)$ is the a priori probability of y_i. In the special case where the variables X_i are independent, we obtain the naïve Bayes classifier (Koller and Friedman, 2009)

$$\hat{y} = argmax_{i \in \{1,\ldots,K\}} P(y_i) \cdot \prod_{j=1}^{L} P\left(x_j|y_i\right) \tag{5}$$

It is important to keep in mind the assumptions that lead to the simplifications of Equations (4) and (5). Medical diagnosis

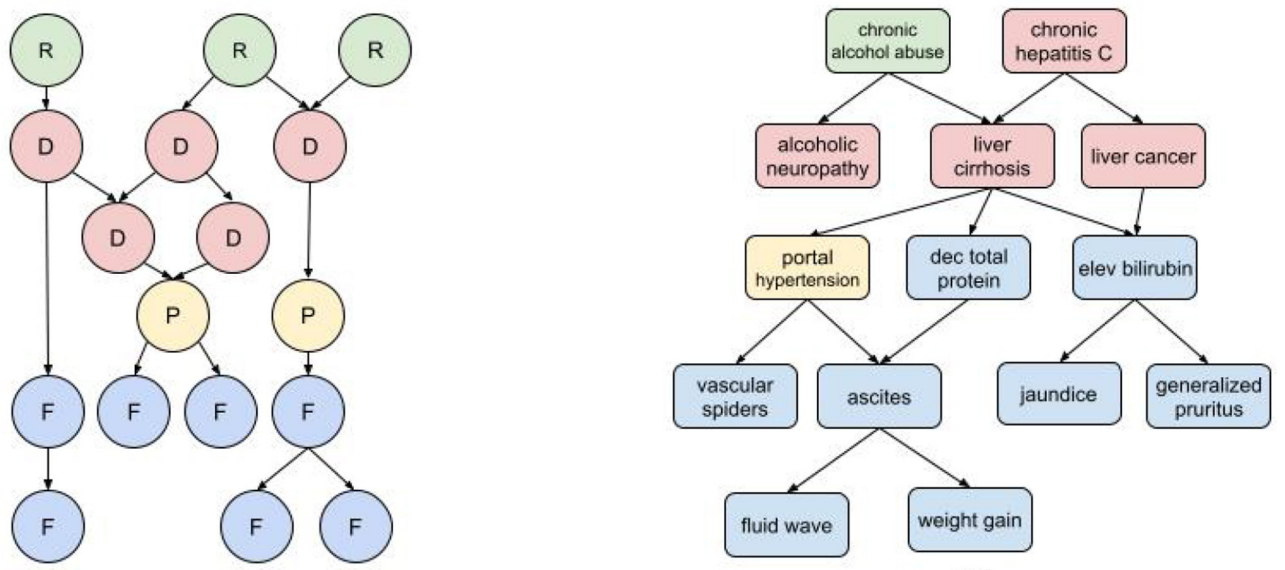

FIGURE 1 | Diagnostic BN hierarchy **(A)** Generic fragment where each node represents a risk factor (R) disease (D), pathophysiological state (P), or findings (F); **(B)** BN fragment for liver cirrhosis.

is one domain in which these assumptions are not always valid, resulting in excessively degraded classification.

In a medical diagnostic BN (**Figure 1**) the input nodes represent all known risk factors and findings (i.e., symptoms, examination results, and test results), while the output nodes are all possible diagnoses. There may also be intermediate nodes representing pathophysiological states or mechanisms. As indicated by the causal arrows, risk factors increase the likelihood of diseases; diseases cause other diseases, pathophysiological states, and findings; physiological states cause findings (and sometimes other physiological states); and findings may cause other findings. For a given set of patient inputs we want to determine the most probable diagnoses using both forward and backward inference.

The characterization of nodes as risk factors, findings, pathophysiological states, and disorders can be governed by somewhat arbitrary nosological distinctions. For example, dehydration is a pathophysiological state with multiple findings (e.g., decreased urine output, dry mucus membranes, dizziness, hypotension), and can be caused by multiple disorders such as acute gastroenteritis and uncontrolled diabetes. But dehydration is also used as a diagnosis when other causal disorders are ruled out, and it can be attributed to, e.g., prolonged exertion in heat without sufficient hydration (a risk factor). The findings of dehydration can be attributed to its causal disorders, but they tend to cluster as a distinct subpopulation in patients with the causal disorders that develop dehydration. The distinction between risk factors and findings can also be ambiguous. For example, obesity is both a risk factor for developing type II diabetes and also a finding of diabetes and other metabolic disorders. And while some findings can cause other findings, it's important not to confuse temporal progression with causality. For example, in an infectious disorder, fever may precede a rash, but doesn't cause it.

Figure 2 shows typical diagnostic BN configurations. In **Figure 2A** a disorder causes 2 findings (F_1, F_2). These findings may be considered conditionally independent, as in pulmonary embolus (PE) causes cough and syncope (the 2 symptoms result from distinct pathophysiologic pathways); or they may be conditionally dependent, as in pulmonary embolus causes cyanosis and syncope (both result from a common pathophysiologic pathway of a PE subset, massive embolism causing circulatory obstruction). In **Figure 2B** two marginally independent disorders cause a single finding, e.g., pneumonia and acute bronchitis both cause cough. In **Figure 2C**, two causally related disorders each cause the same finding, e.g., chronic hepatitis causes cirrhosis and both disorders cause hyperbilirubinemia and jaundice; or acute bronchitis precipitates a COPD flare and both cause cough. In **Figure 2D**, two causally related disorders each explain a distinct subset of the patient findings, e.g., deep vein thrombosis causes pulmonary embolus, with patient findings leg edema (caused by DVT) and dyspnea (caused by PE).

BNs have been a popular choice for medical diagnosis because of their ability to model complex domains and to provide a sound basis for their inference. Compared to pure ML solutions, BNs can incorporate derived medical knowledge (e.g., published studies, textbooks, expert opinion), and do not require huge raw datasets. Fundamental problems with traditional Bayesian implementations include:

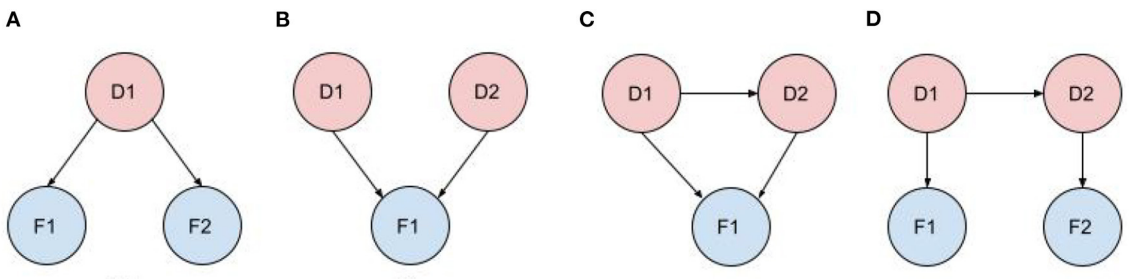

FIGURE 2 | Typical diagnostic BN configurations. **(A)** A disorder causes 2 findings; **(B)** Independent disorders both cause a finding; **(C)** Causally related disorders cause the same finding; **(D)** Causally related disorders each explain a subset of the patient findings.

- Severe scalability problems due to the large number of nodes required for a diagnostic network with a large number of diagnoses and/or findings (Cheng and Druzdzel, 2000; Heckerman, 2013). A general medical diagnosis BN (e.g., for primary care) may have thousands of diagnoses and tens of thousands of findings. The richer the model, the larger and more complex the DAG becomes, and the more data is required to populate the CPTs. Furthermore, high accuracy requires that many of the findings be modeled as continuous or categorical random variables which can make the CPTs very large.
- Inability to model large-scale knowledge representations (Koller and Pfeffer, 1997). The BN DAG represents a single semantic dimension (causality), but other relationships are required to represent the diagnostic process. Of specific interest in diagnosis is the ability to model inheritance hierarchies. For example, to diagnose "brain tumor or neoplasm" or one of its many subtypes, a conventional BN would require the parent disorder and all of its descendants to each independently be represented in the DAG. This presents not only complexity issues but also defies basic diagnostic heuristics, e.g., that "brain tumor" shouldn't "compete" in the differential diagnosis with its child, "dominant temporal lobe tumor".
- Failure to capture the semantic overlap or partial synonymy among findings. Semantic overlap is an inevitable byproduct of a complex ontology. When semantic overlap occurs, findings cannot be considered independent, and they jointly fail to deliver the same diagnostic power that is implied by the assumption of independence. For example, if a chest x-ray shows left atrial enlargement (LAE), then an echocardiogram showing LAE may provide slightly more information since it has a higher specificity, but not as much as if the x-ray had not been discerned. Similarly, if we first discerned that the echocardiogram shows LAE, then the x-ray has little to no additional diagnostic value. The effect of semantic overlap in a system that assumes findings are independent can cause overconfidence or premature closure, leading the system to conclude that a specific disease is the correct diagnosis when in fact there is insufficient evidence for that claim. One approach that has been proposed to partially address this problem is to introduce an intermediate node that represents the collective effect of a set of correlated findings (Yu et al., 1988; Nikovski, 2000; Velikova et al., 2014).
- Failure to capture higher order statistics among finding nodes of a given disorder, e.g., how findings vary with duration of symptoms, age, gender, and other risk factors. For example, gender *per se* has little effect on the likelihood of psoriatic arthritis (PA), but males with PA are significantly more likely to present with involvement of a single joint.
- Failure to capture causal relationships among disorder nodes (Richens et al., 2020). The assumption that a patient's findings must be explained by a single disorder rather than the simultaneous occurrence of multiple causally linked disorders can cause underconfidence (diffidence), leading the system to fail to rank the correct diagnosis or diagnoses as the top disorder(s) even after sufficient information was presented for that claim. For an in-depth discussion of diffidence and over-confidence detection in diagnostic systems, see (Hilden et al., 1978).

MATERIALS AND METHODS

This paper describes the MidasMed DA, a prototype system based on a novel BN with improved diagnostic modeling. A comprehensive description of the diagnostic engine that powers the MidasMed DA is outside of the scope of this paper. However, we provide highlights of the solution architecture and key innovations that address the fundamental limitations of traditional implementations listed above, and advance the state-of-the-art in AI diagnosis.

The solution architecture consists of the following key components:

- A rich semantic model that captures entity data and relationships among entities of the medical ontology that is largely independent of implementation constraints. The semantic model is instantiated as an object-oriented model for efficient diagnostic computations.
- A diagnostic engine that for each diagnostic request dynamically generates a sparse BN, and then applies a Bayesian

classifier to generate a differential diagnosis. The classifier implements disorder subtype hierarchies to recursively and efficiently generate a differential diagnosis with the maximum disorder specificity supported by the data. For example, if warranted by the data, the system will report "anteroseptal acute myocardial infarction" instead of the less specific "acute myocardial infarction." Note that for many disorders, optimum treatment depends on knowing the specific subtype.
- A "Best Next Finding" module that generates a set of additional findings to discern (from the patient or clinician) in order to most quickly and economically refine the diagnosis.

The semantic model describes the medical ontology and the relationships among its concepts using statistical, logical, and heuristic data. The model can be edited and viewed using a web-based content management system (CMS), and is stored in a semantic SQL database. A constructor algorithm generates an object-oriented model from the semantic assertions in the database, resulting in a Data Transfer Object (DTO). The DTO may be serialized for storage and transport to the server running the diagnostic engine. The DTO represents an in-memory object-oriented image of the semantic model that enables rapid and efficient diagnostic computation in real-time. The DTO abstractly represents the global BN, although other (more efficient) data structures are used to hold the node objects. Each node encapsulates all the information it needs to discover its graph neighbors *via* pointers to other nodes.

Our diagnostic model focuses primarily on the following aspects: (1) dependencies among disorders, (2) subtype relations within a disorder family, (3) the characterization of each disorder in terms of its relevant findings and risk factors, (4) statistical correlation and semantic overlap among findings, and (5) finding contingency hierarchies stemming from the relative semantic scope of each finding and the linear progression of the diagnostic interview. Each of these topics in described in the following sections.

Inter-disorder Dependency

Disorder dependency is important to model because a patient may present with symptoms of both a causal disease and its complication(s). For example, a patient might present with deep venous thrombosis (DVT) in a leg, combined with symptoms of pulmonary embolus, a life-threatening complication of DVT. In cases where the initial cause is insidious or insufficiently bothersome, or when the cause and its complication(s) occur in rapid succession, the causal disorder may not have been previously diagnosed. We do not want the classifier to "punish" a disorder for not explaining findings of its co-presenting dependent disorder(s); rather, such combinations of findings often provide high confidence for the diagnosis of a *combination* of causally linked disorders. Therefore, our classifier is designed to identify single disorders or clusters of dependent disorders that best explain the patient findings. Of course two *independent* disorders may also jointly explain the patient findings; however, the probability of such an event is generally much lower.

We used the term *Multi-Disease Model* (MDM) to describe a classifier that detects and accounts for clusters of dependent disorders in the differential diagnosis. One of the consequences of MDM is that co-occurring dependent disorders may each explain some of the same finding(s). We therefor need a mechanism for describing how the joint interaction among disorders affects the presentation of their common findings. We use the term *equivalent sensitivity* to describe the sensitivity of a finding that is relevant to multiple dependent disorders that are all assumed to be present (with appropriate extensions for categorical and continuous findings). To illustrate this case, suppose D_1 causes D_2, and both share common a finding F_1 with sensitivities $s_{1,1} = P(F_1|D_1)$ and $s_{1,2} = P(F_1|D_2)$. The cluster consisting of D_1 and D_2 has 3 configuration: $\{D_1^+, D_2^-\}$, $\{D_1^-, D_2^+\}$, and $\{D_1^+, D_2^+\}$, where the +/– indicate whether the disorder is present or absent. When both disorders are present, F_1 will have an equivalent sensitivity for the configuration that depends on (a) the nature of F_1, (b) the sensitivities $s_{1,1}$ and $s_{1,2}$, and (c) whether or not F_1 arises in D_1 and D_2 due to shared or distinct pathophysiological mechanisms. For example, if F_1 is body temperature, D_1 causes hypothermia and D_2 causes fever (an admittedly unusual case), then we would expect the patient temperature (given that she has both D_1 and D_2) to be $s_{1,1} < \bar{s}_1 < s_{1,2}$. On the other hand, if D_1 and D_2 both cause fever, and due to the same underlying mechanism, then we expect $\bar{s}_1 \approx \max(s_{1,1}, s_{1,2})$. But if D_1 and D_2 both cause fever due to different mechanisms, we might expect $\bar{s}_1 > \max(s_{1,1}, s_{1,2})$. Now suppose F_1 is *time to diagnosis*, with the corresponding question "How long ago did your symptom(s) begin?". If D_1 has a gradual onset with a distribution centered on "months to years", while D_2 has a shorter onset, say "days to weeks" then the equivalent sensitivity will satisfy $\bar{s}_1 \approx \max(s_{1,1}, s_{1,2})$, because the patient will most likely associate the beginning of the problem with the onset of D_1, which started first.

To formally describe MDM, consider a cluster of dependent disorders. To qualify, each cluster member must have at least one link to another cluster member, and must explain at least one abnormal patient finding. A disorder may belong to at most one cluster, for if it belonged to multiple clusters those would be merged into a single cluster. A disorder with no dependencies is called a *singleton* (cluster of size 1). Let $D_1, \ldots, D_N$ be members of cluster C, and $F_1, \ldots, F_M$ be the known patient findings. The *configurations* of C are all permutations of the cluster disorders in which some are present and others are absent. For the net probability of C (all configurations) we have:

$$P\left(C|f_1,\ldots,f_M\right) = \sum_j P\left(C_j^+ \middle| f_1,\ldots,f_M\right) = \sum_j I(C_j^+) \cdot \prod_{i=1}^{M} \bar{s}_{ij} \tag{6}$$

where $I(C_j^+)$ is the joint incidence (prior probability) of the disorders in C_j^+ co-occurring, and $\bar{s}_{ij}$ is the equivalent sensitivity for finding F_i in C_j^+. The probability of cluster disorder D_k is the sum of the probabilities of all configurations in which it is present, i.e.,

$$P\left(D_k|f_1,\ldots,f_M\right) = \sum_j P\left(C_j^+ \middle| f_1,\ldots,f_M\right) \cdot \delta_{jk} \tag{7}$$

where $\delta_{jk} = 1$ *if* $D_k \in C_j^+$ and 0 otherwise. While the total number of configurations may be very large (since C may be large) this does not present a computational problem, since the vast majority of configurations can be discarded using pruning heuristics with negligible effect on the accuracy of the cluster probability computation. Note that given the set of all contending diagnoses across all clusters, the cluster probabilities sum up to 1.0 but the disorder probabilities do not, due to co-occurrence among the disorders.

Disorder Subtype Hierarchies

The ability to model disorder subtypes is important in diagnosis, because disorder subtypes may have different prognoses and/or require different treatments (e.g., viral vs. bacterial meningitis). We use the term *subtypy* to define a framework for describing the disorder inheritance hierarchy. Note that inheritance hierarchies in diagnosis are statistical and not directly analogous to the programming concept of object-oriented inheritance. In diagnosis, the ancestor represents a statistical aggregate of its descendants or variants, and while it may be convenient to think of a subset of findings as manifest in the parent and passed on to the children, there are usually variations in how these findings are expressed (or not) in each child. For example, conjunctival injection is always present in infectious conjunctivitis, and inherited to both subtypes gonococcal (bacterial) conjunctivitis and viral conjunctivitis. However, conjunctival hemorrhages are more common in the viral variant, while eyelid edema and purulent discharge are more common the bacterial variant. Furthermore, a Gram stain of the gonococcal conjunctivitis discharge may identify Gram-negative diplococci, but it is irrelevant to the viral variant. So the Gram stain test finding is relevant to the parent (infectious conjunctivitis), but not to its viral child. In summary, a child attribute is always represented by the parent, but not necessarily vice versa, and the manifestation in the parent is a statistical aggregate of its children.

Because each parent represents the statistical aggregate of its children, and the probability of each child varies based on the patient findings, we must compute all sensitivities dynamically for each new set of patient findings, and we must do so by starting at the very bottom of the hierarchy tree (the "leaves" or childless disorders). To see why this is the case, consider a simple example with parent disorder meningitis and its children viral and bacterial meningitis. The prior probability (incidence) of meningitis in the U.S. is ∼9.25e-5. Approximately 82% of cases are viral and 18% bacterial. Consider the finding "CSF culture positive for bacteria." This finding is relevant to bacterial meningitis with $s_{bm} \approx 0.95$ and is not relevant to viral meningitis, so we assign a noise sensitivity, e.g., $s_{vm} = 0.02$, and compute the sensitivity in the parent as the weighted sum: $s_m = (I_{vm} \cdot s_{vm} + I_{bm} \cdot s_{bm})/I_m = 0.82 \cdot 0.02 + 0.18 \cdot 0.95 = 0.187$. Now suppose this finding was determined to be positive in the patient. The posterior relative probability of the children is now $P_{vm} = I_{vm} \cdot s_{vm} = 0.82 \cdot 9.25\text{e-}5 \cdot 0.02 = 1.152\text{e-}6$ and $P_{bm} = I_{bm} \cdot s_{bm} = 0.18 \cdot 9.25\text{e-}5 \cdot 0.95 = 1.58\text{e-}5$. The relative probability of the children has changed from 0.82/0.18 to 0.07/0.93, and $s_m = 0.07 \cdot 0.02 + 0.93 \cdot 0.95 = 0.88$. Similarly, if the finding was negative in the patient then $P_{vm} = I_{vm} \cdot (1 - s_{vm}) = 0.82 \cdot 9.25\text{e-}5 \cdot 0.98 = 7.43\text{e-}5$, $P_{bm} = I_{bm} \cdot (1 - s_{bm}) = 0.18 \cdot 9.25\text{e-}5 \cdot 0.05 = 8.32\text{e-}7$, the relative probability ratio is 0.99/0.01 and $s_m = 0.99 \cdot 0.02 + 0.01 \cdot 0.95 = 0.03$.

From the end user perspective it is desirable for the diagnostic process to proceed from the general to the specific (e.g., from "stoke or TIA" to "cortical posterior cerebral artery stroke, dominant") progressively as more of the relevant patient findings are discerned. To do so, we use a heuristic called *Child Better than Next* that replaces a parent disorder by all its direct children provided that the relative probability of at least one of the children exceeds that of the next disorder in the differential diagnosis stack. This requires the disorders to be ranked by descending relative probability, and for the stack to be resorted after each replacement.

Disorder Findings Dependencies

Each finding is modeled as binary, discrete multi-valued (categorical), or a continuous random variable. We use the term "finding" broadly to include risk factors, and distinguish between them by selecting the appropriate interaction model (e.g., reflecting direction of causality) when computing their impact on disorder probabilities.

While some findings may justifiably be modeled as conditionally independent for a given disorder (Naïve Bayes), this is not the case in general. Frequently, findings vary with other findings that are not directly relevant to the index disorder. In such cases we can write:

$$P\left(f_1|D\right) = P\left(f_1|f_2, \ldots, f_L, D\right) \tag{8}$$

where F_1 is relevant to D and $P\left(f_1|D\right)$ can be described by a multidimensional probability distribution, with *factor findings* $F_2, \ldots, F_L$ that are not necessarily directly relevant to D, but act as factors in the computation of its finding probabilities. Common factor findings are age, gender, and time-to-diagnosis; however, many findings have their unique factor findings. For example, **Figure 3** depicts the distribution of serum glucose for diabetic ketoacidosis (DKA) as a function of factor findings "current pregnancy" and "recent heavy alcohol consumption".

Inter-findings Dependencies

Failure to capture semantic overlap or disjunction can cause significant distortion unless inter-finding dependencies are properly managed. At the root of the problem is the basic concept of finding *diagnostic power*. The diagnostic power of a finding represents how much information it contributes to the likelihood of a disorder relative to contending disorders. That is, given what we already know about the likelihood of a disorder from its prior probability (incidence) and previously ascertained findings, how much *additional* information does a new finding provide? We define diagnostic power using a measure called the *probability factor* (PF), which is the ratio of the probability of the finding in the disorder relative to its prevalence in the general population. **Table 4** in Supplementary Materials shows how this measure relates to other popular measures that quantify the discriminating power of a finding.

To illustrate the problem of semantic overlap, consider a patient complaining of pain, edema (swelling), and erythema

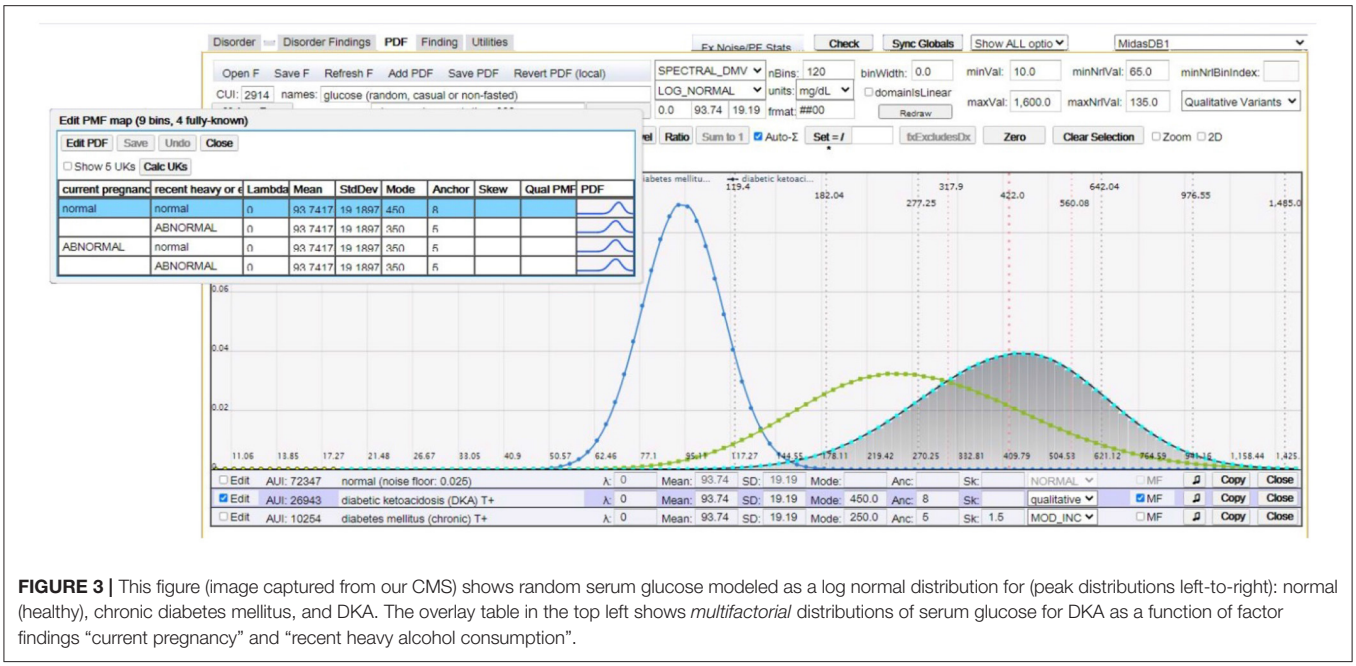

FIGURE 3 | This figure (image captured from our CMS) shows random serum glucose modeled as a log normal distribution for (peak distributions left-to-right): normal (healthy), chronic diabetes mellitus, and DKA. The overlay table in the top left shows *multifactorial* distributions of serum glucose for DKA as a function of factor findings "current pregnancy" and "recent heavy alcohol consumption".

(redness) at the knee. These findings collectively represent aspects of knee joint inflammation in rheumatoid, traumatic, or reactive arthritis. Note however that these findings are not correlated or even jointly relevant for all disorders that cause knee pain. For example, L4 lumbar disc herniation can cause knee pain, but not edema or erythema.

We address semantic overlap by defining an intermediate node called an *xopathy* (a generalization of terms such as neuropathy, dermopathy, or arthropathy). The xopathy framework enables us to represent a set of findings that are conditionally dependent with respect to an index disorder using an interim aggregate node. The xopathy sensitivity represents the incidence of the xopathy in the population of patients with the disorder. The xopathy sensitivity can also be interpreted as the conditional probability that one or more of the xopathy findings is present given the index disorder.

Let D represent a disorder with conditionally *dependent* findings $F_1, \ldots, F_L$. We construct an xopathy Xop with the findings as its members, and each having a sensitivity $s_i = P(f_i|xop)$. We are also given the xopathy sensitivity, $s_{Xop} = P(xop|D)$. Our goal is to compute dynamic sensitivities $s_1^*, \ldots, s_K^*, K \leq L$ for each *known* finding that satisfy

$$P\left(f_1, \ldots, f_K|D\right) = \prod_{i=1}^{K} s_i^* \tag{9}$$

The actual algorithms for computing $\{s_i^*\}$ are beyond the scope of this paper. However, we provide a brief outline of the process with key equations.

Step 1: Compute the *independent* xopathy diagnostic power (probability factor), PF_{indep}, as the product of the finding PFs. This represents the diagnostic power we would introduce into the disorder probability computation if we assumed the findings were independent. As noted earlier, PF_{indep} will generally be greater than the *desired* diagnostic power when the findings are correlated.

$$PF_{indep}\left(Xop\right) = \prod_{i=1}^{K} PF\left(f_i|Xop\right) \tag{10}$$

where $\left(f_i|Xop\right) = s_i/n_i$, n_i is the prevalence of F_i in the general population, and s_i is the finding sensitivity relative to the xopathy. Note that the findings are independent relative to the xopathy (but not the disorder), which allows us to use the Naïve Bayes assumption in Equation (10).

Step 2: Determine the maximum allowed PF for this xopathy, $PF_{\max}\left(Xop\right)$. If PF_{indep} exceeds $PF_{\max}$ then apply compression to decrease finding sensitivities. We denote the compressed sensitivities $\{\ddot{s}_i\}$. The compression algorithm must satisfy several constraints, such as preserving the relative magnitude of the original sensitivities ($s_i > s_j \rightarrow \ddot{s}_i > \ddot{s}_j$), and ensuring that positive findings remain so $\left(\frac{s_i}{n_i} > 1 \rightarrow \frac{\ddot{s}_i}{n_i} > 1\right)$.

Step 3: Reflect the xopathy sensitivities to the disorder. The sensitivities $\{\ddot{s}_i\}$ represent the conditional probability of the findings on the xopathy, but what we really want is sensitivities conditioned on the disorder per Equation (9). Let $x_0 = s_{Xop} = P(xop|D)$, $\hat{s} = \left(\prod_{i=1}^{K} \ddot{s}_i\right)^{\frac{1}{K}}$, and $\hat{n} = \left(\prod_{i=1}^{K} n_i\right)^{\frac{1}{K}}$, where $\hat{s}$ and $\hat{n}$ represent the geometric means of $\{\ddot{s}_i\}$ and $\{n_i\}$, respectively. For simplicity, in this derivation we're interpreting s_i as the probability of the finding F_i in its *known* state. If the finding is negative then $s_i = 1 - P(F_i \; is \; positive)$.

We initialize the algorithm as follows:

$$\begin{cases} \tilde{s}_1 = x_0 \cdot \hat{s} + (1 - x_0) \cdot \hat{n} \\ x_1 = x_0 \cdot \frac{\hat{s}}{\tilde{s}_1} \end{cases} \tag{11}$$

Note that $\tilde{s}_1$ is the expected sensitivity over the two mutually exclusive disorder subpopulations: the xopathy population with prior probability x_0, and the complementary population with prior probability $(1 - x_0)$. With each discerned finding, the probability that the patient belongs to the xopathy subpopulation changes. If the finding was positive the xopathy probability increases and if it was negative it decreases.

Similarly, for the remaining iterations, $j = 2, .., K$ we have:

$$\begin{cases} \tilde{s}_j = x_{j-1} \cdot \hat{s} + (1 - x_{j-1}) \cdot \hat{n} \\ x_j = x_{j-1} \cdot \frac{\hat{s}}{\tilde{s}_j} \end{cases} \tag{12}$$

Similar to $\hat{s}$, we define $\check{s} = \left(\prod_{i=1}^{K} \tilde{S}_i\right)^{\frac{1}{K}}$ as the geometric mean of the raw disorder sensitivities computed in Equation (12). Finally, we normalize the $\{\tilde{s}_j\}$ using the scaling factor $R = \frac{\check{s}}{\hat{s}}$ in order to preserve the xopathy diagnostic power achieved in Step 2. The final sensitivities $\{s_i^*\}$ for Equation (9) are:

$$\begin{cases} s_i^* = R \cdot \ddot{s}_i \; for \; R \leq 1.0 \\ s_i^* = \frac{R \cdot \ddot{s}_i}{1 + \ddot{s}_i(R-1)} \; for \; R \; > 1 \end{cases} \tag{13}$$

The second form of s_i^* in Equation (13) uses the function $f(x) = \frac{R \cdot x}{1 + x(R-1)}$ to guarantee that the sensitivity never exceeds 1.0. While previous work has described the use of intermediate nodes to express the aggregate sensitivity of correlated findings (Yu et al., 1988; Nikovski, 2000; Velikova et al., 2014), we are unaware of other successful attempts to express the diagnostic power and sensitivity of the intermediate node as independent finding sensitivities for the disorder per Equation (9). This process is critical to avoid semantic disjunction in MDM computations. To see why this is the case, consider dependent disorders D_1 and D_2. Suppose findings F_1 and F_2 are relevant to both disorders, but are only conditionally *dependent* with respect to D_1. If we were to replace F_1 and F_2 by an xopathy node $Xop(F_1, F_2)$ as a finding of D_1, then the disorder cluster $\{D_1, D_2\}$ would have 3 findings instead of 2, thus creating semantic disjunction and rendering the equivalent sensitivities incorrect.

Finding Contingency Hierarchies

The finding contingency hierarchy represents a formalization of the "drill-down" conventions of the medical interview. The top finding (e.g., "chest pain") is usually followed by more specific findings like *quality or character* of the pain (e.g., sharp, dull, stabbing, burning, pressing), exacerbating factors (e.g., cough or exercise), relieving factors (e.g., drinking water or sitting up), etc. For many "top level" findings like chest pain or skin rash there may be tens of additional secondary or contingent findings that need to be discerned to obtain a clear picture of the disease state.

We say that finding F_c is *contingent* on F_p (and F_p is a *prerequisite* of F_c) if F_c has no meaning unless F_p has been discerned. Usually, F_c won't have any meaning unless F_c takes on specific state(s). For binary findings, this condition is always that the prerequisite finding must be positive. For example, we can't ask about chest pain quality if the patient has denied chest pain. Note that a prerequisite finding may have multiple contingents, and that a contingent finding may also have multiple prerequisites. Furthermore, contingencies may be chained or nested to multiple levels.

In some cases the contingency chain must be queried in a specific order to create a coherent interview that makes sense to the patient. For example, if the patient complains of a skin lesion, we cannot ask "How deep is the ulcer?" unless we first determine that the lesion is, indeed, an ulcer. Similarly, if the patient complains of abdominal pain, there is no point asking "Is the pain relieved by antacids?" (suggests a peptic ulcer) unless we first discern that the pain is located in the upper abdomen. Similarly, we cannot ask "Which came first, the abdominal pain or the nausea & vomiting?" until we have discerned that both findings were reported.

Finding contingency chains present an interesting dilemma, namely, what probability to assign to contingent findings whose prerequisites are irrelevant to an index disorder. To illustrate this scenario, suppose the patient presents with 2 positive findings, F_1 and F_2 and that there are 3 contending disorders, D_1, D_2, and D_3. Suppose F_1 is relevant to all 3 disorders and F_2 is relevant only to D_1 and D_2. For simplicity assume all disorders have the same incidence, all findings have a sensitivity of 0.3 to all relevant disorders, and that all findings have a noise sensitivity of 0.02. The relative probabilities of the disorders at this point are $P(D_1)/P(D_2)/P(D_3) = 0.3^2/0.3^2/0.3 \cdot 0.02$. The relative probability of D_3 has decreased by approximately an order of magnitude. Now suppose F_2 has contingent finding F_{21} that is positive in the patient, and only relevant to D_1. The updated relative probabilities are $P(D_1)/P(D_2)/P(D_3) = 0.3^3/0.3^2 \cdot 0.02/0.3 \cdot 0.02^2$. The decrease of $P(D_2)$ relative to $P(D_1)$ seems justified, because given F_2, D_1 matches the finding pattern better than D_2. However, D_3 has essentially been punished twice for not explaining the prerequisite finding. Each time we query another finding in the F_2 contingency chain the relative probability of D_3 will decrease by the probability factor 0.3/0.02, and very quickly D_3 will be discarded from consideration. We use the term *"don't care" finding* to mean a positive contingent finding for a prerequisite that is irrelevant to the index disorder. In our example, F_{21} is a "don't care" condition for D_3. We further stipulate that the relative probability of a disorder should be minimally impacted by its "don't care" findings. The solution we implemented was to derive a weak positive sensitivity to "don't care" findings.

The MidasMed Diagnostic Engine and Web App

The diagnostic engine is implemented as a web server that receives stateless diagnostic requests from a client, and returns a response consisting of a probability ranked differential diagnosis and a ranked list of the best next findings to discern. The first step is to generate a list of all valid diagnoses that explain at least one abnormal patient finding. The disorder list is used to create a dynamic sparse BN. It is sparse, because it contains only valid diagnoses for the given request. As described earlier, the conditional probabilities for each parent disorder are represented as statistical aggregates of the children. Note that there is no need to compute the entire finding conditional probability

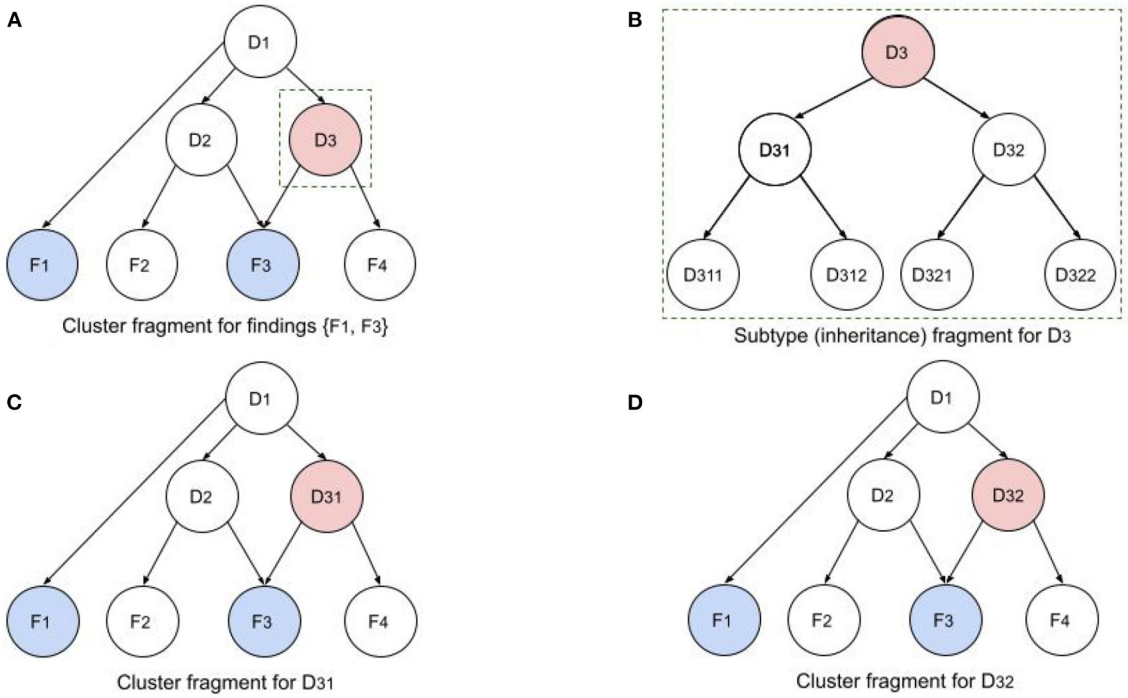

FIGURE 4 | Illustration of recursive BN computations for disorder cluster and subtype fragments. **(A)** Cluster fragment for patient findings F_1 and F_3 and disorder subtype ancestors. **(B)** Subtypy tree for disorder D_3. **(C,D)** D_3 in original network has been replaced by its children D_{31} and D_{32} to compute the cluster probabilities with the two children.

distribution, only the probability of the patient value. A recursive computation is then initialized with the ancestor disorders of each subtype family. MDM computations are applied, and the disorders are placed in a stack and ranked by descending relative probability. The Child Better than Next heuristic is then applied recursively (starting at the top of the disorder stack), by replacing the next qualified parent and all its siblings by all their children, updating the relative probabilities, and resorting the stack. Note that only the MDM cluster containing the parent(s) needs to be recomputed with each replacement. The resulting final differential diagnosis offers the user the appropriate diagnostic subtype specificity for the known findings.

Figure 4 illustrates a fragment of a single iteration in this recursive process. **Figure 4A** shows a cluster fragment for patient findings F_1 and F_3. Note that F_3 is relevant to both D_2 and D_3, so it will require an equivalent sensitivity for configurations in which both disorders are present. In the next iteration (if the Child Better than Next criterion is satisfied) D_3 will be replaced by children D_{31} and D_{32}. In the following iteration D_{31} and D_{32} (siblings) will be replaced by all their children (D_{311}, D_{312}, D_{321}, and D_{322}). Note that the network in **Figure 4A** depicts causality (e.g., D_1 causes D_2 and D_3), while the network in **Figure 4B** depicts disorder subtypes (e.g., D_3 is a supertype of D_{31} and D_{32}). Subtypes of a single parent (siblings) are considered mutually exclusive, so $P(D_{31})$ is computed using the configurations of the cluster in **Figure 4C**. However, the probabilities of the dependent disorders (D_1 and D_2) are computed from the configurations of both **Figures 4C,D**, by summing the probabilities of all configurations in which they appear. Similarly, in the next recursion, configurations will be computed with D_{311}, D_{312}, D_{321}, and D_{322}.

The innovations described above combine to produce a nuanced approach to diagnosis that we assert results in substantially greater accuracy than existing solutions in that the differential diagnosis probabilities are more consistent with the evidence available to support them. We further assert that with diagnostic guidance based on Bayesian probabilities, heuristics, and estimated costs, the differential diagnosis converges to the correct diagnosis more efficiently, potentially translating into time and cost savings.

Our prototype system (MidasMed) currently recognizes a limited subset of 200 common adult primary care disorder subtype families (760 total diagnoses) spanning a variety of systems (respiratory, dermatology, neurology, musculoskeletal, etc.), and 4,000 findings (We estimate these encompass approximately half of the disorders a competent primary care physician should be able to recognize.). The semantic network is defined using statistical and logical analysis of epidemiological data, case series, journal articles, textbooks, and other online resources.

MidasMed includes a user-friendly web app for both patients and clinicians using dual vocabularies and default application settings for the two distinct user groups. For example, by default patients and lay caregivers are presented only with

history questions in lay terminology, while professional users are asked all finding types (including exam and test results) using professional terminology. The user interface is interactive, and is designed to give the user maximum flexibility and control. Throughout the encounter, patient findings can be augmented, edited or deleted. The user can choose from 3 ways of entering new findings to refine the initial differential:

1. Search: The user selects her own findings from a global findings list.
2. Guide Me: MidasMed asks a short series of the best next questions to discern.
3. Drill Down: The user selects a disorder from the differential diagnosis to view, rank, and select undiscerned findings for that disorder. This allows the user to focus on a condition of particular concern due to urgency or severity, and answer the questions that will most efficiently rule it in or out.

Experimental Paradigm

In this research we compare the performance of MidasMed to that of physicians and six other publicly accessible online diagnostic aids: Ada, Babylon, Buoy, Isabel, Symptomate, and WebMD. To facilitate a comparison with previous studies, we used a set of publicly available case vignettes (Semigran et al., 2015) that were tested on 23 symptom checkers in 2015, physicians (Semigran et al., 2016) and on three physicians and the Babylon DA in 2020 (Baker et al., 2020). The vignettes are available online in the format of **Table 1**. (See **Table 5** in the Supplementary Materials for a complete list of vignettes and also a link to the vignettes file).

As in the previous studies (Semigran et al., 2015; Baker et al., 2020), we used only the information from the "Simplified (*added symptoms*)" column of the vignette file, and excluded vignettes based on conditions on which MidasMed had not yet been educated (in conformance with the methodology of Baker et al., 2020). This resulted in a test set of 30 vignettes, the same number used in Baker et al. (2020). We note that none of the vignettes had been used in the training, education or parameterization of MidasMed.

We regarded the diagnosis presented in the "Diagnosis" column of the vignette file as the true or "target" diagnosis, except in 2 cases where no final diagnosis was provided but was clearly implied (the implied diagnosis was used), and 2 cases where multiple causally linked disorders were implied by the vignette history (either implied diagnosis was accepted). We did not find descriptions of how these problematic vignettes were treated in the previous articles. These exceptional cases are clearly identified in **Table 5** in the Supplementary Materials.

In two cases the diagnosis provided for the vignette seemed inadequately substantiated by the simplified vignette history in our clinical opinion. For presumed consistency with the previously reported research, we nonetheless regarded it as the target diagnosis. These cases are also identified in the Supplementary Materials (**Table 5**).

MidasMed is an incomplete prototype, and therefore has not been publicized or promoted, but is publicly accessible (for a limited time) for evaluation and feedback at midasmed.com, and the vignette cases created for this article are publicly accessible *via* the application for anyone to view and experiment with (see instructions in the **Supplementary Materials**). At this writing, MidasMed recognizes only 200 adult disorder families. A complete list of supported diagnoses can be found in the app at midasmed.com from the Options (hamburger) menu.

For this study we used all of the adult vignette cases from the source file on which MidasMed has been educated, plus three pediatric cases for which the presentation is very similar to that in adults. Since MidasMed only accepts patient ages ≥ 18, the ages of the three pediatric patients were transposed to 18 years.

All the other DAs evaluated are publicly promoted as diagnostic aids for the general public. (One limits the age to ≥ 16, for which the age of the two younger patients was also transposed to that minimum age). Since none of the vignettes are based on rare disorders, we assumed the other DAs to be capable of recognizing all the target diagnoses.

The data for physicians and the Babylon DA were taken from Baker et al. (2020), and were not independently replicated in this study. For each of the other diagnostic assistants one of us (D. Jones, MD, board certified in emergency medicine with 25 years' primary care experience) entered only the "Simplified (*added symptoms*)" findings for each vignette into the online DAs (See the **Supplementary Materials** for links to all the DAs). Note that these simplified vignettes were designed to reflect only the history findings and observations that a patient could enter. For each DA we recorded (a) the fraction of cases for which the target diagnosis was #1 in the list of diagnoses provided; and (b) the fraction for which the target diagnosis was in the top 3 disorders of the list.

RESULTS

The results of our research are presented in **Table 2**.

Limitations Regarding Our Results

Although MidasMed aspires to be a complete diagnostic aid for both patients and clinicians, and therefore includes the physical examination and test findings required to definitively diagnose the disorders on which it has been educated, only history findings were entered in this study. The objective here was to quantify the ability to identify the correct diagnosis based on sparse patient histories, as are readily available directly from patients online.

With only 30 cases, the statistical reliability of the results is low, as reflected in the broad confidence intervals. The original study for which the vignettes were created (Semigran et al., 2015) included 45 vignettes, but only the 27 adult plus 3 pediatric disorders on which MidasMed has been educated were tested in this study, and only 30 in the study (Baker et al., 2020) that produced the physician and Babylon data reported here.

It is possible that as the breadth of disorders covered by MidasMed is increased, and the correct diagnosis must compete with a greater number of similar disorders, accuracy will decline. However, since (a) the disorders presently covered by MidasMed were selected because they are among the most common, and (b) the vignette

TABLE 1 | Sample vignette.

Diagnosis	Vignette	Simplified (*added symptoms*)
Requires emergent care (*n* = 15)		
Appendicitis	A 12-year-old girl presents with sudden-onset severe generalized abdominal pain associated with nausea, vomiting, and diarrhea. On exam she appears ill and has a temperature of 104°F (40°C). Her abdomen is tense with generalized abdominal pain, nausea, tenderness and guarding. No bowel sounds are present.	12 y/o f, sudden onset severe abdominal pain, nausea, vomiting, diarrhea, T = 104

TABLE 2 | Performance comparison summary results for 7 DAs and physicians.

Physician or DA	Vignettes tested	Target diagnosis ranked #1			Target diagnosis in the top 3		
		Fraction	Percent (%)	95% CI[e]	Fraction	Percent (%)	95% CI[e]
Physicians[a]	90[b]	68/90[b]	75.3	65.4–84.0	81/90[b]	90.3	81.9–95.3
Ada	30	22/30	73.3	54.1–87.7	27/30	90.0	73.5–97.9
Babylon[a]	30	21/30	70.0	50.6–85.3	29/30	96.7	82.8–99.9
Buoy	21[c]	11/21	52.4	29.8–74.3	15/21	71.4	47.8–88.7
Isabel[d]	30	15/30	50.0	31.3–68.7	21/30	70.0	50.6–85.3
MidasMed	30	28/30	93.3	77.9–99.2	29/30	96.7	82.8–99.9
Symptomate[d]	30	21/30	70.0	50.6–85.3	26/30	86.7	69.3–96.2
WebMD[d]	30	20/30	66.7	47.2–82.7	28/30	93.3	77.9–99.2
All DAs	201	138/201	67.7	61.8–75.0	175/201	87.1	81.6–91.4
Top 3 DAs[f]	90	71/90	78.9	69.0–86.8	82/90	91.1	83.2–96.1

[a] *The Babylon and physician tests were not replicated in this study, but were transcribed from Baker et al. (2020), which used the same methodology.*
[b] *In the Babylon study three physicians were tested, but only percent data were reported; therefore 95% CI's were computed assuming a total of 90 vignettes (30 per doctor).*
[c] *For 9 of the 30 disorders presented, Buoy gave no proposed diagnoses; only triage recommendations (e.g., "Contact a medical professional" or "Call 911!").*
[d] *Isabel, Symptomate, and WebMD are the only DAs tested both in the original paper (Semigran et al., 2015) and this study.*
[e] *CI intervals were computed using Clopper-Pearson exact method for binomial probability distributions.*
[f] *For a larger sample size to compare with physicians, we combined the top 3 DAs we tested (Ada, MidasMed, and Symptomate).*

diagnoses are mostly common disorders, adding the less common disorders is unlikely to hinder the recognition of the vignette disorders. Rather, it will be difficult (probably impossible) to correctly identify an uncommon disorder (e.g., bronchiectasis or idiopathic pulmonary fibrosis) as the most likely diagnosis based on only sparse vignette histories such as were used here, some of which contain only 3 or 4 common findings.

It was difficult to make perfectly fair comparisons of the different DAs due to differences in their user interface (UI) approaches. For example, some apps (e.g., MidasMed, Ada) offer an "unknown" option for (virtually) every follow-up question queried, making it easy to limit the information entered strictly to the items provided in the simplified vignettes. However, other DAs (e.g., Buoy, Symptomate), presented follow-up questions that required an affirmative or negative answer to proceed. In those cases (i.e., when forced to provide information not in the vignette), we attempted to err in the direction of aiding the DA under test, by answering as a typical patient with the target disorder would most likely answer. In a few cases, it was not possible to enter all history items for a specific vignette because an item was both (a) not accessible in the DAs search facility (despite trying multiple synonyms), and (b) not queried *via* follow-up questions presented by the DA.

DISCUSSION

Canadian physician Sir William Osler (1849–1919), "the father of modern medicine," is known for saying, "Listen to your patient, he is telling you the diagnosis." This message repeats in the medical school maxim, "90% of the diagnosis comes from the history, 9% from your examination, and 1% from tests" (Gruppen et al., 1988; Peterson et al., 1992). This maxim has been forgotten in today's over-stressed healthcare system. Too rushed to take a comprehensive history, doctors often compensate by ordering test panels, referring to specialists, and scheduling more follow-up visits; "Next patient, please." Patients on the receiving end are justifiably frustrated and open to alternatives. But with the growing role of telehealth, where the ability to perform exams or order stat tests is limited, patient history should regain its role as the primary factor in the diagnostic equation. There is also a broader trend toward democratizing access to medical information, or "eHealth" *via* phone apps, wearables, and inexpensive measurement devices, giving patients more control over care options.

In this study we performed a prospective validation of a novel Bayesian diagnostic assistant (MidasMed), and compared it to five online DAs (Ada, Buoy, Isabel, Symptomate, and WebMD) and to the accuracy previously reported for the Babylon DA and physicians. MidasMed was able to identify the correct diagnosis

as most likely with 93% accuracy, significantly outperforming physicians (75%) on the same vignettes (Baker et al., 2020).

We attribute the superior performance of MidasMed to a diagnostic model that moves beyond the "leaky noisy OR gate" assumption of conditional independence among the BN nodes (Henrion, 1987), and to reducing semantic overlap and disjunction that are common in the medical literature and can lead to significant distortion in estimated probabilities of the outcomes. These simple vignettes and our scoring technique did not give MidasMed credit for diagnosing co-present causally related disorders. In particular, it is noteworthy that for the two vignettes that imply the causal co-occurrence of multiple disorders, MidasMed produced estimated relative probabilities for these disorders whose sum approaches 200%, implying a high likelihood of co-occurrence (See the **Supplementary Materials**, for instructions to access the cases online).

It appears from our results that the accuracy of online DAs has improved significantly in the 6-year interim since the original paper (Semigran et al., 2015) evaluated the study vignettes. In that paper, the best-performing symptom checker listed the target diagnosis first only 50% of the time, and in the top three only 67% of the time; and the average performance of 19 symptom checkers in that study for the top 1 and top 3 was only 34 and 51%, respectively. Whereas in this study, the best performance was 93% (top 1) and 97% (top 3); and the average DA performance was 68 and 86%, respectively, showing significant improvement. Furthermore, in this study the performance of the top three DAs combined was 78.9% (top 1) and 91.1% (top 3), comparing very favorably with physicians (75.3 and 90.3%, respectively). Note that in the later comparison we use the 90 vignette aggregates, with similar narrower confidence intervals.

We note several differences in test methodology that may have contributed to the *apparent* accuracy improvements relative to Semigran et al. (2015) for previously tested DAs. First, in Semigran et al. (2015), all data was entered by non-clinicians, who may not have been as facile at matching symptoms to their various DA synonyms as the physician-testers in this study and in Baker et al. (2020). However, that method may give a better estimate of "read world" performance with real patients seeking diagnosis. Second, responses to "mandatory" questions (without which the interview does not proceed, but are not answered by the vignette) may have been entered inadvertently in a way that "punished" the target diagnosis, whereas in this study we explicitly answered such questions to favor the target diagnosis. Third, in Semigran et al. (2015) all 45 vignettes in the source file were used to test all DAs without verifying support for the target diagnosis. These factors may have contributed to the lower scores in the earlier study.

Future Work

At this time MidasMed recognizes a limited set of disorders spanning all organ systems, but lacks comprehensive coverage for any specific system. To complete our technology validation, we plan next to expand its education to *in-depth coverage* of a major organ system (e.g., gastrointestinal and hepatobiliary disorders), and verify that (a) it continues to recognize *most* disorders as the likely diagnosis based on history alone, (b) it recognizes *all* disorders with high accuracy when exam findings and tests are included, and (c) it guides the user efficiently from the initial differential to the definitive diagnosis by optimizing a preset criterion (e.g., diagnostic utility-to-cost ratio). When sufficient data has been acquired, we will apply statistical reliability measures (e.g., Hilden et al., 1978) to assess the confidence and diffidence of the DA's probability estimates.

Although the goal of this paper was limited to the comparison of the diagnostic accuracy of currently available online diagnostic assistants using standardized vignettes, we hope in future work to present our diagnostic innovations in greater detail, and to explicitly measure and compare the accuracy contribution of individual algorithmic innovations (e.g., our modeling of dependencies among findings, modeling of subtypy relationships among disorders, use of continuous probability distributions, etc.).

In this work, to facilitate an apples-to-apples comparison with prior results, we tested on a small set of case vignettes previously tested in Semigran et al. (2015, 2016), Baker et al. (2020). We hope in future work to test across multiple DAs using larger sets of test cases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

DJ contributed to the study design and data analysis, entered the symptoms into the diagnostic assistants, and contributed to the writing of the paper. AJ designed the diagnostic software involved, participated in the study design and data analysis, and contributed to the writing of the paper. Both authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.727486/full#supplementary-material

REFERENCES

Antonucci, A. (2011). "The imprecise noisy-OR gate," in *14th International Conference on Information Fusion* (Chicago, IL).

Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., et al. (2020). A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. *Front. Artif. Intell.* 3, 543405. doi: 10.3389/frai.2020.543405

Barnett, G. O., Famiglietti, K. T., Kim, R. J., Hoffer, E. P., and Feldman, M. J. (1998). "DXplain on the Internet," in *Proceedings of the AMIA Symposium* (Orlando, FL), 607–611.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., et al. (2020). "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI). doi: 10.1145/3313831.3376718

Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., et al. (2019). Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 9, e027743. doi: 10.1136/bmjopen-2018-027743

Cheng, J., and Druzdzel, M. (2000). AIS-BN: an adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *J. Artif. Intell. Res.* 13, 155–188. doi: 10.1613/jair.764

Fagiuoli, E., and Zaffalon, M. (1998). 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artif. Intell.* 106, 77–107. doi: 10.1016/S0004-3702(98)00089-7

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA J. Summits Transl. Sci. Proc.* 2020, 191–200.

Graber, M. L. (2012). The incidence of diagnostic error in medicine. *BMJ Qual. Saf.* 22, ii21–ii27. doi: 10.1136/bmjqs-2012-001615

Gruppen, L., Woolliscroft, J., and Wolf, F. (1988). The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. *Res. Med. Educ.* 27, 242–247.

Heckerman, D. (2013). A tractable inference algorithm for diagnosing multiple diseases. *arXiv preprint arXiv:1304.1511*. doi: 10.1016/b978-0-444-88738-2.50020-8

Henrion, M. (1987). "Practical issues in constructing a Bayes' belief network," in *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence* (Seattle, WA).

Hilden, J., Habbeivia, J. D. F., and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods Inform. Med.* 17, 227–237. doi: 10.1055/s-0038-1636442

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press. Available online at: https://djsaunde.github.io/read/books/pdfs/probabilistic%20graphical%20models.pdf

Koller, D., and Pfeffer, A. (1997). "Object-oriented Bayesian networks," in *UAI'97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (Providence, RI), 302–313.

Lemmer, J., and Gossink, D. (2004). Recursive noisy OR-a rule for estimating complex probabilistic interactions. *IEEE Trans. Syst. Man Cybernet. B* 34, 2252–2261. doi: 10.1109/TSMCB.2004.834424

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* 26, 900–908. doi: 10.1038/s41591-020-0842-3

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Meyer, A., Payne, V., Meeks, D., et al. (2013). Physicians' diagnostic accuracy, confidence, and resource requests. *JAMA Intern. Med.* 173, 1952–1958. doi: 10.1001/jamainternmed.2013.10081

Millenson, M., Baldwin, J., Zipperer, L., and Singh, H. (2018). Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. *Diagnosis* 5, 105–195. doi: 10.1515/dx-2018-0009

Nikovski, D. (2000). Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans. Knowledge Data Eng.* 12, 509–516. doi: 10.1109/69.868904

Peterson, M., Holbrook, J., Von Hales, D., Smith, N., and Staker, L. (1992). Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *West J. Med.* 156, 163–165.

Richens, J., Lee, C., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* 11, 3923. doi: 10.1038/s41467-020-17419-7

Rowland, S. P., Fitzgerald, J. E., Holme, T., Powell, J., and McGregor, A. (2020). What is the clinical value of mHealth for patients? *NPJ Digit. Med.* 3:4. doi: 10.1038/s41746-019-0206-x

Semigran, H., Levine, D., and Nundy, S., and Mehrotra, A. (2016). Comparison of physician and computer diagnostic accuracy. *JAMA Intern. Med.* 176, 1860–1861 doi: 10.1001/jamainternmed.2016.6001

Semigran, H., Linder, J., Gidengil, C., and Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351, h3480. doi: 10.1136/bmj.h3480

Shwe, M., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., et al. (1990). "A probabilistic reformulation of the quick medical reference system," in *Proc Annu Symp Comput Appl Med Care* (Washington, DC), 790–794.

Tehrani, A., Lee, H., Mathews, S., et al. (2013). 25-year summary of U.S. malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank. *BMJ Qual. Saf.* 22, 672–680. doi: 10.1136/bmjqs-2012-001550

Van Veen, T., Binz, S., Muminovic, M., Chaudhry, K., Rose, K., Calo, S., et al. (2019). Potential of mobile health technology to reduce health disparities in underserved communities. *West J. Emerg. Med.* 20, 799–802. doi: 10.5811/westjem.2019.6.41911

Velikova, M., van Scheltinga, J. T., Lucas, P. J. F., and Spaanderman, M. (2014). Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare *Int. J. Approx. Reason.* 55, 59–73. doi: 10.1016/j.ijar.2013.03.016

Yu, H., Haug, P. J., Lincoln, M. J., Turner, C., and Warner, H. R. (1988). "Clustered knowledge representation: increasing the reliability of computerized expert systems," in *Proceedings of the Annual Symposium on Computer Application in Medical Care* (Washington, DC), 126–130.

Yu, S., Ma, A., Tsang, V., et al. (2019). Triage accuracy of online symptom checkers for accident and emergency department patients. *Hong Kong J. Emerg. Med.* 27, 217. doi: 10.1177/1024907919842486

Zagorecki, A., Orzechowskia, P., and Hołowniaa, K. (2013). A system for automated general medical diagnosis using Bayesian networks. *Stud. Health Technol. Inform.* 192, 461–465. doi: 10.3233/978-1-61499-289-9-461

Conflict of Interest: AJ and DJ are with Eureka Clinical Computing, the creator of the MidasMed DA.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 09 August 2022
DOI 10.3389/fmed.2022.928642

OPEN ACCESS

EDITED BY
Jingjing You,
The University of Sydney, Australia

REVIEWED BY
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Ramadhan J. Mstafa,
University of Zakho, Iraq

*CORRESPONDENCE
Qin Yanguo
qinyg@jlu.edu.cn

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 26 April 2022
ACCEPTED 11 July 2022
PUBLISHED 09 August 2022

CITATION
Xiongfeng T, Yingzhi L, Xianyue S, Meng H, Bo C, Deming G and Yanguo Q (2022) Automated detection of knee cystic lesions on magnetic resonance imaging using deep learning.
Front. Med. 9:928642.
doi: 10.3389/fmed.2022.928642

Automated detection of knee cystic lesions on magnetic resonance imaging using deep learning

Tang Xiongfeng, Li Yingzhi, Shen Xianyue, He Meng, Chen Bo, Guo Deming and Qin Yanguo*

Department of Orthopaedics, The Second Hospital of Jilin University, Changchun, China

Background: Cystic lesions are frequently observed in knee joint diseases and are usually associated with joint pain, degenerative disorders, or acute injury. Magnetic resonance imaging-based, artificial intelligence-assisted cyst detection is an effective method to improve the whole knee joint analysis. However, few studies have investigated this method. This study is the first attempt at auto-detection of knee cysts based on deep learning methods.

Methods: This retrospective study collected data from 282 subjects with knee cysts confirmed at our institution from January to October 2021. A Squeeze-and-Excitation (SE) inception attention-based You only look once version 5 (SE-YOLOv5) model was developed based on a self-attention mechanism for knee cyst-like lesion detection and differentiation from knee effusions, both characterized by high T2-weighted signals in magnetic resonance imaging (MRI) scans. Model performance was evaluated *via* metrics including accuracy, precision, recall, mean average precision (mAP), F1 score, and frames per second (fps).

Results: The deep learning model could accurately identify knee MRI scans and auto-detect both obvious cyst lesions and small ones with inconspicuous contrasts. The SE-YOLO V5 model constructed in this study yielded superior performance (F1 = 0.879, precision = 0.887, recall = 0.872, all class mAP0.5 = 0.944, effusion mAP = 0.945, cyst mAP = 0.942) and improved detection speed compared to a traditional YOLO model.

Conclusion: This proof-of-concept study examined whether deep learning models could detect knee cysts and distinguish them from knee effusions. The results demonstrated that the classical Yolo V5 and proposed SE-Yolo V5 models could accurately identify cysts.

KEYWORDS

knee joint, cyst, effusion, magnetic resonance imaging, deep learning

Introduction

Benign cysts are frequently encountered during body examinations or advanced knee imaging. Cysts can be categorized into various types, including Baker's cysts, proximal tibiofibular joint cysts, meniscal cysts, and intraosseous cysts at the insertion of the cruciate ligaments (1). Intra- and periarticular cyst-like lesions are secondary phenomena likely to be observed in painful or osteoarthritis (OA) affected knees (2). They are strongly associated with intra-articular pathologies or complications of various disorders, such as trauma, meniscus injury, infection, inflammatory arthritis, and malignant lesions (3). Cysts and joint effusion are also key features in two semi-quantitative assessments of knee OA, the Whole-Organ Magnetic Resonance Imaging Score (WORMS) and the MRI Osteoarthritis Knee Score (MOAKS) (4, 5). Such fluid accumulation may range from benign to minimally symptomatic and poses a diagnostic dilemma if one is unaware of the potential diagnoses and pitfalls (3). Therefore, it is crucial to develop an appropriate differential diagnosis of knee cystic lesions to guide further evaluation and treatment of OA.

Magnetic resonance imaging is commonly used to confirm whether lesions are cystic due to its superior soft-tissue contrast and multi-planar imaging capabilities compared to other imaging modalities (1). MRI can help delineate the location of lesions concerning anatomic structures and, with the application of contrast, determine if lesions are cystic or solid (6). Typically, cysts located around the knee are encapsulated fluid collections with low T1-weighted signals and high T2-weighted signals on MR scans, similar to benign intra-articular fluid collections, effusions, or certain types of soft-tissue tumors (7–10). Radiologists and clinicians must familiarize themselves with the MRI features of the cyst and cyst-like lesions to accurately diagnose the disease, develop treatment plans, and manage patients more effectively.

Artificial intelligence and deep learning are increasingly utilized in the medical field both in medical imaging and biomedical analysis (11, 12). The role of AI in medical imaging of knee joints has been described in many primary publications (13), with an emphasis on OA-related research, such as auto-segmentation of knee joint tissue (14, 15), and auto-detection of cartilage lesions, meniscus injuries, and anterior cruciate ligament tears (16–19). The deep learning models for such detection demonstrated relatively superb accuracy, ranging between 70 and 100% across various studies, suggesting that such methods exhibit the potential to rival human-level performance in decision-making tasks related to the MRI-based diagnosis of knee injuries. These methods promote the growth of medical enterprises and help create more intelligent medical services.

Most of the current deep learning research on the knee joint focuses on knee OA and acute knee injuries, but few studies have examined knee joint cysts, cyst-like lesions, or joint effusion. In 2018, a deep convolutional neural network (CNN) was applied to the segmentation of knee joint anatomy, achieving dice coefficients between 0.7 and 0.8 for both joint effusion and Baker's cyst for each joint (20). A more recent study constructed a dense neural network (CNN) for detecting effusions, defined as nonzero MOAKS-ES scores, from limited MRI scans (21). It was demonstrated that NNs could classify knee effusions from low-resolution images with similar accuracy to human radiologists, suggesting that automated evaluation of scans from low-cost, low-field scanners could help assess knee effusions. Other than these two publications, there is no other literature on applying deep learning to cyst detection. It remains unclear whether deep learning techniques can detect cysts and distinguish them from effusions.

Most of the current deep learning research about knee joints focuses on knee osteoarthritis and acute knee injuries, and very few studies examine knee joint cysts, cyst-like lesions, or joint effusion. In 2018, a deep convolutional neural network was applied to the segmentation of knee joint anatomy in a study published by Liu et al. (20). Using the deep learning model, 20 subjects in sagittal frequencies selected fat-suppressed 3D fast spin echo sequences were segmented using 12 different joint structures, and a Dice coefficient between 0.7 and 0.8 was achieved for both joint effusion and Baker's cyst for each joint. This is the first attempt at deep learning used on joint effusions and cysts. In 2022, Harvard University Bragi Sveinsson carried out a study that created a dense NN (CNN) for detecting effusions, defined as nonzero MOAKS-ES scores, from limited MRI scans (21). Additionally, it was proved that neural networks can classify knee effusions with similar accuracy to that offered by human radiologists utilizing low-resolution images, suggesting that automated assessment of images from low-cost, low-field scanners may be useful for assessing knee effusions. Other than the two publications mentioned above, there are no other literature reports on the application of deep learning to cyst detection. It is not clear whether deep learning technology can be used to detect cysts and the performance of identifying them from effusions.

The present study introduced a deep learning model for the auto-detection of knee cystic lesions to address this knowledge gap. It evaluated the model's performance in differentiating knee cysts from knee effusions, which could facilitate the early diagnosis and prevention of knee cysts in mass detection by clinicians. To our knowledge, this is the first attempt at automatically detecting knee cysts and distinguishing them from knee effusions using deep learning methods. Because of the limited amount of data, Mosaic augmentation was used in data preprocessing to increase the volume of training data. To enhance the ability to detect cysts of various sizes, Yolo-V5 was used as a backbone network alongside a featured pyramid architecture for detection. An attention mechanism, the SE

TABLE 1 Patient demographics (mean ± s.d.).

Basic information	Total subjects (n = 282)	Female (n = 192)	Male (n = 90)	P-value
Age(years)	52.87 ± 13.22	52.95 ± 13.18	52.95 ± 13.24	–
Height(cm)	164.60 ± 7.63	164.19 ± 7.63	164.61 ± 7.64	0.085
Weight(kg)	68.75 ± 11.87	68.75 ± 11.90	68.81 ± 11.90	0.239
BMI(kg/m2)	25.30 ± 3.55	25.30 ± 3.55	25.36 ± 3.73	0.720
Left/Right	143/139	101/91	42/48	–

module, was added to the model to enhance the contribution of information-rich features in the feature extraction process.

Materials and methods

The Institutional Review Board of the Second Hospital of Jilin University approved this retrospective study (No. SB2021-012).

Patient data selection

All knee MRIs were acquired at the Second Hospital of Jilin University between January 2021 and October 2021. An in-house RIS/PACS search engine was used to identify candidates who met the following inclusion and exclusion criteria. The inclusion criteria were: (I) MRI scan of the knee for space-occupying lesions or swelling, or pain in a knee joint; (II) patient is over 18 years old; and (III) a formal description of a cystic lesion or uncertain space-occupying lesion in the written radiology report. The exclusion criteria were: (I) patient not consenting to usage of their data; (II) patient is under 18 years old; (III) patient with fracture of a knee joint; (IV) images with excessive movement or beam hardening artifacts as described in the report; and (V) images with knee surgery implants. For patients with more than one MRI examination, only the most recent MR scan was selected.

Data were retrieved for subjects diagnosed with knee cysts or effusions on the imaging report. If there was uncertainty about including a case, a decision was made after reviewing the original image. A total of 282 cases were included in the final analysis. Patient demographics are listed in Table 1. A detailed data selection flowchart is outlined in Figure 1.

Data process

Magnetic resonance imaging was performed on a GE Discovery MR750 3.0T scanner using a sagittal proton density-weighted fat suppression sequence (PD-FS) [Field of view (FOV) = 160 mm × 160 mm; matrix = 512 × 512; number of slices = 20; voxel resolution = 0.35 × 0.35 × 4.5 mm; slice thickness = 3.5 mm; interslice gap = 4.5 mm; repetition time (TR) = 2,600 ms; echo time (TE) = 34.0 ms; flip angle = 90°]. A total of 5,640 sagittal PD-FS images from all subjects were included in this dataset.

Digital Imaging and Communications in Medicine (DICOM) images were converted to one-channel grayscale PNG images to standardize the format of the image files before training. Images were then rescaled to 256 × 256 pixels, and pixel values were normalized between 0 and 1. Two physicians verified that no information related to knee cyst enlargement and effusion was lost in the PNG format images.

Subsequently, regions of interest (ROIs) of cyst lesions and effusions were annotated using the LabelImg image data annotation software by two resident physicians under the supervision of the chief physician. If the annotation was questionable, the final determination was decided by negotiation with the review panel. Background information surrounding the ROIs was removed whenever possible. Annotation files were stored in Pascal-VOC format during the process. Subsequently, the images and their associated annotation files were divided into a training set, a validation set, and a test set in a ratio of 6:2:2 in the enhanced data set through a Python script. The data

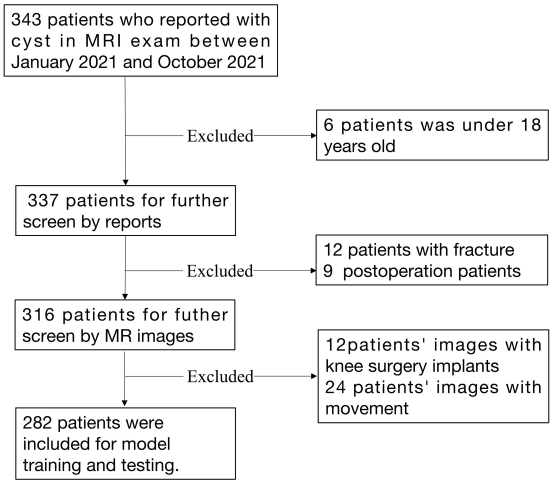

FIGURE 1
Flowchart of subject inclusion/exclusion and data selection.

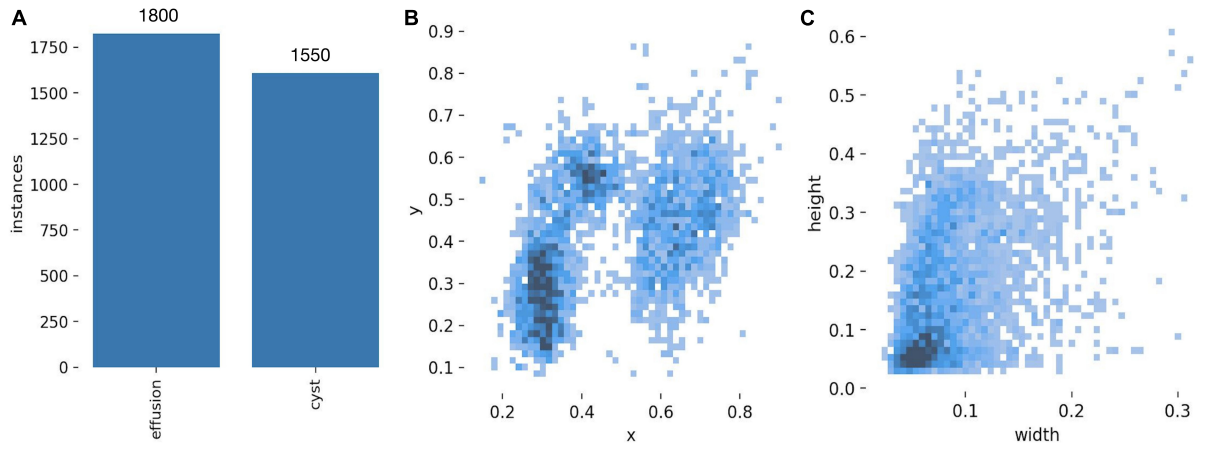

FIGURE 2
(A) Distribution of cyst category. **(B)** Distribution of centroids of cysts and effusions. **(C)** Size distribution of cysts and effusions.

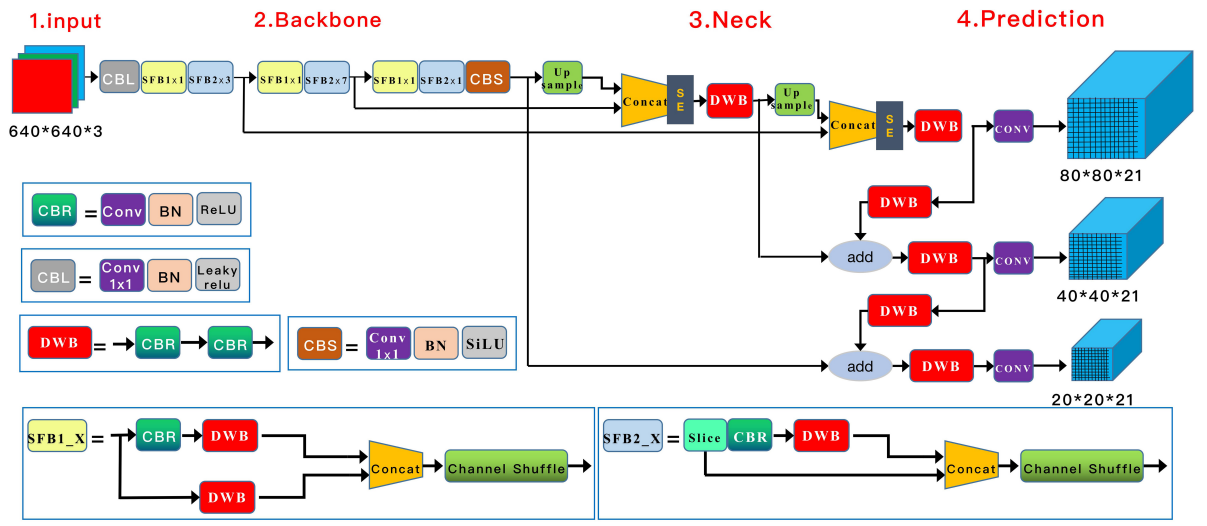

FIGURE 3
SE-YOLOv5 model architecture for cyst detection.

distribution of each lesion category and characteristic is shown in Figure 2.

Deep learning model structure

A Squeeze-and-Excitation (SE) inception attention-based YOLO v5 algorithm (Yolo V5-SE) was adopted to detect knee cyst targets. Similar to the general Yolo v5 algorithm, the architecture of our model was composed of four parts, input, backbone, neck, and prediction, with adjustments in the input and neck parts. In the input preprocessing stage, the images were resized to 640 × 640 × 3, and mosaic data augmentation was applied to increase the number of training samples. Through operations such as flipping, zooming, and color gamut modification, this strategy allowed smaller cyst elements to be detected in a smaller field of sensation, thus enhancing the likelihood of detecting small targets. A couple of SE-inception modules were added after the Concat module in the neck structure (22). The architecture of the model is shown in Figure 3.

Model training and evaluation

During model training, the learning rate was set to 0.0001 to accelerate model convergence. Stochastic gradient descent (SGD) was used for hyperparameter tuning, and the learning

rate momentum was set to 0.90, considering the small number of samples in the cystic lesion dataset. A cosine annealing decay strategy was used for a learning rate change. Cross-entropy was used as the loss function for the model, with batch size set to 8 and training epochs set to 300. The training process was controlled by the early stop method. The training was stopped to prevent over-fitting when the loss value of the validation set did not decrease within 15 epochs. The environment configuration used in the experiment is shown in Table 2.

Validation metrics, including accuracy, precision, recall, mean average precision (mAP), and F1 score, were calculated and visualized in Python to evaluate model performance in cyst and effusion detection (Figure 4). The formulas for the metrics are described below.

$$\text{Precision} = (\text{TP})/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = (\text{TP})/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} \quad (3)$$

TABLE 2 The environment configuration used in the experiment.

Environment	Detail
Central Processing Unit(CPU)	Intel i7-8700k
Opertating system	Window 10
Graphic Processing Unit(GPU)	NVIDIA Geforce GTX1080i 11G
Pytorch version	Pytorch1.8.1 Opencv 4.5.0

$$\text{AP} = \int_0^1 \text{P(R)dR} \quad (4)$$

$$\text{mAP} = \frac{1}{\text{C}} \sum_{\text{i}}^{\text{C}} = 1\text{AP(i)} \quad (5)$$

$$\text{F score} = 2 * (\}\text{precision} * \text{recall})/(\}\text{precision} + \text{recall}) \quad (6)$$

True positives (TP) denote correctly identified cysts, false positives (FP) denote incorrectly identified cysts, and false negatives (FN) denote missed cysts. AP describes average precision; P(R), which denotes the precision P of different recall rates R, corresponds to the P–R curve's area under the curve. The constant C in Eq. 5 has a value of 2, representing cysts and effusions as two separate lesions. The number of average precisions (AP) in each category, which is the number of APs in each category when intersection over union (IoU) is 0.5, is denoted as the mean average precision (mAP). Among these metrics, mAP is the most comprehensive index for evaluating model performance, with higher mAP values corresponding to better model performance.

Furthermore, we compared the performance of our Yolo V5-SE model with that of a general Yolo V5 model by comparing the validation metrics. All statistical tests were performed with SPSS Statistics 26.0 (IBM Corp, Armonk, NY, United States).

Results

Figure 2A shows that the proportion of effusions and cysts was relatively balanced, suggesting that model performance was

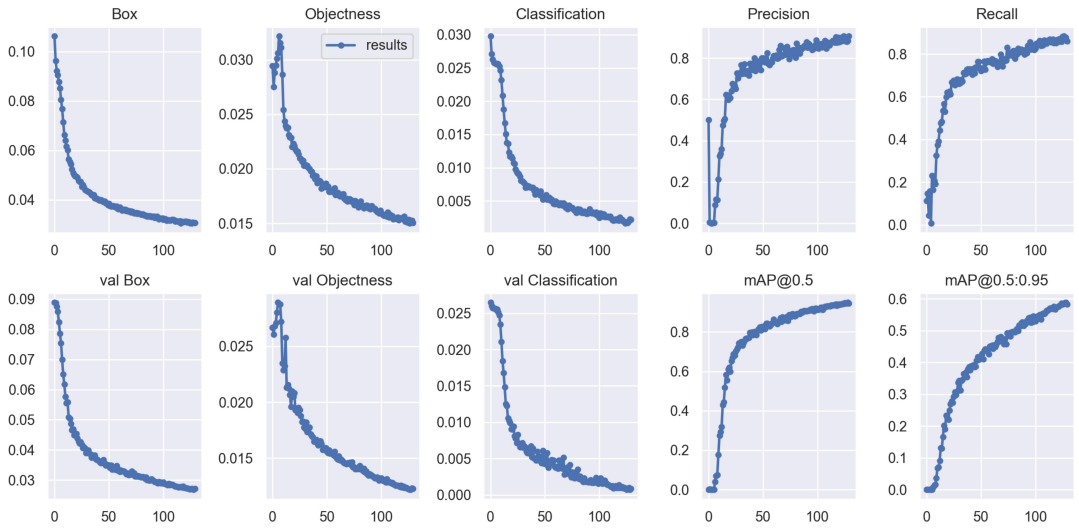

FIGURE 4
Validation metrics for SE-YOLO V5. The horizontal axis denotes the number of iterations.

unlikely to be biased by an imbalanced class distribution. Few lesion centroids were concentrated near the image center, and the distribution of lesion targets was fairly uniform (Figure 2B). Small target lesions accounted for many lesions (Figure 2C).

Validation metrics demonstrated that the model's performance gradually steadied with the training process, indicating that the model converged quickly and yielded good performance. To assess model performance, our proposed SE model was compared against a classical model, YOLOv5, on a series of performance metrics (Table 3). The SE-YOLO V5 model we presented was superior in all performance metrics (F1 = 0.879, precision = 0.887, recall = 0.872, all class mAP0.5 = 0.944, effusion mAP = 0.945, cyst mAP = 0.942). The fps for SE-Yolo v5 was 90.9, suggesting that it could handle more images per unit time. The P–R curves and confusion matrices for these two models are shown in Figures 5, 6. Figure 7 shows example model prediction results compared to the ground truth, indicating that cyst lesions were correctly detected and distinguished from effusions.

TABLE 3 Performance metrics of the SE model and the traditional model.

Metrics	SE-Yolo V5s	Yolo V5s	*P*-value
All class F1 score	0.879 ± 0.002	0.832 ± 0.010	0.002**
All class Precision	0.887 ± 0.011	0.843 ± 0.012	0.011*
All class Recall	0.872 ± 0.014	0.821 ± 0.018	0.018*
All class mAP 0.5	0.944 ± 0.002	0.898 ± 0.011	0.002**
Cyst F1 score	0.875 ± 0.004	0.819 ± 0.016	0.005**
Cyst Precision	0.873 ± 0.012	0.822 ± 0.017	0.014*
Cyst Recall	0.878 ± 0.006	0.818 ± 0.027	0.021*
Cyst mAP 0.5	0.942 ± 0.005	0.893 ± 0.019	0.011*
Effusion F1 score	0.883 ± 0.006	0.843 ± 0.005	0.001**
Effusion Precision	0.902 ± 0.011	0.864 ± 0.008	0.014*
Effusion Recall	0.865 ± 0.022	0.822 ± 0.009	0.037*
Effusion mAP 0.5	0.945 ± 0.001	0.901 ± 0.004	<0.001***

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Discussion

This proof-of-concept study aimed to demonstrate the feasibility of a deep learning system for the auto-detection and classification of knee cysts. The SE-YoloV5 attention model was constructed, trained, and evaluated on clinical MR images. Analysis of model performance indicated that this approach promises to improving diagnostic accuracy.

Deep learning offers excellent performance for segmenting multi-tissue knee joints and detecting ACL, cartilage, or meniscus injuries (15–17). However, few papers have addressed cysts and effusions of the knee joint, which are associated with high morbidity and could also serve as biomarkers for degenerative disorders or acute injuries, like knee osteoarthritis and meniscus injuries. Considering the importance and the potential pitfalls of knee cyst diagnosis, it is beneficial to develop an auto-diagnostic system for cyst detection, which may be used as a primary or supplementary tool to speed up diagnosis and enhance accuracy. Two papers have explored the application of deep learning in cyst segmentation and effusion estimation (20, 21); Zhou et al. (20) demonstrated the application of deep learning in Baker's cyst and joint effusion auto-segmentation and achieved a dice coefficient of 0.736. Raman reported the feasibility of classifying knee effusion based on neural networks, which could achieve an average accuracy of 62%, comparable to a radiologist in a small test dataset (21). Other than the two publications mentioned above, there are no other literature reports on the application of deep learning to cyst detection. To our

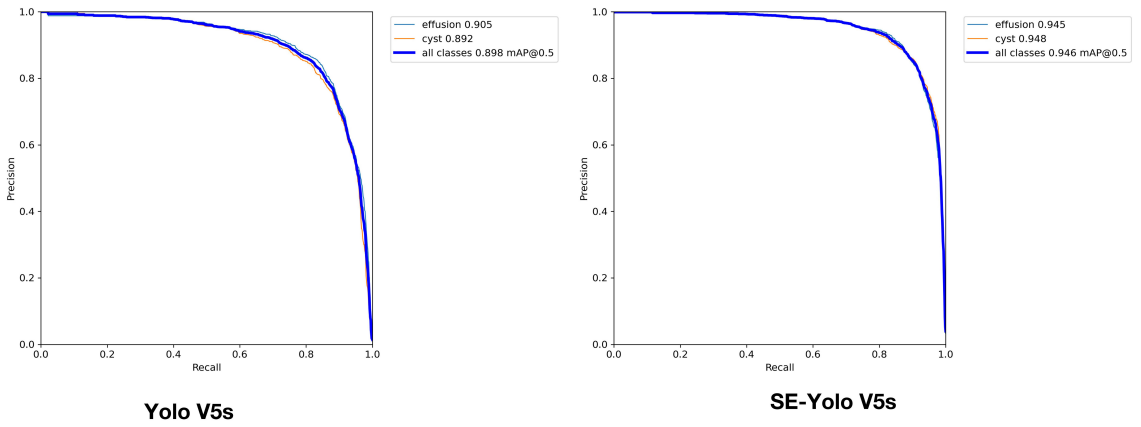

FIGURE 5
P-R curves of YOLO V5 and YOLO V5-SE.

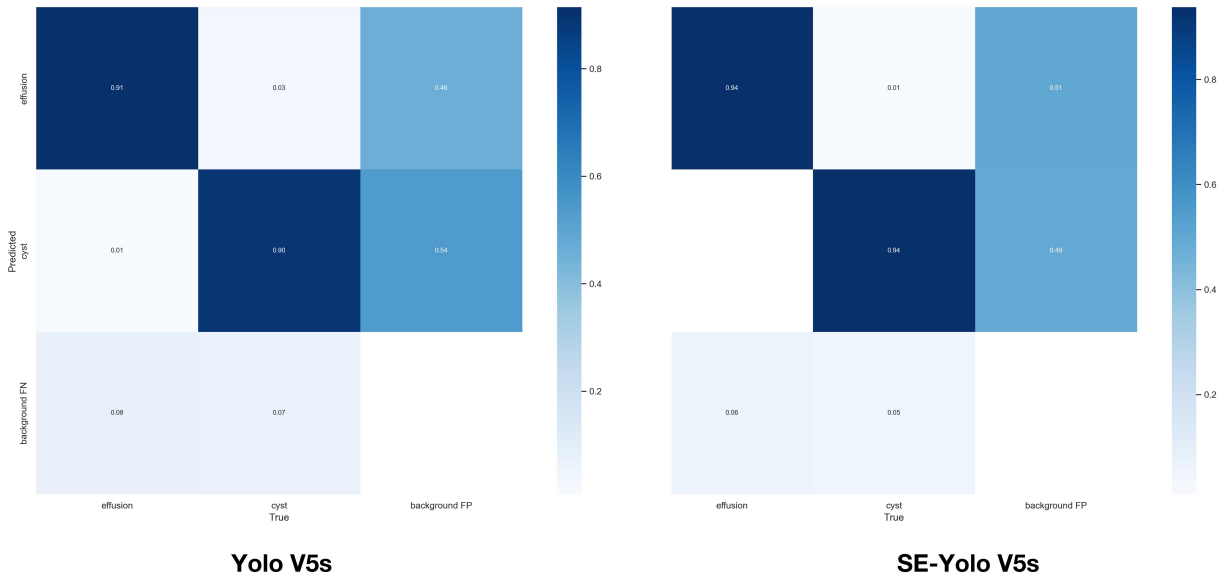

FIGURE 6
Confusion matrices of YOLO V5 and YOLO V5-SE.

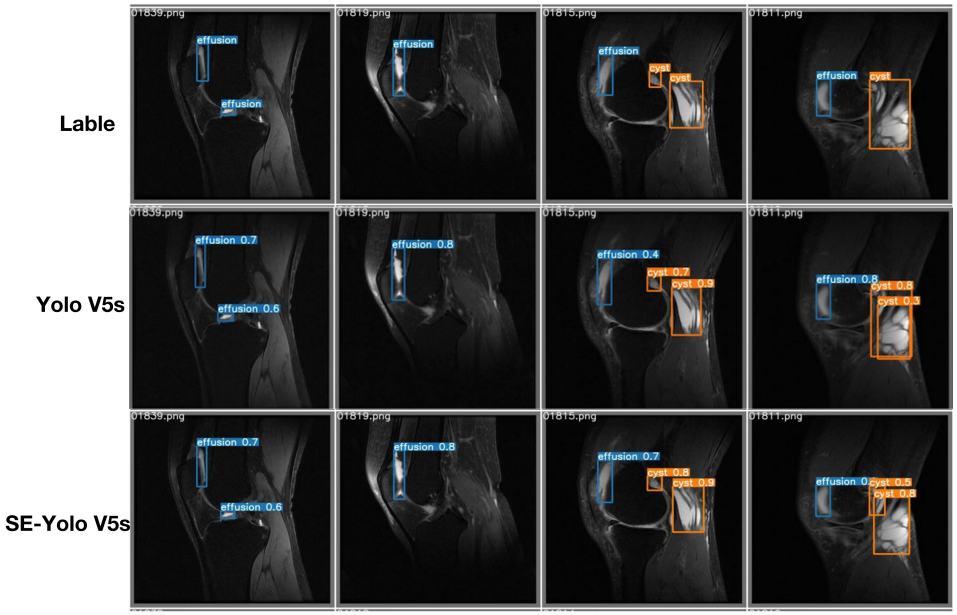

FIGURE 7
Example prediction outcomes of YOLO V5 and YOLO V5-SE compared with the ground truth.

knowledge, this paper is the first to use deep learning in knee cyst detection.

Cyst detection is an object detection task in nature. Object detection is a primary computer vision task that entails determining where particular objects are in an image and classifying them. YOLO, a new algorithm deployed in 2015 (23), redefined object recognition as a regression problem that can be performed in a single neural network. Yolo has been updated to version five and is regarded as the state-of-the-art algorithm for object detection (24). It has been applied in many daily life aspects, such as the detection of surface knots (25) and real-time vehicles (26), as well as in various medical fields, including face mask recognition (27), breast tumor detection and classification (28), and chest abnormality detection (29). This study showed

that the basic deep learning model Yolo V5 could handle the cyst-detection task, attaining F1, precision, and mAP scores of 0.832, 0.843, and 0.821, respectively. After the attention SE module was added to the Yolo V5 model, the resulting attention-based model SE-Yolo V5 achieved better accuracy and higher speed of 0.879, 0.87, and 0.944 for F1 score, precision, and mAP, respectively. Small target lesions accounted for a significant proportion of our dataset, but the proposed model was also capable of detecting them accurately, as illustrated in **Figure 7**.

This paper aimed to demonstrate the feasibility of utilizing deep learning in general knee cyst detection. Despite its promise, there are several limitations to the presented model. First, there are many cyst types, such as Baker's cysts, meniscal cysts, and intraosseous cysts at the insertion of the cruciate ligaments, but these different cyst sub-types were not explicitly classified in this study. Neither did we verify whether deep learning performed equally well in these sub-groups. We may enroll more kinds of knee cysts in the future and evaluate the model's performance on different cyst types. Second, our data was relatively limited, and model performance was not compared with human diagnosis. Nevertheless, the model prediction proved efficient and reliable, suggesting that the model may become a valuable tool for radiologists and clinicians, subject to further study and multi-center validation. Third, the cysts were easily classified based on the reports or images, but there was no general standard for diagnosing inherent effusions, which might be a caveat for the model, radiologist opinions, and the ground truth labels. Last but not least, the uncertainties and interpretability of the model should be mentioned, and we will explore them in further studies. To explore the model in the external datasets or public datasets.

Conclusion

This proof-of-concept study examined whether deep learning models could detect knee cysts and distinguish them from knee effusions and demonstrated that the classical Yolo V5 and proposed SE-Yolo V5 models could identify cysts with high accuracy. This study suggested that cutting-edge deep learning methods constitute a promising avenue of research to develop AI-assisted auto-detection systems to facilitate radiological and clinical diagnosis of knee pathologies.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the local Institutional Review Board (The Second Hospital of Jilin University). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

TX and QY: conceptualization. TX, HM, and CB: methodology and formal analysis. HM, GD, CB, and SX: validation. HM, GD, and CB: investigation. HM and TX: resources and data curation. TX: writing—original draft preparation. SX and LY: writing—review and editing. QY: supervision and funding acquisition. TX, SX, and QY: project administration. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the National Natural Science Foundation of China (U21A20390), the Special Foundation for Science and Technology Innovation of Jilin (20200601001JC), the Health Service Capacity Building Projects of Jilin Province (05KA001026009002), and Nature Science Foundation of Jilin Province (20200201536JC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Telischak NA, Wu JS, Eisenberg RL. Cysts and cystic-appearing lesions of the knee: A pictorial essay. *Indian J Radiol Imaging.* (2014) 24:182–91. doi: 10.4103/0971-3026.134413

2. Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. *Neurocomputing.* (2016) 214:376–81. doi: 10.1016/j.neucom.2016.06.023

3. Stein D, Cantlon M, Mackay B, Hoelscher C. Cysts about the knee: Evaluation and management. *J Am Acad Orthop Surg.* (2013) 21:469–79. doi: 10.5435/jaaos-21-08-469

4. Peterfy C, Guermazi A, Zaim S, Tirman P, Miaux Y, White D, et al. Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis. *Osteoarthr Cartil.* (2004) 12:177–90.

5. Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). *Osteoarthr Cartil.* (2011) 19:990–1002. doi: 10.1016/j.joca.2011.05.004

6. Beaman FD, Peterson JJ. MR imaging of cysts, ganglia, and bursae about the knee. *Magn Reson Imaging Clin N Am.* (2007) 15:39–52. doi: 10.1016/j.mric.2007.02.001

7. McCarthy CL, McNally EG. The MRI appearance of cystic lesions around the knee. *Skeletal Radiol.* (2004) 33:187–209.

8. Marra MD, Crema MD, Chung M, Roemer FW, Hunter DJ, Zaim S, et al. MRI features of cystic lesions around the knee. *Knee.* (2008) 15:423–38. doi: 10.1016/j.knee.2008.04.009

9. Hayashi D, Roemer FW, Dhina Z, Kwoh CK, Hannon MJ, Moore C, et al. Longitudinal assessment of cyst-like lesions of the knee and their relation to radiographic osteoarthritis and MRI-detected effusion and synovitis in patients with knee pain. *Arthritis Res Ther.* (2010) 12:1–9. doi: 10.1186/ar3132

10. Perdikakis E, Skiadas V. MRI characteristics of cysts and "cyst-like" lesions in and around the knee: What the radiologist needs to know. *Insights Imaging.* (2013) 4:257–72. doi: 10.1007/s13244-013-0240-1

11. Le NQK. Potential of deep representative learning features to interpret the sequence information in proteomics. *Proteomics.* (2022) 22:e2100232. doi: 10.1002/pmic.202100232

12. Tng SS, Le NQK, Yeh HY, Chua MCH. Improved prediction model of protein lysine crotonylation sites using bidirectional recurrent neural networks. *J Proteome Res.* (2022) 21:265–73. doi: 10.1021/acs.jproteome.1c00848

13. Si L, Zhong J, Huo J, Xuan K, Zhuang Z, Hu Y, et al. Deep learning in knee imaging: A systematic review utilizing a checklist for artificial intelligence in medical imaging (CLAIM). *Eur Radiol.* (2022) 32:1353–61. doi: 10.1007/s00330-021-08190-4

14. Gaj S, Yang M, Nakamura K, Li X. Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks. *Magn Reson Med.* (2020) 84:437–49. doi: 10.1002/mrm.28111

15. Panfilov E, Tiulpin A, Nieminen MT, Saarakkala S, Casula V. Deep learning-based segmentation of knee MRI for fully automatic subregional morphological assessment of cartilage tissues: Data from the osteoarthritis initiative. *J Orthop Res.* (2022) 40:1113–24. doi: 10.1002/jor.25150

16. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* (2018) 15:e1002699. doi: 10.1371/journal.pmed.1002699

17. Astuto B, Flament I, Namiri KN, Shah R, Bharadwaj U, Link TM, et al. Automatic deep learning-assisted detection and grading of abnormalities in knee MRI studies. *Radiol Artif Intell.* (2021) 3:e200165. doi: 10.1148/ryai.2021200165

18. Li Z, Ren S, Zhou R, Jiang X, You T, Li C, et al. Deep learning-based magnetic resonance imaging image features for diagnosis of anterior cruciate ligament injury. *J Healthc Eng.* (2021) 2021:4076175. doi: 10.1155/2021/4076175

19. Tack A, Shestakov A, Ludke D, Zachow S. A multi-task deep learning method for detection of meniscal tears in MRI data from the osteoarthritis initiative database. *Front Bioeng Biotechnol.* (2021) 9:747217. doi: 10.3389/fbioe.2021.747217

20. Zhou Z, Zhao G, Kijowski R, Liu F. Deep convolutional neural network for segmentation of knee joint anatomy. *Magn Reson Med.* (2018) 80:2759–70. doi: 10.1002/mrm.27229

21. Raman S, Gold GE, Rosen MS, Sveinsson B. Automatic estimation of knee effusion from limited MRI data. *Sci Rep.* (2022) 12:3155. doi: 10.1038/s41598-022-07092-9

22. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Piscataway, NJ: IEEE) (2018). p. 7132–41.

23. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: (2016). p. 779–88.

24. Thuan D. *Evolution of Yolo Algorithm and Yolov5: The State-Of-The-Art Object Detection Algorithm*. Bachelor's thesis. Oulu: Oulu University of Applied Sciences (2021).

25. Fang Y, Guo X, Chen K, Zhou Z, Ye Q. Accurate and automated detection of surface knots on sawn timbers using YOLO-V5 Model. *BioResources.* (2021) 16:5390–406.

26. Wu T-H, Wang T-W, Liu Y-Q. Real-time vehicle and distance detection based on improved yolo v5 network. *Proceedings of the 3rd World Symposium on Artificial Intelligence (WSAI)*. (Guangzhou: IEEE) (2021). p. 24–8.

27. Yang G, Feng W, Jin J, Lei Q, Li X, Gui G, et al. Face mask recognition system with YOLOV5 based on image recognition. *Proceedings of the IEEE 6th International Conference on Computer and Communications (ICCC)*. (Chengdu: IEEE) (2020). p. 1398–404.

28. Mohiyuddin A, Basharat A, Ghani U, Peter V, Abbas S, Naeem OB, et al. Breast tumor detection and classification in mammogram images using modified YOLOv5 network. *Comput Math Methods Med.* (2022) 2022:1359019. doi: 10.1155/2022/1359019

29. Yap MH, Hachiuma R, Alavi A, Brüngel R, Cassidy B, Goyal M, et al. Deep learning in diabetic foot ulcers detection: A comprehensive evaluation. *Comput Biol Med.* (2021) 135:104596. doi: 10.1016/j.compbiomed.2021.104596

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fdata.2022.941451

OPEN ACCESS

EDITED BY
Francesco Napolitano,
University of Sannio, Italy

REVIEWED BY
Cristian Axenie,
Technische Hochschule Ingolstadt, Germany
OPhir Nave,
Jerusalem College of Technology, Israel

*CORRESPONDENCE
Helena Coggan
helena.coggan.21@ucl.ac.uk

SPECIALTY SECTION
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

RECEIVED 11 May 2022
ACCEPTED 17 August 2022
PUBLISHED 12 September 2022

CITATION
Coggan H, Andres Terre H and Liò P (2022) A novel interpretable machine learning algorithm to identify optimal parameter space for cancer growth. *Front. Big Data* 5:941451. doi: 10.3389/fdata.2022.941451

A novel interpretable machine learning algorithm to identify optimal parameter space for cancer growth

Helena Coggan [1]*, Helena Andres Terre [2] and Pietro Liò [2]

[1]Department of Mathematics, University College London, London, United Kingdom, [2]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom

Recent years have seen an increase in the application of machine learning to the analysis of physical and biological systems, including cancer progression. A fundamental downside to these tools is that their complexity and nonlinearity makes it almost impossible to establish a deterministic, *a priori* relationship between their input and output, and thus their predictions are not wholly accountable. We begin with a series of proofs establishing that this holds even for the simplest possible model of a neural network; the effects of specific loss functions are explored more fully in Appendices. We return to first principles and consider how to construct a physics-inspired model of tumor growth without resorting to stochastic gradient descent or artificial nonlinearities. We derive an algorithm which explores the space of possible parameters in a model of tumor growth and identifies candidate equations much faster than a simulated annealing approach. We test this algorithm on synthetic tumor-growth trajectories and show that it can efficiently and reliably narrow down the area of parameter space where the correct values are located. This approach has the potential to greatly improve the speed and reliability with which patient-specific models of cancer growth can be identified in a clinical setting.

KEYWORDS

cancer, neural networks, white-box machine learning, interpretability, parameter optimization

1. Introduction

The application of neural networks to the modeling of cancer has seen a flood of interest in recent years (Sanoob et al., 2016; Hsu et al., 2018; Ghazani et al., 2021; Kwak et al., 2021; Kumar et al., 2022). The hope is to be able to use patient-specific data to generate accurate predictions of tumor growth and treatment response, in order to guide the clinician in their prognosis and choice of treatment regime (Rockne et al., 2019; Kumar et al., 2022). From a modeling perspective, a tumor is a system of interacting objects (tumor cells, fibroblasts, etc.) which influence each other's behavior according

to certain rules. It should therefore be possible to use tumor-growth data to derive a system of equations to describe the trajectory of cancer, which can then be extrapolated into the future to predict the course of a particular disease. Over the last few years, neural networks have become the natural first choice of most scientists when tasked with extracting such equations from large datasets (Benzekry, 2020; Kurz et al., 2021). However, when we resort to machine learning to build models and predict the behavior of any system, we sacrifice a crucial attribute: *explainability*. The sheer vastness of a neural network, which may contain many tens of thousands of continually-adjusted interacting weights, makes the effort of deducing the impact of any single component on a network's output almost impossible. In addition, we must consider the neural network's various nonlinearities, which interfere with any attempt to construct an analytically solvable description of its processes (and thus to account for its decision-making). One example is the common Rectified Linear Unit (ReLU), and its many cousins [the parameterized ReLU (Xu et al., 2015), the "leaky" ReLU (Maas et al., 2013), etc.], which may or may not act on an input as it makes its way through the system. Any attempt to construct a gradient of the output with respect to the input will have to contend with the resulting discontinuities. Less analytically troublesome, but still exhausting, are backpropagation algorithms: ADAM (Kingma and Ba, 2014), for instance, adjusts each weight not simply in response to its current effect on the output but to all of its past effects, which will create a new set of complex nonlinearities in any differential equation aimed at describing a the workings of a network.

The best that can be hoped for, then, is to gain a "general idea" of the effect of each network attribute, using hyperparameter tuning (Yuan et al., 2021). This is an obviously risky approach: sampling a few points in the hyperspace of all possible hyperparameter values does not give us a complete picture of the dependence of the output on our choice of values. Without a complete picture of this dependence, we can never be sure that the relationships predicted by a network reflect physical reality or are simply a product of its own internal calibration. This is the crucial issue, and why, as long as a neural network remains a "black box," its output can never be fully understood or trusted, especially in a clinical setting where the results of a model may guide cancer treatment and thus affect a patient's length and quality of life. A lack of explainability is a significant impediment to the adoption of machine learning and other computational approaches in a clinical setting. It also hinders the clinician's ability to fully interact with and analyse ML-derived predictions: not knowing where they come from, it is very difficult to rigorously deduce what any set of values assigned to a tumor "mean," or to "sanity-check" them against clinical expertize. To reliably incorporate computational methods into cancer treatment, we must either develop some picture of the workings of a neural network, or move away from stochastic gradient descent altogether, to an algorithmic approach whose decision-making processes are transparent and accountable. A great deal of interesting work has been done in recent years to achieve this first goal, attempting to render explainable the workings of black-box neural networks (Rudin, 2019; Kazhdan et al., 2020; Dujon et al., 2021; Magister et al., 2021). The general approach of such papers is either to deduce the emergent rules of the neural network from its behavior, or to induce such strong biases in its workings that it is naturally directs to the correct area of parameter hyperspace (as with the physics-inspired neural networks discussed in Karniadakis et al., 2021). Such *a posteriori* attempts to harness or constrain the chaotic nonlinear workings of a neural network, however, are no replacement for an *a priori* understanding of its rules and aims. Without this, no result derived from such a network can be considered mathematically rigorous, which becomes an increasingly serious problem as the area of application approaches the hard sciences. The aim of this paper is to explore the difficulties inherent in this promising research, and to place some mathematical limits on the degree to which black-boxes can be truly, *a priori* explained. We also develop a computational method of fitting a model to cancer-growth data which is built around explainability first and foremost, excising nonlinearity and stochasticity where possible, and find that such a method can usefully direct and improve the efficiency of standard machine-learning techniques.

This paper is laid out as follows. We demonstrate first that it is impossible to truly account for the workings of even the simplest imaginable neural network, and then introduce an alternative "white-box" algorithm which can be used to quickly and reliably identify candidate equations for tumor growth. By using this algorithm, we can explainability identify the region of "parameter space"—and thus, in a sense, the "type" of tumor growth—appropriate to a particular disease. After this step has been applied, we are no longer "fighting blind," and may leave more detailed fitting to neural networks. With this algorithm, we can both significantly reduce the time taken to fit patient-specific models of tumor growth and provide meaning to their parameters. The goal of explainability, then, does not have to slow down machine learning techniques, but can aid them in their search for appropriate models.

2. Materials and methods

2.1. Theory: The barriers to an analytically explainable neural network

In the following section we consider a idealized mathematical model of the graph neural network during its training process, without activation functions and with inductive biases sufficient to describe a physical system of N interacting objects. Each object within the system is represented by a node with two properties: the input "representation" value

x_i (which may represent size, position, age, etc.), and the target property, whose true value is y'_i. By considering many values of x_i and y'_i, we aim to learn the relationship $y_i(x_1, x_2, ...x_N)$ between them; the goal is to produce a value of y_i as close as possible to y'_i on the training data. All properties in this model are one-dimensional for simplicity, but the mathematics behind it may easily be extended to multidimensional systems. Since we are describing observable quantities, we assume all properties are real.

A real graph neural network will use several layers of interconnected weights and activation functions to represent the relationship between any two objects; a separate computational layer will then learn how each object aggregates the information it receives from the rest of the system. In our model, we condense this operation into a single relationship, which we assume is of the form

$$y_i = \sum_{jks} w_{ijks} x_i^k x_j^s \tag{1}$$

where $1 \leq j \leq N$ and k, s in principle range over all integers, so that we are considering the product of two Taylor expansions. In practice, because we cannot store infinite sums, we choose some combinations of j, k, s to describe our system. w_{ijks} are coefficients which we will adjust according to a loss function. This form encodes a number of physical assumptions: firstly, that the relationship y_i is continuous and differentiable; secondly, that it consists of a number of sub-relationships y_{ij}, which combine additively; and thirdly, that the relationship y_{ij}, which describes the effect of object j on object i, is dependent only on the properties of those nodes (i.e., on x_i and x_j) and on no others, i.e., that each object interacts with every other object independently. Less obvious is that we are assuming the relationship is also *local*. Though we presumably have many values of x_i from different time-points, the relationship y_i depends on the value of the representations $\{x_i\}$ only at a single time-point. The system does not know about its previous states, and is assumed to have time-translational symmetry.

Having given the weights w_{ijks} some initial values, we now adjust them continuously according to their contribution to our loss function L, which describes the total "wrongness" of our current guesses:

$$\frac{\partial w_{ijks}}{\partial t} = -\alpha \frac{\partial L}{\partial w_{ijks}} \tag{2}$$

We say the system has converged when no further adjustments remain to be made, i.e., when

$$\frac{\partial w_{ijks}}{\partial t} = \frac{\partial L}{\partial w_{ijks}} = 0 \tag{3}$$

for all weights.

What is the impact of our choice of loss function on the value of the relationships $\{y_i\}$ at convergence? We will use a slightly modified and generalized version of the loss function used by Cranmer et al. (2020), and include one "error" term designed to penalize divergence from target values, and another term, commonly referred to as the "regularization" term (Xu et al., 2015), designed to penalize the overall complexity of the system. The general form of our loss function is

$$L = \sum_i \left| y_i - y'_i \right|^m + \beta \sum_{ijks} \left| w_{ijks} \right|^n \tag{4}$$

Clearly, there are three adjustable hyperparameters here: the positive integers m, n, and the real and positive β. For the loss function closest to that used by Cranmer et al., m = 1 and n = 2, it can be shown that there are two possible values for convergence, depending on the value of the parameter β and the target value y'_i. The proof is as follows and is based on a self-consistency argument.

We have at convergence

$$\frac{\partial L}{\partial w_{ijks}} = \frac{\partial \left| y_i - y'_i \right|}{\partial y_i} \frac{\partial y_i}{\partial w_{ijks}} + 2\beta w_{ijks} = 0 \tag{5}$$

and $\frac{\partial |y_i - y'_i|}{\partial y_i} = 1$ if $y_i \geq y'_i$ and -1 otherwise, i.e., $\frac{\partial |y_i - y'_i|}{\partial y_i} = \frac{y_i - y'_i}{|y_i - y'_i|}$, and $\frac{\partial y_i}{\partial w_{ijks}} = x_i^k x_j^s$, so we have convergence when

$$\frac{\partial L}{\partial w_{ijks}} = \frac{y_i - y'_i}{|y_i - y'_i|} x_i^k x_j^s + 2\beta w_{ijks} = 0 \tag{6}$$

i.e., if $y_i \geq y'_i$ we have $(y_i - y'_i)(x_i^k x_j^s + 2\beta w_{ijks}) = 0$, and if $y_i < y'_i$ we have $(y_i - y'_i)(x_i^k x_j^s - 2\beta w_{ijks}) = 0$. So convergence at $y_i = y'_i$ is possible for *any value of* w_{ijks}.

For $y_i \geq y'_i$ we also have a solution for convergence at $w_{ijks} = -\frac{x_i^k x_j^s}{2\beta}$. Now we can use our self-consistency argument, because y_i is defined by its contributing weights: thus this solution is possible if

$$y_i = \sum_{jks} w_{ijks} x_i^k x_j^s = \sum_{jks} -\frac{x_i^{2k} x_j^{2s}}{2\beta} \geq y'_i \tag{7}$$

which is to say we can have a different kind of convergence—what we will call "information-free" convergence—at $y_i = \sum_{jks} -\frac{x_i^{2k} x_j^{2s}}{2\beta}$ provided that $y'_i \leq \sum_{jks} -\frac{x_i^{2k} x_j^{2s}}{2\beta} \leq 0$ for all j, k, s combinations used to describe our system. An identical argument for the $y_i < y'_i$ case allows such information-free convergence at $y_i = \sum_{jks} \frac{x_i^{2k} x_j^{2s}}{2\beta}$ if $y'_i > \sum_{jks} \frac{x_i^{2k} x_j^{2s}}{2\beta} \geq 0$.

In summary, then, if $\left| y'_i \right| \leq \frac{\sum_{jks} x_i^{2k} x_j^{2s}}{2\beta}$, then convergence is only reached at $y_i = y'_i$ for all i, with no restriction placed upon the weights w_{ijks}. We refer to this as "absolute convergence."

If any target value falls outside of those restrictions (i.e., $|y_i'| > \frac{\sum_{jks} x_i^{2k} x_j^{2s}}{2\beta}$ for any i), then in addition to absolute convergence, we have a second possibility: that relationship y_i may converge at $|y_i'| = \frac{\sum_{jks} x_i^{2k} x_j^{2s}}{2\beta}$. This is, of course, a completely meaningless value, independent of y_i' and indeed of any individual property of the node i. This is why we refer to this possibility as "information-free" (I-F) convergence. It, too, places no restriction on the value of the weights; the system is not guaranteed to be made any simpler, which of course would be little reassurance, given that the relationship it describes is essentially "random."

From this, we see that we can mitigate the possibility of I-F convergence by setting

$$\beta \ll \frac{\sum_{jks} x_i^{2k} x_j^{2s}}{2}$$

thus widening the range of values of y_i' within which only absolute convergence is possible; and I-F convergence is avoided entirely by setting $\beta = 0$. What, then, is the point of having a regularization term in this model at all, if not for its original intended purpose of making the result 'simpler'? The answer is that it makes convergence *faster*. The speed of convergence of this loss function is determined by

$$\frac{\partial L}{\partial t} = \sum_{ijks} \frac{\partial L}{\partial w_{ijks}} \frac{\partial w_{ijks}}{\partial t} = -\alpha \sum_{ijks} \left(\frac{\partial L}{\partial w_{ijks}}\right)^2 \tag{8}$$

as the weights are adjusted according to $\frac{\partial w_{ijks}}{\partial t} = -\alpha \frac{\partial L}{\partial w_{ijks}}$ within our model. In the limit $\beta \to 0$, $\frac{\partial L}{\partial t} \to -\sum_i \alpha \left(\frac{y_i - y_i'}{|y_i - y_i'|}\right)^2 = -\sum_i \alpha$, i.e., decline is constant and at a rate proportional to α and to the number of objects in the system. Conversely, in the limit $\beta \to \infty$, $L \to \beta \sum_{ijks} w_{ijks}^2$ and $\frac{\partial L}{\partial t} \to -\alpha \sum_{ijks} 4\beta^2 w_{ijks}^2 = -4\alpha\beta L$, so $L = L_0 e^{-4\alpha\beta t}$, and convergence is exponential with time.

This example is simple but illustrative: even within this toy model, the loss function does not have an intuitive effect on convergence values. For the general even-power case $m = n$, it can be shown similarly (proof in Appendix, Section 1) that at convergence,

$$y_i = \frac{y_i'}{1 + \frac{\beta^{\frac{1}{n-1}}}{\sum_{jks} \left(x_i^k x_j^s\right)^{\frac{n}{n-1}}}} \tag{9}$$

with a corresponding equation for weights. We see now the *scale* on which the value of β should be considered: what governs the final output guess is the ratio $\frac{\beta^{\frac{1}{n-1}}}{\sum_{jks} \left(x_i^k x_j^s\right)^{\frac{n}{n-1}}}$. In the limit of large n, since n is even, the denominator tends to $\sum_{jks} \left|x_i^k x_j^s\right|$, which we may think of as the "sum of the total information in the subsystem i." In that limit, the effect of increasing β is blunted by the fact that the relevant quantity is its $n-1$-th root. In the limit $\beta^{\frac{1}{n-1}} \ll \sum_{jks} \left(x_i^k x_j^s\right)^{\frac{n}{n-1}}$, we recover absolute convergence, $y_i \to y_i'$; in the limit $\beta^{\frac{1}{n-1}} \gg \sum_{jks} \left(x_i^k x_j^s\right)^{\frac{n}{n-1}}$, all weights in the subsystem i and the output guess y_i tend to zero. There is no possibility of information-free convergence to a non-zero value. This would seem, then, to be a much more appropriate choice of loss function. In Appendix (Section 1), we briefly discuss the general even-power m, n case, the case $m = n = 2$, and in Appendix (Section 3) we note the behavior of the more niche subcase of elastic regularization (Li et al., 2020).

Until now, we have discussed the effect of loss function hyperparameters on convergence values within an idealized linear model of a neural network. We will now attempt to incorporate the structure of a real neural network into our model—i.e., that of layers of nodes mediated by activation functions.

We model a simple two-layer network. We have two inputs, x_i and x_j, which are fed into a hidden layer of nodes. The node indexed by k within this layer has output

$$v_k = a_{ki} x_i + a_{kj} x_j + b_k \tag{10}$$

and our final guess y (we will drop the subscript i for the moment) is made by combining the outputs of the hidden layer, each fed through an activation function:

$$y = \sum_k c_k \phi\left(v_k\right) + \delta \tag{11}$$

for the activation function used in the rectified linear unit, $\phi(x) = \max(x, 0)$. We will use the loss function (4) with $m = n = 2$ which has bounded error, no information free-convergence, and whose error decays exponentially with time (proof in Appendix, Section 1). Here, it becomes:

$$L = \left(y - y'\right)^2 + \beta \sum_k a_{ki}^2 + a_{kj}^2 + b_k^2 + c_k^2 + \delta^2 \tag{12}$$

At convergence we obtain a self-consistency equation for the node outputs v_k:

$$v_k = \frac{\left(y - y'\right)^2}{\beta^2} \left(x_i^2 + x_j^2 + 1\right) \phi\left(v_k\right) \tag{13}$$

This imposes either $v_k = 0$ or, for $v_k > 0$, $|y - y'| = \frac{\beta}{\sqrt{x_i^2 + x_j^2 + 1}}$, i.e. a minimum error at convergence that tends to infinity with β. Further, constructing the guess y directly from our convergence equations for c_k, we obtain the result (full proof in Appendix, Section 2) that for target guesses within the range

$$|y'| < \frac{\beta + 1}{\sqrt{x_i^2 + x_j^2 + 1}} \tag{14}$$

convergence is *impossible*. Even taking the limit $\beta \to 0$ cannot eliminate this effect entirely, and the range to which it applies widens without bound as $\beta \to \infty$. This is worth restating: in the simplest realistic model of a neural network that incorporates activation functions, there are ranges of representations and target values—unalterable input data—for which convergence becomes mathematically impossible, and the learning process will never terminate. In practice, of course, real networks do not converge only when the gradient of the loss function with respect for each weight is precisely zero: we will consider the network converged when the magnitude of the gradient of each weight has reached some small value ε. From the standpoint of the white-box modeler, unfortunately, this is hardly any better. If there is some large number N_w of weights in the system, then all we can say with certainty is that convergence occurs somewhere within a high-dimensional hyperspace of volume $(2\varepsilon)^{N_w}$, which leaves us with a very large number of possible configurations of the system, of which the "correct" one will be chosen stochastically. The system has become unexplainable once again.

How do we build an algorithm which does not run into these analytical difficulties, and has explainability as its central goal? If our aim is to construct a procedure that can correctly analyze a physical system, whose workings are completely mathematically transparent, and which is guaranteed to converge, our analysis above suggests we should move away from the realm of gradient descent and nonlinear units entirely, and begin from first principles. We follow this approach in the section below.

2.2. A white-box algorithm for characterizing tumor growth

Suppose that we have chosen some i, j, k, s combinations to describe our system, so that we assume relationships are of the form

$$y_i = \sum_{jks} w_{ijks} x_i^k x_j^s = \sum_m f_{im} z_{im} \tag{15}$$

where we have condensed the weights w_{ijks} and terms $x_i^k x_j^s$ corresponding to the combinations $\{(i, j, k, s)\}$ into M_i weights and terms f_{im}, z_{im} corresponding to the object i. We will assume that we have samples of $\{x_i\}$ and $\{y_i'\}$ for all objects, and for several configurations of the system. In all methods discussed above, we considered each timepoint independently; here we will combine them, and attempt to find the coefficients $\{f_{im}\}$ which produce the most accurate guesses across all timepoints and objects.

This raises two immediate concerns. One is a degrees-of-freedom issue: if we have M_i coefficients, then we can only guarantee accuracy at M_i time-points. However, if we actually have deduced the physical laws obeyed by our system, this should not matter; the correct relationships will hold at all time-points and not just the ones they were determined from. If we have chosen the wrong terms z_{im}, our guess $y_i(t)$ will diverge from the target values $y_i'(t)$ at times far away from those used to deduce the coefficients.

The second problem is one of "interpretability." In theory, if we have M_i time-points, we have as many equations as variables, and we can determine our coefficients by simple linear algebra: if we define a vector $\vec{Y}_i'$ of target values such that $(\vec{Y}_i')_j = y_i'(t_j)$ and a matrix $\underline{\underline{Z_i}}$ given by $(\underline{\underline{Z_i}})_{jk} = z_{ij}(t_k)$, such that each row describes the value of a single term at each time-point, then our coefficients are straightforwardly given by solving the equation

$$(\underline{\underline{Z_i}})^T \cdot \vec{F}_i = \vec{Y}_i' \tag{16}$$

for a vector $\vec{F}_i$ whose entries are the coefficients f_{im}. However, this would involve the calculation of the matrix inverse of $(\underline{\underline{Z_i}})^T$, which is both computationally fraught and analytically problematic. There is no easy general formula for the inverse of an N-by-N matrix, and so it is all but impossible to discern how the values of our chosen terms influence our final coefficients. Once we introduce the matrix inverse into our algorithm, it becomes a black box once again; it is impossible to construct, say, a useful differential equation in a single datapoint $z_{ij}(t_k)$, if that term is incorporated into a matrix which is then inverted.

Instead we use Cramer's rule, first written down in 1,752 and of which there are many proofs widely available (including that in Brunetti, 2014). The coefficients are given by

$$f_{im} = \frac{\left|\underline{\underline{S_{im}}}\right|}{\left|\underline{\underline{Z_i}}\right|}$$

where square brackets indicate determinants and the matrix $\underline{\underline{S_{im}}}$ is defined by

$$(\underline{\underline{S_{im}}})_{jk} = \begin{cases} z_{ij}(t_k), j \neq m; \\ y_i'(t_k), j = m \end{cases} \tag{17}$$

This produces coefficients which exactly solve, for all chosen timepoints t_k (which we assume are randomly chosen from a dataset of possible observations),

$$y_i(t_k) = \sum_m f_{im} z_{im}(t_k) = y_i'(t_k) \tag{18}$$

The great benefit of this technique is that a determinant is linear in all values it involves. By avoiding the matrix inverse, we have ensured that the coefficient is differentiable in every element of data that contributes to it, and thus the effect of each piece of data on our conclusions is exactly quantifiable. This part of the algorithm is a completely "white box."

The above procedure predicts the coefficients $\{f_{im}\}$ that best describe the system when presented with a set of terms $\{z_{im}\}$; we must still develop a process for choosing between sets of terms.

The simplest and best procedure is simply to try each possible set of terms sequentially and choose the set of terms $\{z_{im}\}$ which has the lowest error according to the loss function

$$L = \sum_{i,t} \left(y_i(t) - y_i'(t)\right)^2 \tag{19}$$

where the sum is over all timepoints in the dataset, not simply the randomly-chosen timepoints used to deduce the coefficients. This is a straightforward way of determining the "goodness of fit" of our model, and has no hyperparameters, because we have eliminated the regularization term. Here, there is a much easier, more intuitive way of measuring the complexity of our system: the number of terms in our polynomial description, M_i, which we control directly. We could make our loss function L_n instead of L_2 for $n \geq 2$ and even; clearly, this would have the effect of valuing a polynomial description with a large number of small errors over one with a small number of large errors, which may be desirable or not depending on the needs of the clinician.

We must, therefore, try each set of terms sequentially, however naive an approach that may initially seem. Any attempt to navigate the space of possible terms $\{z_{im}\}$ through stochastic gradient descent using the loss function L is doomed to failure, since we cannot move in infinitesimal increments through z_{im}, but must jump between discrete sets of input data combinations, which may involve changes in value so large as to render gradients useless. Further, in order to determine the gradient of the loss function with respect to an input term z_{im}, we must also consider its effect on the entire set of deduced coefficients $\{f_{im}\}$, which will require two matrix determinant evaluations for every coefficient. At this point, the calculation of the gradient at each point becomes much more computationally expensive than simply calculating the loss for each set of terms, which is guaranteed to terminate, since the space it is exploring is finite. A brief analysis of cost, and an additional *generalizability* metric assessing the suitability of a particular description-length M_i, is included in Appendix (Section 4).

2.3. Experiment: Fitting models of tumor growth

We now investigate the advantages of this algorithm when applied to real-world cancer data. For the remainder of this paper we will be following the work of Kühleitner et al. (2019). In this paper, the authors considered longitudinal time-series data of the growth of a tumor. Human breast cancer cells were injected into nude mice, and the resulting tumor volume $v(t)$ was observed over 114 days, in a study by Worschech et al. (2009) (shown in Figure 1). Kühleitner et al. (2019) aimed to find the best parameter fit for a Bertalanffy-Pütter model from the

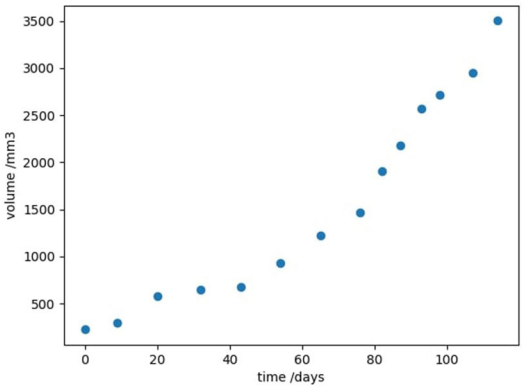

FIGURE 1
Experimental data showing the growth of tumor volume with time, in a mouse model of human breast cancer, taken from Kühleitner et al. (2019).

observed tumor data; that is to fit the non-negative parameters p, q, a, b in the first-order differential equation

$$\frac{dv}{dt} = pv^a - qv^b \tag{20}$$

The Bertalanffy-Pütter model (Ohnishi et al., 2014) is a general class of tumor-growth model which encompasses other, more specific tumor models, including the Verhulst model (Verhulst, 1838) ($a = 1.0, b = 2.0$) and the Gompertz model ($a = 1.0, b > 1.0$) (Gompertz, 1833). Per Kühleitner, it has been experimentally observed that tumors tend to shrink when they become very large; to ensure this behavior, only exponent-pairs $a < b$ are considered. They were examined at intervals of 0.01, so that ($a = 0.01n, b = a + 0.01m$) for all valid non-negative integers n, m that placed (a, b) within the highlighted range. For every exponent-pair, the authors fitted the best coefficient-pair (p, q) through a painstaking process of stochastic gradient descent and simulation (simulated annealing), using the same L_2 loss function (2), otherwise known as the sum of squared error (SSE), defined in our algorithm. Having chosen a trial pair (p, q), they solve the equation numerically over 144 days, sum the square of the errors, make a partially-stochastic adjustment to (p, q), and simulate again. Their final best fit was ($p = 5 \cdot 10^{-4}, q = 5.6 \cdot 10^{-7}, a = 1.62, b = 2.44$), obtained at a cost of roughly 1 week of CPU time. Our objective is to repeat this study by applying our algorithm to fit coefficients of the Bertalannfy-Pfutter model to this data using SSE as our loss function. We make these choices for ease of comparison, but the algorithm could in theory work with any differential-equation model and any loss function. If we were to use a stochastic differential equation (SDE), for example, we could generate a maximum likelihood function for a model defined by a given set of parameters, which would allow us to use

likelihood-dependent loss functions, such as the Akaike and Bayesian Information Criteria.

3. Results

3.1. Identifying regions of good fit with real-world data

We have a single output guess, $y_i'(t) = y'(t) = \frac{dv}{dt}$, obtained using numpy.gradient's (Cranmer et al., 2020) first-order approximations at each timepoint instead of by precise and repeated simulation; we have a single input representation, $x_i(t) = x(t) = v(t)$, the observed tumor volume. Because we are fitting to a *known model* here instead of unknown dynamics, we do not need to involve the generalizability metric or decide between numbers of terms; instead we can simply try each (a, b) pair sequentially, deduce our coefficients (p, q) using Cramer's rule, and output an error L using the sum of the squares of the errors of the gradient at each timepoint according to that prediction. As we are only deducing two coefficients, we choose two timepoints at random; to make sure our predictions are an accurate reflection of the entire dataset, we repeat the procedure above 20 times for each (a, b) pair (to ensure that each datapoint has a 95% chance of being selected at least once), and choose the deduced coefficient pair (p, q) with the lowest error. We consider all exponent-pairs at 0.01 intervals where $a < b \leq 3.0$, the highest value considered by Kühleitner et al. (2019). Our algorithm runs very quickly on a standard laptop (requiring just under seven minutes to terminate), and efficiently explores the space of possible parameters for the roughly 45,000 possible exponent pairs, returning the accuracy surface. Because we only have two coefficients to fit per exponent pair, this surface can be visualized in three dimensions (see Figure 2); this is an advantage of the Bertalanffy-Pütter model.

Because our target values are imprecise approximations to the true growth rate, the algorithm cannot perfectly identify the actual accuracy minimum. However, this surface shows us intuitively how the model behaves in various regions of the (a, b) space. We can see, for example, that the model behaves asymptotically badly as the exponents increase past 2.5, and that no effort should be expended trying to identify (p, q) pairs there. We can also see a "valley" of low error in the center, which might be understood as a "region of good fit," where exponent pairs generally describe the system well. We can also use this algorithm to identify regions of overfit, by plotting the best values of p and q obtained at each point in (a, b) space (see Figure 3).

We see that all regions where $a, b < 1.0$ should be ignored, as the coefficients "hit a wall" as soon as that threshold is passed: they become rapidly unstable (and, in the case of q, unphysically negative) with respect to small changes in exponent pairs, which suggests that region provides a poor model of the system, since any good mathematical model of a biological system should not be so acutely sensitive to small changes in its terms. This allows us to narrow down the promising region of (a, b) in space to the section of the valley where $a, b > 1.0$, and we can explore that region further using precise simulation to identify the best coefficient-pair (p, q). Further, we have a good idea of where those coefficients should lie: for the authors' final best exponent pair $(a = 1.62, b = 2.44)$ we obtain $(p = 3 \cdot 10^{-4}, q = 3 \cdot 10^{-7})$ to their $(p = 5 \cdot 10^{-4}, q = 5.6 \cdot 10^{-7})$, which is remarkably close given that their gradients are derived from careful simulation and ours from crude first-order approximation. We have narrowed down the space of possible hyperparameters by several orders of magnitude in a matter of minutes; what remains can then be explored more precisely.

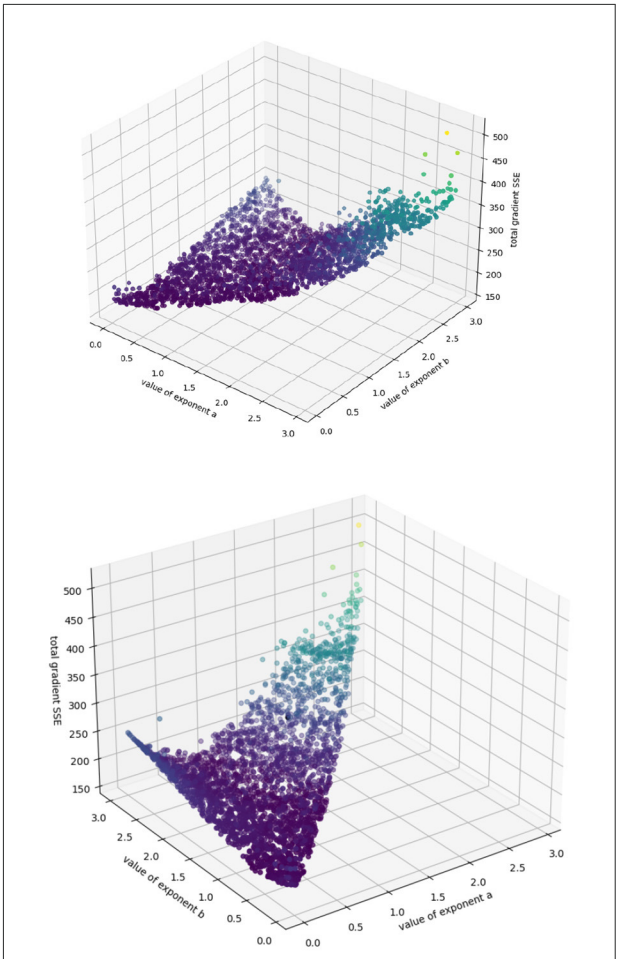

FIGURE 2
Sum of squared error from extrapolation from fitted (p, q) values for each exponent-pair value; color simply corresponds to height for highlighting purposes.

3.2. Recovering parameters from synthetic data

We can test the algorithm's accuracy further by using this surface to identify trial parameters, generate synthetic data using

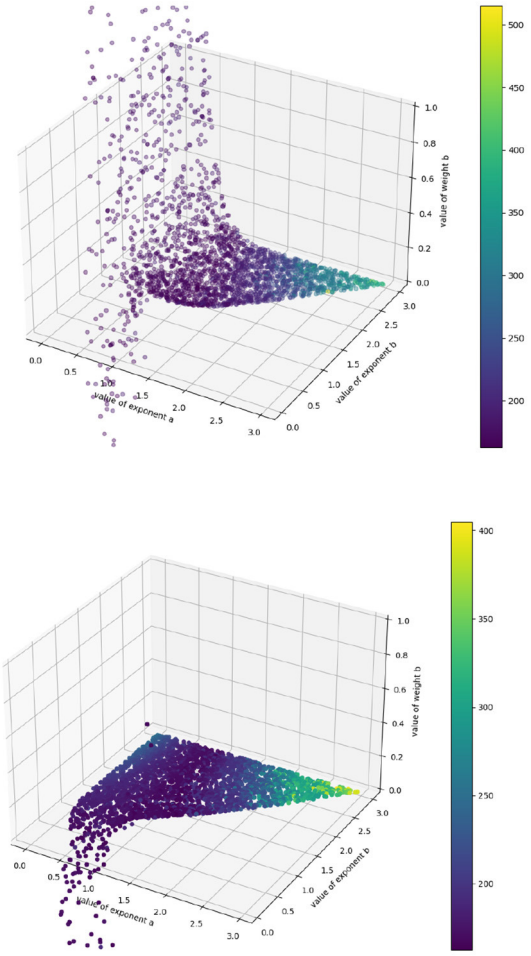

FIGURE 3
Fitted p (above) and q (below) values for each exponent-pair value. The colorbar corresponds to SSE—purple datapoints have lower error, yellow ones higher. The varying limits come from the fact that, to generate each plot, we randomly choose 2,500 points out of 45,000 to display.

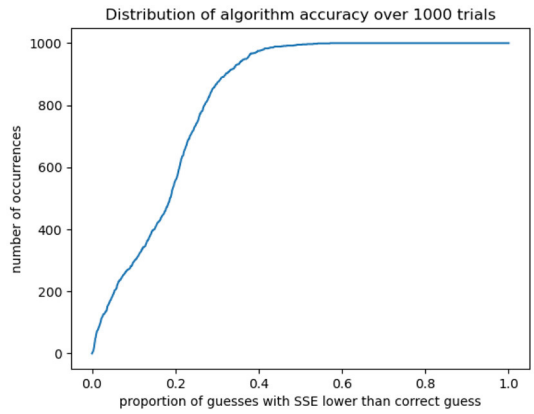

FIGURE 4
For 1,000 synthetically generated test cases, we calculate whether the correct exponent-pair occurs within the lowest-SSE "x" percent of the space. In most cases the algorithm can isolate roughly two-fifths of the original space which may then be explored in more detail for a closer-fitting solution.

those parameters, and using the algorithm to retrieve them. We assume that every set of (a, b, p, q) parameters with SSE smaller than that of the "official" Kühleitner solution is biologically realistic, as it fits the tumor growth trajectory at least as closely. We limit ourselves to the region $a, b > 1.0$ and obtain about 5,000 possible sets of parameters, from which we select 1,000 at random. Using the initial tumor volume as our starting point, for every chosen (a, b, p, q) we extrapolate forward according to equation (20).

We then take the tumor volumes at the same timepoints as the original data, to mimic its sparsity. We generate an accuracy surface for each trajectory according to the procedure above (This process took roughly 36 h using the University College London DPS machines). For each "synthetic tumor," we denote the exponent-pair used to generate it as (a_*, b_*), and calculate the fraction of the parameter space $1.0 <= a, b =< 3.0$ with an assigned SSE lower than that calculated for (a_*, b_*). This gives us a neat metric for the degree to which the algorithm "narrows down" the parameter space, depending on how confident the modeler wishes to be that the "correct" parameter values—insofar as any biological system can be said to have a single correct set of underlying parameters—lies within the identified region. Our results are shown in Figure 4. For 999 out of 1,000 trajectories, (a_*, b_*) has an SSE higher than 57% of the parameter space; for 990 trajectories, we can narrow down to 46% of the space, for 950, to 37%; for 900, to 32%; and for 800 to 27%. We see a "threshold effect," demonstrated below: in the vast majority of cases the space can be narrowed down to roughly two-fifths of its original area.

3.3. The effect of noise on algorithmic efficacy

We can also explore the effect of noise on this accuracy, by separating our 1,000 trajectories into five groups of 200 and injecting random noise at each timepoint. For a noise level of 0.01, for example, at each timepoint a random fraction of the tumor volume between 1 and −1% is drawn from a normal distribution and added to the tumor volume. Gradients are then computed and the algorithm is run as previously; we again calculate the proportion of the parameter space with an SSE lower than that assigned to the correct exponents (a_*, b_*). Our results are shown in Figure 5. We see that the "thresholding" effect, by which the correct parameters can be narrowed down

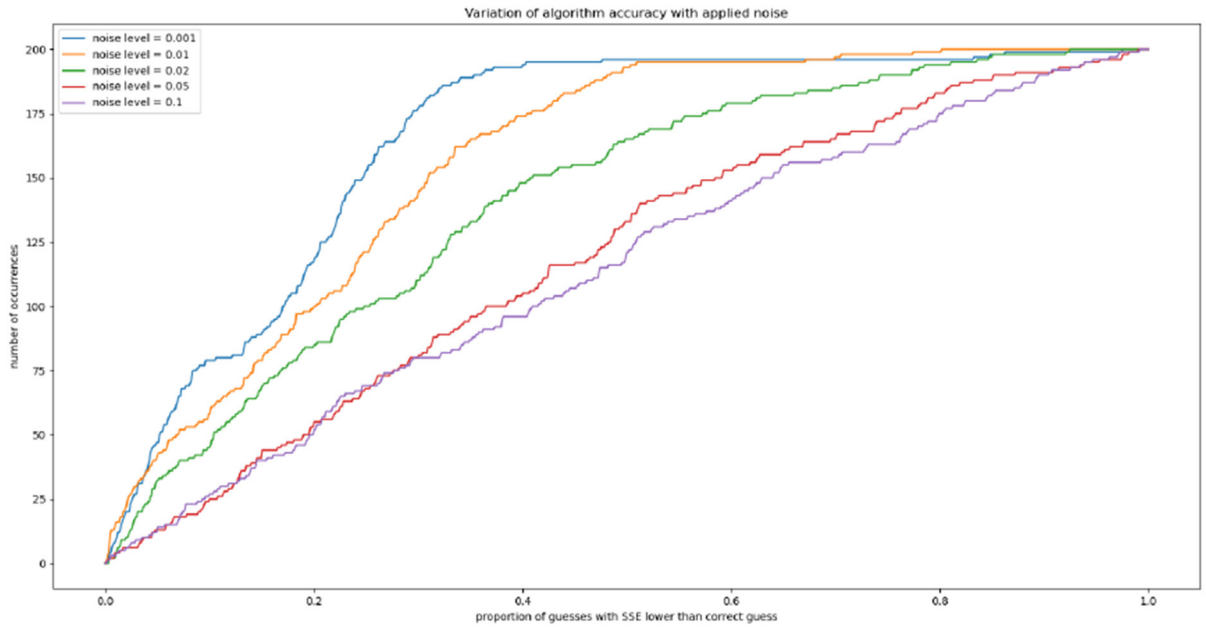

FIGURE 5
The effect of randomly-generated noise on algorithm accuracy.

to a certain proportion of the space with near-certainty, holds up to a noise level of roughly 0.02.

4. Discussion

By attempting to build an algorithm that can interpretably explain the unknown dynamics of an interacting system, we have found an approach that can quickly and easily explore the space of parameters of a differential equation which incorporates a variety of models of tumor growth. On synthetic tumor-growth data, the algorithm can reliably (with a probability of 95%) more than halve the region of parameter space that requires finer searching using less rigorous, more computationally expensive machine learning methods. There is good reason to think the algorithm can be usefully applied to more general models of cancer growth, so long as there are enough datapoints that the compromise of first-order gradient estimation can be safely made. In fact, above approach does not require the underlying equation to be first-order, or indeed to be a differential equation at all; it works for any form, any number of terms, and any number of objects. It provides a first-approximation to the behavior of the system, without the expense of simulation, and it does so without nonlinearity or the use of hyperparameters. It can therefore be applied to a variety of contexts, medical and otherwise.

An important aspect of the above procedure, at least as it applies to cancer modeling, is that it identifies not simply one good fit to the equation—as stochastic gradient descent does—but instead identifies several thousand candidate equations and ranks them by "goodness of fit." This is particularly useful to us because a tumor is not a purely deterministic or mathematical object: it does not obey a single equation for all time, and its behavior is likely best modeled as a combination of, or a movement through, the candidate equations suggested by the algorithm. The ability to *narrow down* the space of model parameters to describe a particular tumor—perhaps successively, through more and more granular exploration—will be of use to clinicians trying to classify and predict the behavior of cancers. Even leaving aside explainability considerations, our algorithm can more than halve the space which must be explored to fit parameters to the tumor using stochastic gradient descent, which is a vital efficiency gain when trying to provide personalized predictions at scale. There are a wide range of complex interacting-differential-equation models of cancer growth to which this algorithm might usefully be applied (for instance, Nave, 2020; Hori et al., 2021; Mascheroni et al., 2021; Nave and Elbaz, 2021), although the algorithm could, again, in principle be used to describe any dynamical system.

In addition to this, across patients, the accuracy surface may provide a useful tool for characterizing particular kinds of cancer, or the effects of certain treatments. It may be that further study reveals that there is a link between the best regions of (a, b) space to describe a tumor and some aspect of its growth or behavior. The ability to associate a set of best-fit (p, q, a, b) parameters to a particular tumor also suggests the possibility of new set of survival metrics, which may correlate directly the prognosis of human patients. This merits further

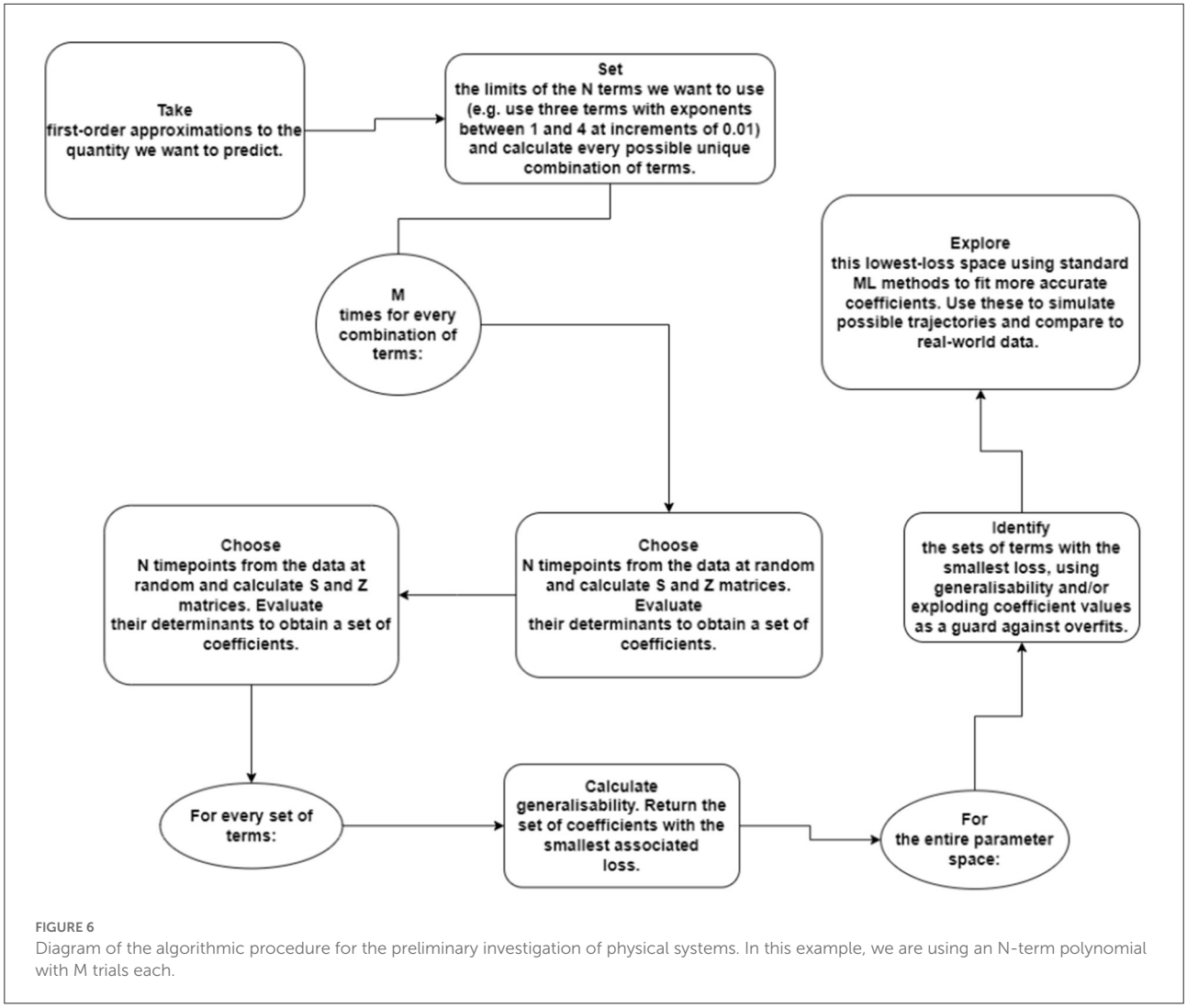

FIGURE 6
Diagram of the algorithmic procedure for the preliminary investigation of physical systems. In this example, we are using an N-term polynomial with M trials each.

investigation. A full diagram of the procedure is included in Figure 6.

A technical aspect of the algorithm worth drawing attention to is its susceptibility to underflow errors, which arises from its calculation of the ratio of two determinants. This is not an issue in any of the cases discussed above, but rapidly compromises any current attempt to apply the algorithm to large systems or to use many terms. If we have M terms in our description, for example, each of the order 10^{-n}, then the coefficients will be ratios of two numbers of order 10^{-nM}. Given that standard Python floating-point precision cannot accurately represent numbers smaller than about 10^{-39} (Rajaraman, 2016), neither n nor M have to become very large before we run into accuracy issues. Further work could implement the algorithm using an arbitrary-precision arithmetic program designed specifically to compute matrix determinants, such as Arb (Johansson, 2017). The algorithm also requires its input data to be sufficiently detailed that the compromise of first-order gradient approximation is worth making. On datasets such as that attached to Laleh et al. (2022), where most trajectories are composed of six or fewer datapoints, attempts to fit exponents result in flat, highly noisy surfaces with no significant curvature. Mouse or *in vitro* models, which can be monitored more or less continuously without the need for painful and invasive scans on human subjects, are our likeliest sources of useful data. However, as scanning methods become more advanced over the next decade (Rockne et al., 2019)—less invasive, less painful, and cheaper to perform regularly on human patients—tumor-volume trajectories will become denser and more amenable to mathematical analysis, and the context in which this algorithm is useful will move from the experimental to the clinical.

5. Conclusion

This paper describes an interpretable method for quickly surveying the parameter space of various differential-equation models. It is precisely the complexity and nonlinearity of

neural networks which make them so useful in problems of classification or recognition, but when human lives are at stake, it is important to develop methods of generating predictions and informing treatments that are built around explainability and *a priori* justification. Clinicians and patients must understand as much as possible where their information is coming from, and mathematical models derived from computational methods must be rigorous. Moreover, as our work on Kühleitner et al. (2019) shows, it is not even clear that immediately resorting to machine learning makes anything *faster*. Slow brute-force adjustment is an inefficient approach when a straightforward algorithm can narrow down the space of possible parameters, and suggest thousands of candidate equations, in a matter of minutes. In addition to the detailed machine learning work currently being done in the field of mathematical oncology (see for instance Bekisz and Geris, 2020), a different approach is needed—the unification of mathematics and machine learning to create a rigorous, explainable justification for the directions in which neural networks should be sent. We suggest the use of this first-order "exploration algorithm" as a first line of defense when modeling the behavior of cancer, to provide an initial understanding of the behavior of a model across its parameter space and significantly reduce the time taken to fit predictive equations. A return to first principles in cancer modeling may yield significant optimization.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s. All code used in the production of these results is available on request.

Author contributions

HC developed proofs and experiments and drafted the paper under the close supervision of PL and with the advice of HA, both of whom also edited the paper. All authors contributed to the article and approved the submitted version.

Funding

HC was supported by a grant from the Engineering and Physical Sciences Research Council, reference EP/W523835/1. The University College London Mathematics Department DPS machines were used to conduct some of the computational experiments in this paper. PL acknowledges funding from HORIZON-EIC (project number 101058004): Chemometric histopathology via coherent Raman imaging for precision medicine.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.941451/full#supplementary-material

References

Bekisz, S., and Geris, L. (2020). Cancer modeling: from mechanistic to data-driven approaches, and from fundamental insights to clinical applications. *J. Comput. Sci.* 46:101198. doi: 10.1016/j.jocs.2020.101198

Benzekry, S. (2020). Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clin. Pharmacol. Therap.* 108, 471–486. doi: 10.1002/cpt.1951

Brunetti, M. (2014). Old and new proofs of cramer's rule. *Appl. Math. Sci.* 8, 6689–6697. doi: 10.12988/ams.2014.49683

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., et al. (2020). "Discovering symbolic models from deep learning with inductive biases," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates), 17429–17442.

Dujon, A. M., Aktipis, A., Alix-Panabières, C., Amend, S. R., Boddy, A. M., Brown, J. S., et al. (2021). Identifying key questions in the ecology and evolution of cancer. *Evol. Appl.* 14, 877–892. doi: 10.1111/eva.13190

Ghazani, M. A., Saghafian, M., Jalali, P., and Soltani, M. (2021). Mathematical simulation and prediction of tumor volume using rbf artificial neural network at different circumstances in the tumor microenvironment. *Proc. Instit. Mech. Eng. Part H* 235, 1335–1355. doi: 10.1177/09544119211028380

Gompertz, B. (1833). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life

contingencies. *Abstracts Pap. Printed Philos. Trans. R. Soc. Lond.* 2, 252–253. doi: 10.1098/rspl.1815.0271

Hori, S. S., Tong, L., Swaminathan, S., Liebersbach, M., Wang, J., Gambhir, S. S., et al. (2021). A mathematical model of tumor regression and recurrence after therapeutic oncogene inactivation. *Sci. Rep.* 11:1341. doi: 10.1038/s41598-020-78947-2

Hsu, C. -H., Manogaran, G., Panchatcharam, P., and Vivekanandan, S. (2018). "A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classifiers," in *2018 IEEE 8th International Symposium on Cloud and Service Computing*. p. 111–115. doi: 10.1109/SC2.2018.00023

Johansson, F. (2017). ARB: efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Trans. Comput.* 66, 1281–1292. doi: 10.1109/TC.2017.2690633

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nat. Rev. Phys.* 3, 422–440. doi: 10.1038/s42254-021-00314-5

Kazhdan, D., Dimanov, B., Jamnik, M., and Liò, P. (2020). Meme: generating RNN model explanations *via* model extraction. *arXiv [Preprint]*. arXiv: 2012.06954. Available online at: https://arxiv.org/pdf/2012.06954.pdf

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv [Preprint]*. arXiv: 1412.6980. Available online at: https://arxiv.org/pdf/1412.6980.pdf

Kühleitner, M., Brunner, N., Nowak, W.-G., Renner-Martin, K., and Scheicher, K. (2019). Best fitting tumor growth models of the von bertalanffy-püttertype. *BMC Cancer* 19:683. doi: 10.1186/s12885-019-5911-y

Kumar, Y., Gupta, S., Singla, R., and Hu, Y.-C. (2022). A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch. Comput. Methods Eng.* 29, 2043–2070. doi: 10.1007/s11831-021-09648-w

Kurz, D., Sànchez, C. S., and Axenie, C. (2021). Data-driven discovery of mathematical and physical relations in oncology data using human-understandable machine learning. *Front. Artif. Intell.* 4:713690. doi: 10.3389/frai.2021.713690

Kwak, M. S., Lee, H. H., Yang, J. M., Cha, J. M., Jeon, J. W., Yoon, J. Y., et al. (2021). Deep convolutional neural network-based lymph node metastasis prediction for colon cancer using histopathological images. *Front. Oncol.* 10:619803. doi: 10.3389/fonc.2020.619803

Laleh, N. G., Loeffler, C. M. L., Grajek, J., Staňková, K., Pearson, A. T., Muti, H. S., et al. (2022). Classical mathematical models for prediction of response to chemotherapy and immunotherapy. *PLoS Comput. Biol.* 18:e1009822. doi: 10.1371/journal.pcbi.1009822

Li, Y., Mark, B., Raskutti, G., Willett, R., Song, H., and Neiman, D. (2020). Graph-based regularization for regression problems with alignment and highly correlated designs. *SIAM J. Math. Data Sci.* 2, 480–504. doi: 10.1137/19M1287365

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Magister, L. C., Kazhdan, D., Singh, V., and Liò, P. (2021). GCexplainer: human-in-the-loop concept-based explanations for graph neural networks. *arXiv [Preprint]*. arXiv: 2107.11889. Available online at: https://arxiv.org/pdf/2107.11889.pdf

Mascheroni, P., Savvopoulos, S., Alfonso, J. C. L., Meyer-Hermann, M., and Hatzikirou, H. (2021). Improving personalized tumor growth predictions using a Bayesian combination of mechanistic modeling and machine learning. *Commun. Med.* 1:19. doi: 10.1038/s43856-021-00020-4

Nave, O. (2020). Adding features from the mathematical model of breast cancer to predict the tumour size. *Int. J. Comput. Math.* 5, 159–174. doi: 10.1080/23799927.2020.1792552

Nave, O., and Elbaz, M. (2021). Artificial immune system features added to breast cancer clinical data for machine learning (ML) applications. *Biosystems* 202:104341. doi: 10.1016/j.biosystems.2020.104341

Ohnishi, S., Yamakawa, T., and Akamine, T. (2014). On the analytical solution for the pütter-bertalanffy growth equation. *J. Theor. Biol.* 343, 174–177. doi: 10.1016/j.jtbi.2013.10.017

Rajaraman, V. (2016). IEEE standard for floating point numbers. *Resonance*. 21, 11–30.

Rockne, R. C., Hawkins-Daarud, A., Swanson, K. R., Sluka, J. P., Glazier, J. A., Macklin, P., et al. (2019). The 2019 mathematical oncology roadmap. *Phys. Biol.* 16:041005. doi: 10.1088/1478-3975/ab1a09

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Sanoob, M. U., Madhu, A., Ajesh, K. R., and Varghese, S. M. (2016). Artificial neural network for diagnosis of pancreatic cancer. *Int. J. Cybern. Inform.* 5, 41–49. doi: 10.5121/ijci.2016.5205

Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.* 10, 113–121.

Worschech, A., Chen, N., Yu, Y. A., Zhang, Q., Pos, Z., Weibel, S., et al. (2009). Systemic treatment of xenografts with vaccinia virus GLV-1h68 reveals the immunologic facet of oncolytic therapy. *BMC Genomics*. 10, 301. doi: 10.1186/1471-2164-10-301

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv [Preprint]*. arXiv: 1505.00855. Available online at: https://arxiv.org/pdf/1505.00853.pdf

Yuan, Y., Wang, W., and Pang, W. (2021). "A systematic comparison study on hyperparameter optimisation of graph neural networks for molecular property prediction," in *GECCO '21: Proceedings of the Genetic and Evolutionary Computation Conference*, 386–394. doi: 10.1145/3449639.3459370

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fmed.2022.945698

OPEN ACCESS

EDITED BY
Jingjing You,
The University of Sydney, Australia

REVIEWED BY
Kurt Svärdsudd,
Uppsala University, Sweden
Andrés Bueno-Crespo,
Catholic University San Antonio of Murcia, Spain

*CORRESPONDENCE
Guillermo Droppelmann
guillermo.droppelmann@meds.cl

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 16 May 2022
ACCEPTED 29 August 2022
PUBLISHED 23 September 2022

CITATION
Droppelmann G, Tello M, García N, Greene C, Jorquera C and Feijoo F (2022) Lateral elbow tendinopathy and artificial intelligence: Binary and multilabel findings detection using machine learning algorithms. *Front. Med.* 9:945698. doi: 10.3389/fmed.2022.945698

Lateral elbow tendinopathy and artificial intelligence: Binary and multilabel findings detection using machine learning algorithms

Guillermo Droppelmann[1,2,3]*, Manuel Tello[4], Nicolás García[5], Cristóbal Greene[6], Carlos Jorquera[7] and Felipe Feijoo[4]

[1]Research Center on Medicine, Exercise, Sport and Health, MEDS Clinic, Santiago, RM, Chile, [2]Health Sciences Ph.D. Program, Universidad Católica de Murcia UCAM, Murcia, Spain, [3]Principles and Practice of Clinical Research (PPCR), Harvard T.H. Chan School of Public Health, Boston, MA, United States, [4]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, [5]MSK Diagnostic and Interventional Radiology Department, MEDS Clinic, Santiago, RM, Chile, [6]Hand and Elbow Unit, Department of Orthopaedic Surgery, MEDS Clinic, Santiago, RM, Chile, [7]Facultad de Ciencias, Escuela de Nutrición y Dietética, Universidad Mayor, Santiago, RM, Chile

Background: Ultrasound (US) is a valuable technique to detect degenerative findings and intrasubstance tears in lateral elbow tendinopathy (LET). Machine learning methods allow supporting this radiological diagnosis.

Aim: To assess multilabel classification models using machine learning models to detect degenerative findings and intrasubstance tears in US images with LET diagnosis.

Materials and methods: A retrospective study was performed. US images and medical records from patients with LET diagnosis from January 1st, 2017, to December 30th, 2018, were selected. Datasets were built for training and testing models. For image analysis, features extraction, texture characteristics, intensity distribution, pixel-pixel co-occurrence patterns, and scales granularity were implemented. Six different supervised learning models were implemented for binary and multilabel classification. All models were trained to classify four tendon findings (hypoechogenicity, neovascularity, enthesopathy, and intrasubstance tear). Accuracy indicators and their confidence intervals (*CI*) were obtained for all models following a K-fold-repeated-cross-validation method. To measure multilabel prediction, multilabel accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) with 95% *CI* were used.

Results: A total of 30,007 US images (4,324 exams, 2,917 patients) were included in the analysis. The RF model presented the highest mean values in the area under the curve (AUC), sensitivity, and also specificity by each degenerative finding in the binary classification. The AUC and sensitivity showed the best performance in intrasubstance tear with 0.991 [95% *CI*, 099, 0.99], and 0.775 [95% *CI*, 0.77, 0.77], respectively. Instead, specificity showed

upper values in hypoechogenicity with 0.821 [95% *CI*, 0.82, −0.82]. In the multilabel classifier, RF also presented the highest performance. The accuracy was 0.772 [95% *CI*, 0.771, 0.773], a great macro of 0.948 [95% *CI*, 0.94, 0.94], and a micro of 0.962 [95% *CI*, 0.96, 0.96] AUC scores were detected. Diagnostic accuracy, sensitivity, and specificity with 95% *CI* were calculated.

Conclusion: Machine learning algorithms based on US images with LET presented high diagnosis accuracy. Mainly the random forest model shows the best performance in binary and multilabel classifiers, particularly for intrasubstance tears.

KEYWORDS

AUC curve, diagnosis, random forest, tennis elbow, ultrasound

Introduction

Lateral elbow tendinopathy (LET) (1), also known as tennis elbow (2), is one of the most frequent musculoskeletal disorders (3). The common extensor tendon, specifically the extensor carpi radialis brevis, is directly involved in the development of this condition (4). LET is a potentially debilitating condition causing significant pain and disability for periods of 12 months or more (5), and in some cases, also generates disruptive sleep (6). This condition is estimated to affect 3.3–3.5 per 1,000 by year (7), affecting individuals during their most productive period (8) and increasing in tennis players with a prevalence of over 40–50% (9). Effective treatment for this tendinopathy is uncertain, with controversial scientific evidence that provides more than 40 modalities (10) in 200 clinical trials and several systematic reviews (11).

Although LET remains primarily a clinical diagnosis (12), the ultrasound (US) findings in common extensor tendon have been well documented in asymptomatic persons (13–17) and LET individuals with tendon structural changes (18–22). However, the degree of these tendon structural changes is highly diverse, with different levels of accuracy (19, 23), making the interpretation of the US imaging a real radiological challenge. For example, a met analysis reported that the US sensitivity and specificity in the detection of common extensor tendon ranged between 64 and 100% and 36 and 100%, respectively (24). Furthermore, this high variability can increase even more if different types of degenerative findings are considered, such as hypoechogenicity, bone changes, neovascularity, calcifications, cortical irregularities (25), and tear (thickness) (26), increasing the lack of precision in the diagnosis by US images. To date, there is still no consensus about what parameters should be considered for the evaluation of changes in the tendon matrix (27).

Recently, artificial intelligence has shown the potential to revolutionize the accuracy of diagnosis by developing a series of classification models (28) and by reducing medical diagnosis variability (29–31). The algorithms based on machine learning and convolutional neural network have been successfully used in pattern recognition in different clinical contexts and specialties, such as neurology (32–34), pulmonary (35–37), cardiovascular (38–42), and oncology (43–51), improving diagnosis accuracy, weighted errors, false-positive rate, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) (52). In radiology, machine learning and convolutional neural network algorithms have been used to detect and classify injury patterns in fractures, cartilage defects, meniscal and anterior cruciate ligament tears, and spinal metastases (53, 54) with excellent performance indices.

Most of the studies mentioned above have used computed tomography scan, magnetic resonance imaging, and X-rays as an image-generating source. For example, fracture detection using a computed tomography scan has been used by Tomita et al. (55) with deep neural networks for automatic detection of osteoporotic vertebral fractures, obtaining an accuracy of 89.2%. Another author (56) that also studied automated detection of posterior-element fractures with deep convolutional networks obtained an AUC of 85.7%. There is also some experience using automatic classification and detection of calcaneus fracture with an accuracy of 98% (57). Couteaux et al. (58), Bien et al. (59), and Roblot et al. (60) developed algorithms to automatically detect knee meniscal tears using convolutional neural networks and deep learning assisted with magnetic resonance imaging, obtaining AUC scores of 90.6, 84.7, and 92%, respectively. A similar performance was obtained by authors in (61), where cartilage lesion detection algorithms were developed, reaching accuracy levels of 91%. In radiography, different applications are considered, such as deep learning classification algorithms for the detection of ossification areas of the hand to estimate skeletal maturity (62), obtaining accuracy results similar to an expert radiologist (63). Another publication evaluated knee osteoarthritis in 3,000 subjects (5,960 knees) from the Osteoarthritis Initiative dataset using deep learning techniques. They achieved an AUC of 93%, although the

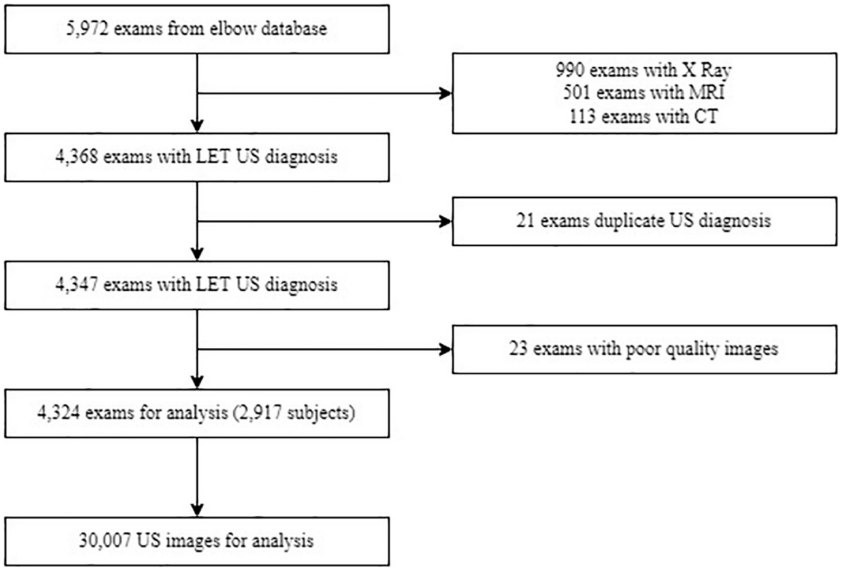

FIGURE 1
Flowchart of data selection and subjects used in the study. Abbreviations: MRI, magnetic resonance imaging; CT, computed tomography scan; LET, lateral elbow tendinopathy; US, ultrasound.

diagnosis is highly dependent on the practitioner's subjectivity, just like US methods (64). As noted earlier, however, US imaging has not been frequently used as an image-generating source.

Machine learning for the medical US continues to be an opportunity (65), especially in musculoskeletal disorders since the US is highly operator-dependent (66) and the applications are dictated by adequate front-end beamforming, compression, signal extraction, and velocity (67), requiring significant training to acquire a level of competence in clinical diagnosis (68) because the images contain multiplicative noise (69). Baka et al. (70) proposed a model to learn the appearance of the bone interface using US images and random forest methods, obtaining a precision of 86%. Another group proposed an algorithm to segment vertebral US images into three regions with a classification rate of 84.7% (71). In tendon, literature is uncommon yet. In 2017, the University of Salford from the United Kingdom reported in an international conference an automatic method to detect and classify Achilles tendon injuries using decision trees, non-linear support vector machines, and ensemble classifiers (69). Kapinski in 2018 (72) reported a novel method for continuous evaluation of reconstructed Achilles tendon healing based on the responses of intermediate convolutional neural network layers. Note that the task of detecting and classifying different conditions as described above can be considered simple since they are based on binary results (an anomaly can only be present or not) (54). This study differs from others that use deep learning or convolutional neural networks because it uses a multilabel, fast, and simplified classifier to find different degenerative patterns simultaneously, such as hypoechogenicity, neovascularity, bony irregularities, and fibrillar disruptions. Currently, no scientific publications have identified ultrasonographic findings using artificial intelligence algorithms.

This article aims to assess multilabel classification models using machine learning algorithms to detect degenerative findings and intrasubstance tear in US images with LET diagnosis.

Materials and methods

Study design

This study was designed as a retrospective and multicentric study. It was written following the Strengthening the Reporting of Observation studies in Epidemiology (STROBE) guideline (73). All patients records with an elbow US exam at MEDS Clinic in Santiago, Región Metropolitana, Chile. This study started on March 1st, 2019.

Subjects

Only images of the common extensor tendon were considered. We selected US images and medical records from patients with a LET diagnosis from January 1st, 2017, to December 30th, 2018. The inclusion criteria were: (1) clinical diagnosis of LET established by orthopedists, sports

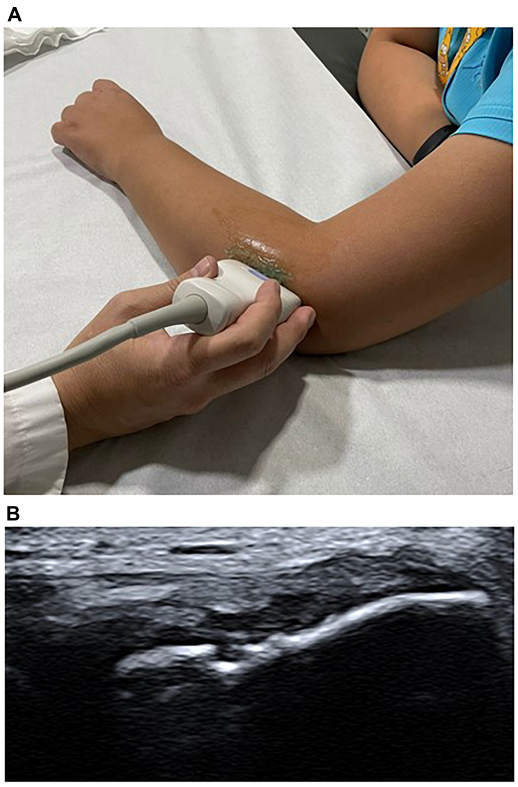

FIGURE 2
Patient evaluation position and an ultrasound (US) finding, respectively. **(A)** Probe positioning in the elbow in the US exploration of the extensor tendon complex. **(B)** US imaging shows intrasubstance tear in extensor tendon complex.

medicine physicians, or any musculoskeletal specialists, (2) US exam made in the medical center of interest, (3) US exam reported by any musculoskeletal radiologist with more than 10 years of experience, and (4) no race or age restriction. Consecutively, exclusion criteria were: (1) US-guided procedures, such as corticoid, stem cell, and platelet-rich plasma injections, (2) previous LET surgery, and (3) duplicate or not distinguishable images, were removed from the dataset. **Figure 1** provides the flowchart to select the subjects.

Ultrasound assessment of common extensor tendon

All common extensor tendons were assessed using an Aplio 500 US system (Toshiba America Medical Systems, Inc, Tustin, CA, USA) equipped with a multifrequency linear transducer was used. A frequency of 18 MHz was chosen. The images were stored as Digital Imaging and Communications in Medicine (DICOM) files and reviewed on a picture archiving and communication system (PACS).

All patients with LET diagnosis were examined in a seated position with flexion elbow in 90 grades with the wrist pronated, and the arm was resting on a table (14).

Greyscale and color Doppler US imaging are standard methods used for assessing tendon structural changes (74). Following the literature recommendations, four common prevalent degenerative findings were selected from US exams, such as hypoechogenicity, neovascularity, enthesopathy, and intrasubstance tear (75). A focal hypoechoic region was defined as being rounded and not associated with tendon disruption. Neovascularity was assessed as the presence of blood flow on color Doppler. Enthesopathy was evaluated as bony abnormalities at the tendon insertion. A linear intrasubstance tear was defined as a linear hypoechoic focus associated with discontinuity of tendon fibers (76–80). Every finding was evaluated with a binary score as present or absent. We recorded when an exam presents more than one degenerative finding. **Figure 2A** shows the evaluation position, and **Figure 2B** represents US finding, in this case, an intrasubstance tear.

Datasets: Ultrasound image and database

Several recommendations were followed for data (images) pre-processing, object detection, and feature extraction (81–83). Two datasets (A and B) were built for training and testing models. The pre-processing step considers eliminating any elements that generated noise in the images, such as uneven lighting, different sizes, or image portions without information (84). Object detection is a specific injury area of interest for the analysis. However, in this case, we considered the common extensor tendon image. Feature extraction is an important step in the construction of any pattern classification and aims at the extraction of the relevant information that characterizes each class (85). According to the 7th International Conference on System Engineering and Technology 2017, texture analysis and classification in US medical images can use feature extraction and texture characteristics for determining echo pattern characteristics (86). One of the most used are intensities distribution (mean intensity and standard deviation), pixel-pixel co-occurrence patterns, and scales granularity. Then the shape contour was extracted where the texture of the pixels was quantified. The US images were labeled manually with four degenerative findings classification outputs findings (hypoechogenicity, neovascularity, enthesopathy, and intrasubstance tear) (65) and complementary patient data such as sex, age, and side of the injury (right or left). The final process consists of a combination between the patient's information and image analysis. Dataset A was image prediction and contained data extraction from 95 morphology characteristics, shapes, and texture variables, where one image corresponding to one diagnostic (30.007 rows). Dataset B was the patient prediction

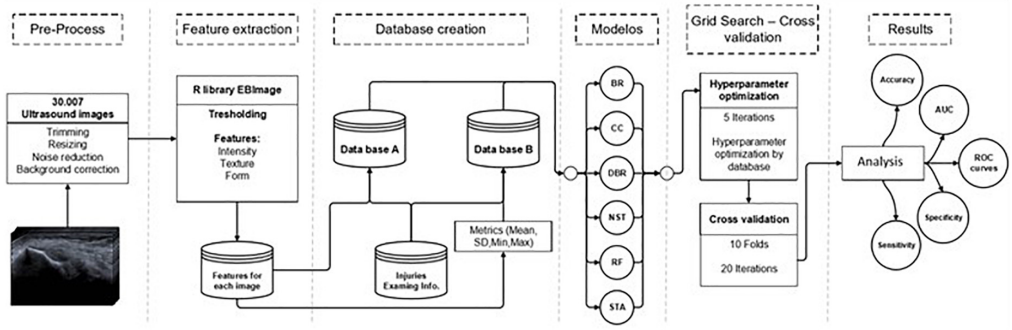

FIGURE 3
Study workflow. Abbreviations: BR, binary relevance model; CC, classifier chains model; DBR, dependent binary relevance model; NST, nested stacking model; RF, random forest; STA, staking generalization; AUC, area under the curve.

TABLE 1 Ultrasound findings comparison between sexes.

Demographic characteristics/ degenerative findings	Female ($N = 1717$) Mean ± SD; *n* (%)	Male ($N = 2607$) Mean ± SD; *n* (%)	*p*-value	Total ($N = 4324$) Mean ± SD; *n* (%)
Age	47.18 ± 11.00	45.99 ± 11.03	<0.001[a]	46.46 ± 11.03
Right side of the injury	1179 (68.88)	1790 (68.66)	0.98	2969 (68.66)
HE	1201 (69.94)	1730 (66.35)	0.0119[b]	2931 (67.75)
NV	636 (37.04)	999 (38.31)	0.4093	1635 (37.79)
E	599 (34.88)	915 (35.09)	0.9411	1514 (35.00)
IST	582 (33.89)	880 (33.75)	0.9521	1462 (33.80)

HE, hypoechogenicity; NV, neovascularity; E, enthesopathy; IST, intrasubstance tear. [a]p-value < 0.001. [b]p-value < 0.01.

and included 380 variables from data extraction, such as median, standard deviation, minimal, and maximal, where one exam corresponds to one diagnostic (4.321 rows). Figure 3 represents the study workflow process.

Machine learning and statistical analysis

Supervised learning was used because most machine learning applications for US involve them. Both datasets were implemented into binary and multilabel classification algorithms in six machine learning methods: Binary relevance model, classifier chains model, nested stacking model, dependent binary relevance model, staking generalization, and random forest.

All models were trained to classify four tendon findings (hypoechogenicity, neovascularity, enthesopathy, and intrasubstance tear) in images with LET diagnosis. First, each pattern was recognized individually and then the four finding simultaneously. Different metrics were conducted to assess the classification of machine learning models. A K-fold-repeated-cross-validation (KFRCV) with ten as the number of folds was used. After this process, means and confidence intervals (*CI*) values were obtained.

Data were analyzed using R version 3.6.2 (R Foundation for Statistical Computing). The following packages were used: "EBImage" for characteristics extraction, "mlr" for each machine learning algorithm, and "randomForest" for the random forest (87–89). Additionally, to measure multilabel prediction (classification) were used multilabel accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) (90). Also, we included a positive predictive value. Differences in US findings between women and men were assessed for significance using the *T*-test and chi-squared test. The significance level was considered $p < (0.05)$ and 95% *CI* for all metrics.

Results

Common extensor tendinopathy

A total of 30,007 US images, 6.9 on average in 4,324 exams, and medical records from 2,917 patients with a LET diagnosis were included in the data analysis in this study. Patients' age was presented with a minimum value of 7 and a maximum of 91 years. Women are older than men in 1 year 47.18 ± 11.00 ($p < 0.001$) and also, they presented statistical differences in hypoechogenicity finding in comparison with men ($p = 0.01$).

TABLE 2 The area under the curve (AUC), sensitivity, and specificity [95% *CI*] values of six machine learning classifiers based on degenerative findings in datasets A and B.

Dataset	Measure	Model	HE [95% *CI*]		NV [95% *CI*]		E [95% *CI*]		IST [95% *CI*]	
A	AUC	BR	0.806	(0.81, 0.81)	0.901	(0.900, 0.902)	0.7482	(0.747, 0.749)	0.963	(0.963, 0.964)
		CC	0.810	(0.81, 0.81)	0.897	(0.896, 0.898)	0.6954	(0.689, 0.701)	0.961	(0.960, 0.963)
		DBR	0.804	(0.8, 0.81)	0.892	(0.891, 0.893)	0.6488	(0.647, 0.650)	0.956	(0.954, 0.958)
		NST	0.806	(0.81, 0.81)	0.901	(0.900, 0.902)	0.7463	(0.745, 0.747)	0.963	(0.963, 0.964)
		RF	0.928	(0.93, 0.93)	0.974	(0.973, 0.974)	0.8993	(0.898, 0.9)	0.991	(0.990, 0.991)
		STA	0.806	(0.81, 0.81)	0.847	(0.846, 0.848)	0.688	(0.686, 0.689)	0.935	(0.934, 0.936)
	SE	BR	0.577	(0.58, 0.58)	0.704	(0.703, 0.704)	0.6568	(0.656, 0.657)	0.760	(0.759, 0.760)
		CC	0.578	(0.58, 0.58)	0.702	(0.701, 0.703)	0.6234	(0.619, 0.627)	0.759	(0.758, 0.76)
		DBR	0.576	(0.58, 0.58)	0.699	(0.698, 0.7)	0.594	(0.593, 0.595)	0.756	(0.754, 0.757)
		NST	0.577	(0.58, 0.58)	0.704	(0.703, 0.704)	0.6556	(0.654, 0.656)	0.760	(0.759, 0.760)
		RF	0.607	(0.61, 0.61)	0.741	(0.740, 0.741)	0.7522	(0.751, 0.752)	0.775	(0.774, 0.776)
		STA	0.577	(0.58, 0.58)	0.676	(0.676, 0.677)	0.6187	(0.617, 0.619)	0.744	(0.743, 0.744)
	SP	BR	0.729	(0.73, 0.73)	0.697	(0.696, 0.697)	0.5913	(0.590, 0.591)	0.703	(0.702, 0.70)
		CC	0.732	(0.73, 0.73)	0.695	(0.694, 0.696)	0.5719	(0.569, 0.574)	0.702	(0.701, 0.703)
		DBR	0.728	(0.73, 0.73)	0.692	(0.691, 0.693)	0.5548	(0.554, 0.555)	0.700	(0.699, 0.701)
		NST	0.729	(0.73, 0.73)	0.697	(0.696, 0.697)	0.5906	(0.590, 0.591)	0.703	(0.702, 0.704)
		RF	0.820	(0.82, 0.82)	0.732	(0.732, 0.733)	0.6469	(0.646, 0.647)	0.715	(0.714, 0.716)
		STA	0.729	(0.73, 0.73)	0.670	(0.670, 0.671)	0.5692	(0.568, 0.569)	0.691	(0.690, 0.691)
B	AUC	BR	0.830	(0.83, 0.83)	0.925	(0.923, 0.927)	0.7811	(0.778, 0.784)	0.960	(0.957, 0.963)
		CC	0.830	(0.83, 0.83)	0.906	(0.901, 0.911)	0.7228	(0.714, 0.731)	0.964	(0.961, 0.966)
		DBR	0.788	(0.79, 0.79)	0.846	(0.842, 0.85)	0.6477	(0.643, 0.652)	0.965	(0.963, 0.967)
		NST	0.830	(0.83, 0.83)	0.926	(0.925, 0.928)	0.781	(0.777, 0.784)	0.960	(0.957, 0.963)
		RF	0.888	(0.89, 0.89)	0.965	(0.964, 0.966)	0.8517	(0.849, 0.854)	0.986	(0.985, 0.987)
		STA	0.829	(0.83, 0.83)	0.870	(0.866, 0.873)	0.7222	(0.717, 0.726)	0.937	(0.935, 0.940)
	SE	BR	0.606	(0.61, 0.61)	0.764	(0.762, 0.765)	0.6821	(0.679, 0.684)	0.804	(0.801, 0.806)
		CC	0.606	(0.61, 0.61)	0.752	(0.749, 0.755)	0.6444	(0.638, 0.650)	0.807	(0.804, 0.809)
		DBR	0.592	(0.59, 0.59)	0.714	(0.712, 0.717)	0.5957	(0.592, 0.598)	0.807	(0.805, 0.809)
		NST	0.606	(0.61, 0.61)	0.765	(0.763, 0.766)	0.6821	(0.679, 0.684)	0.804	(0.801, 0.806)
		RF	0.624	(0.62, 0.63)	0.789	(0.787, 0.790)	0.7279	(0.725, 0.73)	0.821	(0.820, 0.823)
		STA	0.605	(0.6, 0.61)	0.729	(0.727, 0.732)	0.6441	(0.640, 0.647)	0.789	(0.787, 0.791)
	SP	BR	0.723	(0.72, 0.73)	0.660	(0.658, 0.661)	0.5983	(0.597, 0.599)	0.654	(0.653, 0.656)
		CC	0.723	(0.72, 0.73)	0.653	(0.651, 0.655)	0.5779	(0.574, 0.580)	0.656	(0.654, 0.657)
		DBR	0.695	(0.69, 0.7)	0.630	(0.628, 0.632)	0.5516	(0.550, 0.553)	0.656	(0.655, 0.658)
		NST	0.723	(0.72, 0.73)	0.660	(0.659, 0.662)	0.5983	(0.597, 0.599)	0.654	(0.653, 0.656)
		RF	0.763	(0.76, 0.76)	0.675	(0.673, 0.676)	0.623	(0.621, 0.624)	0.663	(0.662, 0.665)
		STA	0.723	(0.72, 0.73)	0.639	(0.637, 0.641)	0.5776	(0.576, 0.579)	0.647	(0.645, 0.648)

AUC, area under the curve; SE, sensitivity; SP, specificity; HE, hypoechogenicity; NV, neovascularity; IST, intrasubstance tear; E, enthesopathy; BR, binary relevance model; CC, classifier chains model; NST, nested stacking model; DBR, dependent binary relevance model; STA, staking generalization; RF, random forest.

The total of exams presented at least one degenerative finding. US features are summarized in Table 1.

Machine learning models for a binary classifier

Table 2 shows the binary classification performance (AUC, sensitivity, and specificity) for both datasets (A and B) in each of the six machine learning algorithms. Main degenerative findings in LET (hypoechogenicity, neovascularity, enthesopathy, and intrasubstance tear) were considered under analysis. Focusing on AUC sensitivity and specificity, most models performed with variability among them. Results were described in most cases with a minimal range of 95% *CI*, demonstrating a robust performance for all models. Notably, the RF model obtained the best results. For example, Table 2 shows dataset A, where random forest presented the highest mean values in AUC, sensitivity, and also specificity by each degenerative finding.

10.3389/fmed.2022.945698

TABLE 3 Multilabel accuracy values of six machine learning classifiers based on degenerative findings in both datasets.

Dataset	Model	Macro AUC		Micro AUC		SE		SP		Accuracy		PPV	
A	BR	0.854	(0.854, 0.855)	0.911	(0.910, 0.911)	0.700	(0.700, 0.701)	0.710	(0.71, 0.710)	0.683	(0.682, 0.684)	0.818	(0.816, 0.821)
	CC	0.841	(0.839, 0.842)	0.891	(0.889, 0.893)	0.691	(0.690, 0.692)	0.700	(0.699, 0.701)	0.691	(0.689, 0.692)	0.790	(0.783, 0.798)
	DBR	0.825	(0.824, 0.826)	0.865	(0.864, 0.866)	0.678	(0.678, 0.678)	0.687	(0.686, 0.687)	0.697	(0.696, 0.698)	0.765	(0.7630, 766)
	NST	0.854	(0.853, 0.854)	0.910	(0.910, 0.911)	0.700	(0.700, 0.700)	0.710	(0.709, 0.710)	0.683	(0.682, 0.684)	0.818	(0.816, 0.821)
	RF	0.948	(0.947, 0.948)	0.962	(0.962, 0.963)	0.725	(0.725, 0.726)	0.736	(0.736, 0.737)	0.772	(0.771, 0.773)	0.891	(0.890, 0.892)
	STA	0.819	(0.818, 0.819)	0.897	(0.897, 0.898)	0.694	(0.693, 0.694)	0.703	(0.703, 0.703)	0.683	(0.682, 0.684)	0.818	(0.816, 0.821)
B	BR	0.874	(0.872, 0.875)	0.918	(0.917, 0.919)	0.735	(0.734, 0.736)	0.682	(0.681, 0.682)	0.665	(0.662, 0.668)	0.804	(0.799, 0.809)
	CC	0.855	(0.853, 0.85)	0.899	(0.897, 0.902)	0.725	(0.724, 0.726)	0.674	(0.673, 0.675)	0.676	(0.673, 0.679)	0.777	(0.772, 0.783)
	DBR	0.811	(0.809, 0.813)	0.847	(0.845, 0.849)	0.696	(0.694, 0.697)	0.651	(0.650, 0.652)	0.658	(0.656, 0.661)	0.770	(0.765, 0.775)
	NST	0.874	(0.873, 0.876)	0.918	(0.917, 0.919)	0.736	(0.735, 0.737)	0.682	(0.681, 0.683)	0.666	(0.663, 0.669)	0.804	(0.799, 0.810)
	RF	0.922	(0.921, 0.923)	0.942	(0.941, 0.943)	0.749	(0.748, 0.750)	0.692	(0.692, 0.693)	0.723	(0.721, 0.726)	0.858	(0.855, 0.862)
	STA	0.839	(0.838, 0.841)	0.898	(0.897, 0.899)	0.724	(0.723, 0.725)	0.673	(0.672, 0.674)	0.663	(0.660, 0.665)	0.808	(0.802, 0.814)

AUC, area under the curve; SE, sensitivity; SP, specificity; PPV, positive predictive value; BR, binary relevance model; CC, classifier chains model; NST, nested stacking model; DBR, dependent binary relevance model; STA, staking generalization; RF, random forest.

The AUC and sensitivity showed the best performance in IST with 0.991 [95% *CI*, 0.99, −0.99], and 0.775 [95% *CI*, 0.77, −0.77], respectively. Instead, specificity showed upper values in hypoechogenicity with 0.821 [95% *CI*, 0.82, −0.82].

A similar situation occurred for dataset B, which showed slightly lower values for the same findings and models. The RF model also demonstrated the best performance for all measures and degenerative features. Table 2 showed the highest AUC and sensitivity values for ISR 0.937 [95% *CI*, 0.93–0.94] and 0.82 [95% *CI*, 0.82, −0.82]. Hypoechogenicity also presented better specificity than other degenerative findings with 0.763 [95% *CI*, 0.72, −0.72].

Machine learning models for a multilabel classifier

In the previous results section, the machine learning models assessed a binary classification for each degenerative finding. Now, these methods used a multilabel classifier to identify the four types of tendon findings simultaneously in both datasets. In this scenario, the diagnosis presented different accuracy levels in all machine learning models. When the diagnosis was based on the combination of degenerative findings, the random forest algorithm again presented the best performances among the selected models. Table 3 shows that the random forest in dataset A presented the highest multilabel accuracy value of 0.772 [95% *CI*, 0.771, 0.773]. Similarly, in the condition represented in dataset B, these results show that the model performs well in testing environments without presenting overfitting issues. Multilabel accuracy value was 0.723 [95% *CI*, 0.721, 0.726]. Additionally, high macro and micro-AUC scores are observed in RF models in both datasets. These results could be explained due to the balance between sensitivity and specificity shown in RF models. Particularly, micro-AUC observed in dataset A of 0.962 [95% *CI*, 0.962–0.963] and 0.942 [95% *CI*, 0.941–0.943] in dataset B results are essential because aggregating the contributions of all classes to compute the average metric.

Diagnosis performance

Figure 4 represents dataset A, and the results show the relation between sensitivity vs. 1-specificity across each degenerative finding using the random forest model. In this figure, the plot shows the higher discriminant capacity of diagnosis detection. Most of the lines are located progressively closer to the upper left-hand corner in ROC space. The intrasubstance tear shows the most

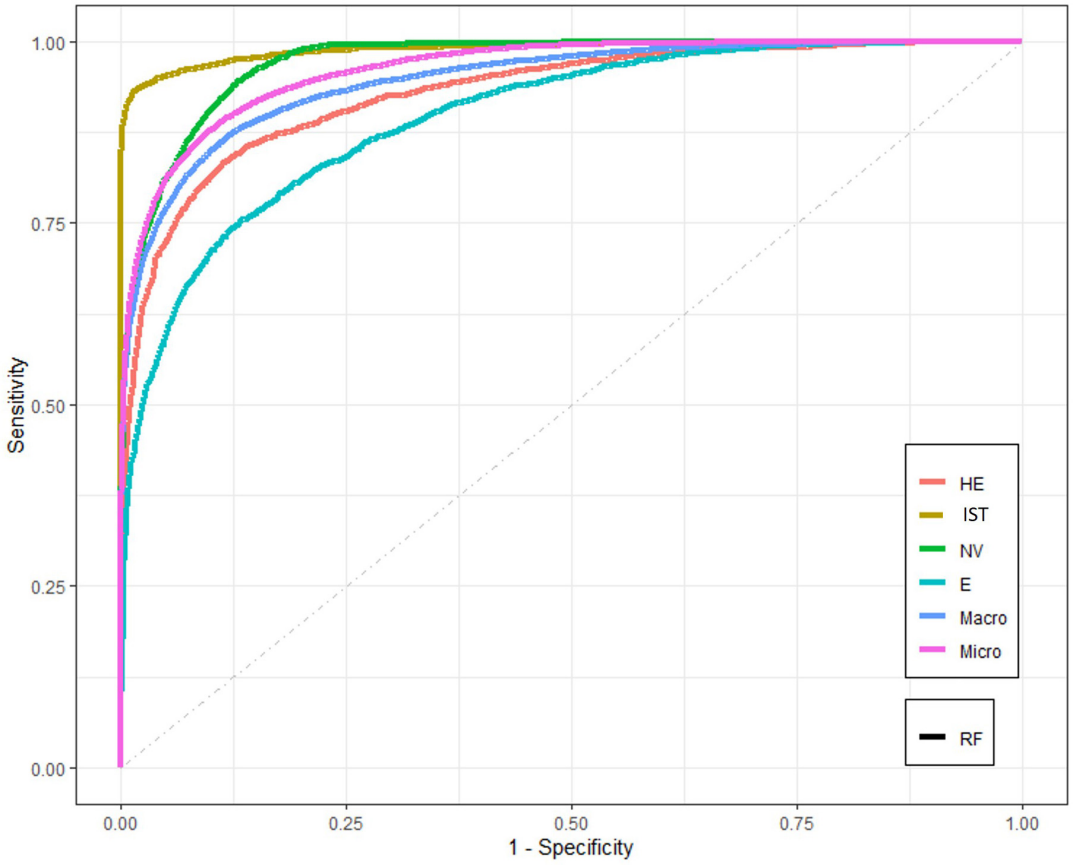

FIGURE 4
The receiver operating characteristic (ROC) curves for RF model for dataset A. Abbreviations: RF, random forest; HE, hypoechogenicity; NV, neovascularity; IST, intrasubstance tear; E, enthesopathy; Macro, macro-AUC; Micro, micro-AUC.

significant discriminate capacity in comparison with the other tendon injuries. However, the enthesopathy finding presented the lowest discriminate capacity in this model.

Discussion

This study is one of the first to present multilabel classification models using machine learning algorithms to detect degenerative findings and intrasubstance tear in US images with LET diagnosis. This retrospective analysis explicitly considered one of the most extensive series of extensor carpi radialis brevis US images, and our machine learning-based tool for diagnosis of LET was trained using the largest dataset so far. The most notable outcomes in this study were obtained by incorporating several machine learning models based on diagnosis know condition. Excellent results and highest values for all degenerative findings were detected in the binary classification performance. Moreover, when the US diagnosis was based on the combination of degenerative findings using a multilabel classifier, the accuracy values presented strong performance too. Our results showed that the random forest algorithm presented the best diagnosis performance, in both binary and multilabel models. These results demonstrate that the implementation of tools derived from artificial intelligence can be used to support the imaging for tendinopathies. Collaborative work between the radiologist and the algorithm could improve the precision of the results, especially if the institution does not have a radiologist specializing in the musculoskeletal area.

Traditionally, US has been demonstrated as a cost-effective tool for detecting abnormalities patterns in tendon structures. Additionally, there is evidence to support the use of US in the detection of LET. A meta-analysis published in 2014 determined that diagnostic test accuracy appears to be highly dependent on numerous variables, such as operator experience, equipment, and stage of pathology. However, US has variable sensitivity and specificity (sensitivity: 64–100%; specificity: 36–100%), decreasing the clinical diagnosis precision (24). Another article published in

the same year reported specifically the sensitivity and specificity for each abnormal US finding using traditional detection method. The hypoechogenicity presented the best combination of diagnostic sensitivity and specificity. It is moderately sensitive sensitivity: 0.64 [95% *CI*, 0.56, 0.72] and highly specific specificity 0.82 [95% *CI*, 0.72, 0.90]. Additionally, neovascularity specificity 1.00 [95% *CI*, 0.97, 1.00)], calcifications specificity 0.97 [95% *CI*, 0.94, 0.99], and cortical irregularities specificity 0.96 [95% *CI*, 0.88, 0.99] have strong specificity for chronic lateral epicondylalgia (25). Our results, particularly for intrasubstance tear detection using the binary algorithm classification in both datasets, demonstrated a superior performance to the traditional US diagnosis methods. In the case of multilabel accuracy, the performance for both indicators was lowest results of specificity and sensitivity than the binary method. This situation could be explained because it is difficult to find a function that minimized the error for more classes. In other words, it increases the variability of the response variable.

For example, in the binary classification, the enthesopathy presented the lowest performance of the six machine learning classifiers. Notably, in the dependent binary relevance model from dataset B, our analysis showed that AUC was 0.647 [95% *CI*, 0.64, 0.65]. This result is quite similar to other reports with a sensitivity of 0.65 and specificity of 0.86 for this finding (77). However, our best result in the binary classification was detecting intrasubstance tear injuries using random forest algorithms. The performance showed an AUC of almost 1.0 (0.99) [95% *CI*, 0.99, 0.99] in contrast with the traditional US methods diagnosis for detecting common extensor tendon tear in the lateral with lower performances in sensitivity, specificity, and accuracy with 64.52, 85.19, and 72.73%, respectively (26).

However, one of our research strengths is the execution of machine learning models using multilabel detection for tendon injury findings. To date, few experiences had been published in the musculoskeletal area using artificial intelligence for tendon pattern detection. Some previous experiences have used Automatic ROI Detection and Classification of the Achilles Tendon ultrasound Images (69), and deep learning models for automatic tracking of the muscle-tendon junction or even measuring muscle atrophy (91). Other disciplines have also used other classification techniques such as neural networks or deep learning convolutional neural networks for image detection, demonstrating excellent results. However, CNN and DL have some drawbacks that should be analyzed when developing predictive models. First, it has been shown that DL requires large datasets to obtain better performance. To handle this, transfer learning is commonly used. However, DL architectures should also be re-trained and model parameters should be optimized, looking out for possible overfitting patterns. Second, DL architectures rely on the high computational performance, and it takes longer to prove results. In this sense, they are more complex to implement, especially in a clinical environment with a high demand for care, so improving diagnostic speed without compromising diagnostic accuracy is crucial for patients and the health system. Therefore, machine learning algorithms are advantageous when speed is of interest. In this case, the execution times of the proposed method were very low, allowing it to be easily implemented in a hospital scenario and re-trained with new data that is daily generated. Finally, the multilabel classification model differs from other algorithms most commonly used in image diagnosis due to the simplicity of its implementation.

This study also has some limitations. Firstly, our images come from the same institution, and patients presented similar socioeconomic conditions. Secondly, we included all static US images from common extensor tendon US per patient, not considering real-time and other structures or tissues. Thirdly, we included tendons with a definitive LET diagnosis, and we did not compare inter and intraobserver variability between radiologists. Fourthly, we considered all images without a region of interest, such as most of the publications. Nevertheless, in a short time, it could be a potential advantage. Finally, we did not repeat the US diagnosis to reduce retrospective bias. However, our radiologist presented more than 10 years of experience.

In conclusion, the random forest model presented the highest sensitivity and specificity in binary and multilabel classifiers for degenerative findings in the common extensor tendon. In particular, intrasubstance tear detections obtained the best performance. Machine learning models could be used to support the US diagnosis of LET.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

This study has been performed following the latest version of the Declaration of Helsinki and the Chilean scientific legislation. The study was approved by the "Comité de Ética Científico Adulto del Servicio Metropolitano Oriente de la ciudad de Santiago de Chile (SSMO)." The Ethics Committee required no informed consent given the nature of the study. The project was approved on August 7th, 2018. No approval number was recorded.

Author contributions

GD: conceptualization, data curation, formal analysis, investigation, methodology, validation, and writing the original draft. MT: data curation, formal analysis, investigation, and visualization. NG: validation, review, and editing. CG: review and editing. CJ: resources and validation. FF: conceptualization, formal analysis, investigation, supervision, review, and editing. All authors contributed to the article and approved the submitted version.

Funding

The authors received no financial support for the research and authorship. The publication was financially supported by the Universidad Mayor.

Acknowledgments

We are grateful for the kind collaboration and assistance of the Sports Medicine Data Science Center MEDS-PUCV. Special thanks to Sandra Mahecha from MEDS Clinic for her support during this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AB-C declared a shared affiliation with one of the authors GD to the handling editor at the time of the review.

Publisher's note

References

1. Stasinopoulos D, Johnson MI. "Lateral elbow tendinopathy" is the most appropriate diagnostic term for the condition commonly referred-to as lateral epicondylitis. *Med Hypotheses.* (2006) 67:1400–2. doi: 10.1016/j.mehy.2006.05.048

2. Struijs PAA, Buchbinder R, Green SE. Tennis elbow. In: Bhandari M editor. *Evidence-Based Orthopedics*. Hoboken, NJ: Wiley-Blackwell (2012). p. 787–95. doi: 10.1002/9781444345100.ch92

3. Shiri R, Viikari-Juntura E, Varonen H, Heliövaara M. Prevalence and determinants of lateral and medial epicondylitis: a population study. *Am J Epidemiol.* (2006) 164:1065–74. doi: 10.1093/aje/kwj325

4. Bunata RE, Brown DS, Capelo R. Anatomic factors related to the cause of tennis elbow. *J Bone Joint Surg Am.* (2007) 89:1955–63. doi: 10.2106/JBJS.F.00727

5. Coombes BK, Bisset L, Vicenzino B. Cold hyperalgesia associated with poorer prognosis in lateral epicondylalgia: a 1-year prognostic study of physical and PS. *Clin J Pain.* (2015) 31:30–5. doi: 10.1097/AJP.0000000000000078

6. Obuchowicz R, Bonczar M. Ultrasonographic Differentiation of Lateral Elbow Pain. *Ultrasound Int Open.* (2016) 2:E38–46. doi: 10.1055/s-0035-1569455

7. Sanders TL, Maradit Kremers H, Bryan AJ, Ransom JE, Smith J, Morrey BF. The epidemiology and health care burden of tennis elbow: a population-based study. *Am J Sports Med.* (2015) 43:1066–71. doi: 10.1177/0363546514568087

8. Roquelaure Y, Ha C, Leclerc A, Touranchet A, Sauteron M, Melchior M, et al. Epidemiologic surveillance of upper-extremity musculoskeletal disorders in the working population. *Arthritis Care Res.* (2006) 55:765–78. doi: 10.1002/art.22222

9. Gruchow HW, Pelletier D. An epidemiologic study of tennis elbow. Incidence, recurrence, and effectiveness of prevention strategies. *Am J Sports Med.* (1979) 7:234–8. doi: 10.1177/036354657900700405

10. Hong Q. Treatment of lateral epicondylitis: where is the evidence?. *Joint Bone Spine.* (2004) 71:369–73. doi: 10.1016/j.jbspin.2003.05.002

11. Bisset LM, Vicenzino B. Physiotherapy management of lateral epicondylalgia. *J Physiother.* (2015) 61:174–81. doi: 10.1016/j.jphys.2015.07.015

12. Zwerus EL, Somford MP, Maissan F, Heisen J, Eygendaal D, Van Den Bekerom MP. Physical examination of the elbow, what is the evidence? A systematic literature review. *Br J Sports Med.* (2018) 52:1253–60. doi: 10.1136/bjsports-2016-096712

13. De Maeseneer M, Brigido MK, Antic M, Lenchik L, Milants A, Vereecke E, et al. Ultrasound of the elbow with emphasis on detailed assessment of ligaments, tendons, and nerves. *Eur J Radiol.* (2015) 84:671–81. doi: 10.1016/j.ejrad.2014.12.007

14. Draghi F, Danesino GM, de Gautard R, Bianchi S. Ultrasound of the elbow: examination techniques and US appearance of the normal and pathologic joint. *J Ultrasound.* (2007) 10:76–84. doi: 10.1016/j.jus.2007.04.005

15. Radunovic G, Vlad V, Micu MC, Nestorova R, Petranova T, Porta F, et al. Ultrasound assessment of the elbow. *Med Ultrasonogr.* (2012) 14:141–6.

16. Pierce JL, Nacey NC. Elbow Ultrasound. *Curr Radiol Rep.* (2016) 4:51. doi: 10.1007/s40134-016-0182-8

17. Barr LL, Babcock DS. Sonography of the normal elbow. *Am J Roentgenol.* (1991) 157:793–8. doi: 10.2214/ajr.157.4.1892039

18. Poltawski L, Jayaram V, Watson T. Measurement issues in the sonographic assessment of tennis elbow. *J Clin Ultrasound.* (2010) 38:196–204. doi: 10.1002/jcu.20676

19. Du Toit C, Stieler M, Saunders R, Bisset L, Vicenzino B. Diagnostic accuracy of power Doppler ultrasound in patients with chronic tennis elbow. *Br J Sports Med.* (2008) 42:872–6. doi: 10.1136/bjsm.2007.043901

20. Maffulli N, Regine R, Carrillo F, Capasso G, Minelli S. Tennis elbow: an ultrasonographic study in tennis players. *Br J Sports Med.* (1990) 24:151–5. doi: 10.1136/bjsm.24.3.151

21. Clarke AW, Ahmad M, Curtis M, Connell DA. Lateral elbow tendinopathy: correlation of ultrasound findings with pain and functional disability. *Am J Sports Med.* (2010) 38:1209–14. doi: 10.1177/0363546509359066

22. Longo UG, Franceschetti E, Rizzello G, Petrillo S, Denaro V. Elbow tendinopathy. *Muscles Ligaments Tendons J.* (2012) 2:115–20.

23. Heales LJ, Broadhurst N, Mellor R, Hodges PW, Vicenzino B. Diagnostic ultrasound imaging for lateral epicondylalgia: a case-control study. *Med Sci Sports Exerc.* (2014) 46:2070–6. doi: 10.1249/MSS.0000000000000345

24. Latham SK, Smith TO. The diagnostic test accuracy of ultrasound for the detection of lateral epicondylitis: a systematic review and meta-analysis. *Orthop Traumatol Surg Res.* (2014) 100:281–6. doi: 10.1016/j.otsr.2014.01.006

25. Dones VC, Grimmer K, Thoirs K, Suarez CG, Luker J. The diagnostic validity of musculoskeletal ultrasound in lateral epicondylalgia: a systematic review. *BMC Med Imaging.* (2014) 4:10. doi: 10.1186/1471-2342-14-10

26. Bachta A, Rowicki K, Kisiel B, Żabicka M, Elert-Kopeæ S, Płomiński J, et al. Ultrasonography versus magnetic resonance imaging in detecting and grading common extensor tendon tear in chronic lateral epicondylitis. *PLoS One.* (2017) 12:e0181828. doi: 10.1371/journal.pone.0181828

27. Matthews W, Ellis R, Furness J, Hing W. Classification of tendon matrix change using ultrasound imaging: a systematic review and meta-analysis. *Ultrasound Med Biol.* (2018) 44:2059–80. doi: 10.1016/j.ultrasmedbio.2018.05.022

28. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell.* (2018) 172:1122–31. doi: 10.1016/j.cell.2018.02.010

29. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5

30. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy workshop. *Radiology.* (2019) 291:781–91. doi: 10.1148/radiol.2019190613

31. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. *Eur J Radiol.* (2018) 102:152–6. doi: 10.1016/j.ejrad.2018.03.019

32. Cascianelli S, Scialpi M, Amici S, Forini N, Minestrini M, Fravolini M, et al. Role of Artificial Intelligence Techniques (Automatic Classifiers) in Molecular Imaging Modalities in Neurodegenerative Diseases. *Curr Alzheimer Res.* (2017) 14:198–207. doi: 10.2174/1567205013666160620122926

33. Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Transl Res.* (2018) 194:56–97. doi: 10.1016/j.trsl.2018.01.001

34. Zhe S, Xu Z, Qi Y, Yu P. Sparse Bayesian multiview learning for simultaneous association discovery and diagnosis of Alzheimer's disease. In: *Proceedings of the National Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press (2015). p. 1966–72. doi: 10.1609/aaai.v29i1.9473

35. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* (2018) 15:e1002686. doi: 10.1371/journal.pmed.1002686

36. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open.* (2019) 2:e191095. doi: 10.1001/jamanetworkopen.2019.1095

37. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology.* (2019) 290:218–28. doi: 10.1148/radiol.2018180237

38. Lee H, Huang C, Yune S, Tajmir SH, Kim M, Do S. Machine friendly machine learning: interpretation of computed tomography without image reconstruction. *Sci Rep.* (2019) 9:15540. doi: 10.1038/s41598-019-51779-5

39. Itu L, Rapaka S, Passerini T, Georgescu B, Schwemmer C, Schoebinger M, et al. A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. *J Appl Physiol.* (2016) 121:42–52. doi: 10.1152/japplphysiol.00752.2015

40. Kolossváry M, De Cecco CN, Feuchtner G, Maurovich-Horvat P. Advanced atherosclerosis imaging by CT: radiomics, machine learning and deep learning. *J Cardiovasc Comput Tomogr.* (2019) 13:274–80. doi: 10.1016/j.jcct.2019.04.007

41. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J.* (2019) 40:1975–86. doi: 10.1093/eurheartj/ehy404

42. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit Med.* (2018) 1:9. doi: 10.1038/s41746-017-0015-z

43. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* (2006) 2:59–78. doi: 10.1177/117693510600200030

44. Wang Z, Yu G, Kang Y, Zhao Y, Qu Q. Breast tumor detection in digital mammography based on extreme learning machine. *Neurocomputing.* (2014) 128:175–84. doi: 10.1016/j.neucom.2013.05.053

45. Ramos-Pollán R, Guevara-López MA, Suárez-Ortega C, Díaz-Herrero G, Franco-Valiente JM, Rubio-Del-Solar M, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *J Med Syst.* (2012) 36:2259–69. doi: 10.1007/s10916-011-9693-2

46. Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing.* (2016) 173:930–41. doi: 10.1016/j.neucom.2015.08.048

47. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. *Med Image Anal.* (2018) 47:45–67. doi: 10.1016/j.media.2018.03.006

48. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, et al. Structure-leveraged methods in breast cancer risk prediction. *J Mach Learn Res.* (2016) 17:85.

49. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed.* (2016) 127:248–57. doi: 10.1016/j.cmpb.2015.12.014

50. Al-Hadidi MR, Alarabeyyat A, Alhanahnah M. Breast cancer detection using K-nearest neighbor machine learning algorithm. In: *Proceedings 2016 9th International Conference on Developments in eSystems Engineering, DeSE*. Liverpool: IEEE (2016). doi: 10.1109/DeSE.2016.8

51. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep*. (2016) 6:27327. doi: 10.1038/srep27327

52. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *J Med Internet Res.* (2019) 7:e10010. doi: 10.2196/10010

53. Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial intelligence in musculoskeletal imaging: current status and future directions. *Am J Roentgenol.* (2019) 213:506–13. doi: 10.2214/AJR.19.21117

54. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiol.* (2020) 49:183–97. doi: 10.1007/s00256-019-03284-z

55. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med.* (2018) 98:8–15. doi: 10.1016/j.compbiomed.2018.05.011

56. Roth HR, Wang Y, Yao J, Lu L, Burns JE, Summers RM. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. In: *Medical Imaging 2016: Computer-Aided Diagnosis*. Bellingham, WA: SPIE (2016). doi: 10.1117/12.2217146

57. Pranata YD, Wang KC, Wang JC, Idram I, Lai JY, Liu JW, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. *Comput Methods Programs Biomed.* (2019) 171:27–37. doi: 10.1016/j.cmpb.2019.02.006

58. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagn Interv Imaging.* (2019) 100:235–42. doi: 10.1016/j.diii.2019.03.002

59. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med.* (2018) 15:e1002699. doi: 10.1371/journal.pmed.1002699

60. Roblot V, Giret Y, Bou Antoun M, Morillot C, Chassin X, Cotten A, et al. Artificial intelligence to diagnose meniscus tears on MRI. *Diagn Interv Imaging.* (2019) 100:243–9. doi: 10.1016/j.diii.2019.02.007

61. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology.* (2018) 289:160–9. doi: 10.1148/radiol.2018172986

62. Koitka S, Demircioglu A, Kim MS, Friedrich CM, Nensa F. Ossification area localization in pediatric hand radiographs using deep neural networks for object detection. *PLoS One.* (2018) 13:e0207496. doi: 10.1371/journal.pone.0207496

63. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology.* (2018) 287:313–22. doi: 10.1148/radiol.2017170236

64. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci Rep.* (2018) 8:1727. doi: 10.1038/s41598-018-20132-7

65. Brattain LJ, Telfer BA, Dhyani M, Grajo JR, Samir AE. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdom Radiol.* (2018) 43:786–99. doi: 10.1007/s00261-018-1517-0

66. Martin K. Special issue on education and training in ultrasound. *Ultrasound.* (2015) 23:5. doi: 10.1177/1742271X14568074

67. van Sloun RJG, Cohen R, Eldar YC. Deep learning in ultrasound imaging. *Proc IEEE.* (2019) 108:11–29. doi: 10.1109/JPROC.2019.2932116

68. Ihnatsenka B, Boezaart AP. Ultrasound: basic understanding and learning the language. *Int J Shoulder Surg.* (2010) 4:55–62. doi: 10.4103/0973-6042.76960

69. Benrabha J, Meziane F. Automatic ROI detection and classification of the achilles tendon ultrasound images. In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*. Liverpool: ACM (2017). p. 1–7. doi: 10.1145/3109761.3158381

70. Baka N, Leenstra S, van Walsum T. Random Forest-Based Bone Segmentation in Ultrasound. *Ultrasound Med Biol.* (2017) 43:2426–37. doi: 10.1016/j.ultrasmedbio.2017.04.022

71. Berton F, Cheriet F, Miron MC, Laporte C. Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images. *Comput Biol Med.* (2016) 72:201–11. doi: 10.1016/j.compbiomed.2016.03.018

72. Kapinski N, Zielinski J, Borucki BA, Trzcinski T, Ciszkowska-Lyson B, Nowinski KS. Estimating achilles tendon healing progress with convolutional neural networks. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G editors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Cham: Springer (2018). doi: 10.1007/978-3-030-00934-2_105

73. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Int J Surg.* (2014) 12:1500–24. doi: 10.1016/j.ijsu.2014.07.014

74. Palaniswamy V, Ng SK, Manickaraj N, Ryan M, Yelland M, Rabago D, et al. Relationship between ultrasound detected tendon abnormalities, and sensory and clinical characteristics in people with chronic lateral epicondylalgia. *PLoS One.* (2018) 13:e0205171. doi: 10.1371/journal.pone.0205171

75. Droppelmann G, Feijoo F, Greene C, Tello M, Rosales J, Yáñez R, et al. Ultrasound findings in lateral elbow tendinopathy: a retrospective analysis of radiological tendon features [version 1; peer review: awaiting peer review]. *F1000Res.* (2022) 11:44. doi: 10.12688/f1000research.73441.1

76. Connell D, Burke F, Coombes P, McNealy S, Freeman D, Pryde D, et al. Sonographic examination of lateral epicondylitis. *Am J Roentgenol.* (2001) 176:777–82. doi: 10.2214/ajr.176.3.1760777

77. Levin D, Nazarian LN, Miller TT, O'Kane PL, Feld RI, Parker L, et al. Lateral epicondylitis of the elbow: US findings. *Radiology.* (2005) 237:230–4. doi: 10.1148/radiol.2371040784

78. Bianchi S, Martinoli C. *Ultrasound of the Musculoskeletal System*. Berlin: Springer (2007).

79. Coombes BK, Bisset L, Vicenzino B. Management of lateral elbow tendinopathy: one size does not fit all. *J Orthop Sports Phys Ther.* (2015) 45:938–49. doi: 10.2519/jospt.2015.5841

80. Vaquero-Picado A, Barco R, Antuña SA. Lateral epicondylitis of the elbow. *EFORT Open Rev.* (2016) 1:391–7. doi: 10.1302/2058-5241.1.000049

81. Sommer C, Gerlich DW. Machine learning in cell biology-teaching computers to recognize phenotypes. *J Cell Sci.* (2013) 126:5529–39. doi: 10.1242/jcs.123604

82. Sun Y, Li L, Zheng L, Hu J, Li W, Jiang Y, et al. Image classification base on PCA of multi-view deep representation. *J Vis Commun Image Represent.* (2019) 62:253–8. doi: 10.1016/j.jvcir.2019.05.016

83. Zhang Z, Sejdiæ E. Radiological images and machine learning: trends, perspectives, and prospects. *Comput Biol Med.* (2019) 108:354–70. doi: 10.1016/j.compbiomed.2019.02.017

84. Buchser W. Assay development guidelines for image-based high content screening, high content analysis and high content imaging. In: Markossian S, Grossman A, Brimacombe K editors. *Assay Guidance Manual*. Bethesda, MD: Eli Lilly & Company (2014).

85. Kumar G, Bhatia PK. A detailed review of feature extraction in image processing systems. In: *2014 Fourth International Conference on Advanced Computing & Communication Technologies*. Rohtak: IEEE (2014). p. 5–12. doi: 10.1109/ACCT.2014.74

86. Nugroho HA, Rahmawaty M, Triyani Y, Ardiyanto I, Choridah L, Indrastuti R. Texture analysis and classification in ultrasound medical images for determining echo pattern characteristics. In: *2017 IEEE International Conference on System Engineering and Technology*. Malaysia: IEEE (2017). p. 23–6. doi: 10.1109/ICSEngT.2017.8123414

87. Sklyar O, Huber W. Image analysis for microscopy screens. *R News.* (2006) 6:12–6.

88. Bosch, B. *Machine Learning in R. Package 'mlr'*. (2021). Available online at: https://cran.r-project.org/web/packages/mlr/mlr.pdf (accessed September 8, 2021).

89. Breiman L, Cutler A. *Breiman and Cutler's Random Forests for Classification and Regression. Package 'randomForest'*. (2022). Available online at: https://cran.rproject.org/web/packages/randomForest/randomForest.pdf (accessed June 15, 2022).

90. Sorower M. *A Literature Survey on Algorithms for Multi-Label Learning*. Corvallis: Oregon State University (2010).

91. Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. *Comput Methods Programs Biomed.* (2019) 182:105063. doi: 10.1016/j.cmpb.2019.105063

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fmed.2022.969724

OPEN ACCESS

EDITED BY
Chris Hodge,
The University of Sydney, Australia

REVIEWED BY
Diego Raimondo,
University of Bologna, Italy
Nguyen Minh Duc,
Pham Ngoc Thach University of Medicine, Vietnam

*CORRESPONDENCE
Zhen Xiao
seriousdoc@163.com
Rui Li
rli@dlut.edu.cn

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 15 June 2022
ACCEPTED 28 September 2022
PUBLISHED 20 October 2022

CITATION
Shen ZW, He YY, Shen ZY, Wang XF, Wang Y, Hua ZY, Jiang N, Song ZJ, Li R and Xiao Z (2022) Novel exploration of Raman microscopy and non-linear optical imaging in adenomyosis. *Front. Med.* 9:969724. doi: 10.3389/fmed.2022.969724

Novel exploration of Raman microscopy and non-linear optical imaging in adenomyosis

Zhuowei Shen[1], Yingying He[2], Zhuoyi Shen[3], Xuefei Wang[2], Yang Wang[4], Zhengyu Hua[5], Nan Jiang[5], Zejiang Song[4], Rui Li[4]* and Zhen Xiao[1]*

[1]Department of Obstetrics and Gynecology, First Affiliated Hospital of Dalian Medical University, Dalian, China, [2]Department of Pathology, Dalian Medical Center for Women and Children, Dalian, China, [3]Department of Information Science and Technology, Wenhua University, Wuhan, China, [4]Department of Physics, Dalian University of Technology, Dalian, China, [5]Department of Pathology, First Affiliated Hospital of Dalian Medical University, Dalian, China

Background: Adenomyosis is a common gynecological disease in women. A relevant literature search found that approximately 82% of patients with adenomyosis chose to undergo hysterectomy. However, women of childbearing age are more likely to undergo surgery to preserve the uterus. Because it is difficult to determine the extent of adenomyosis, it is almost impossible to resect adenomyotic tissue and retain the uterus at the same time.

Materials and methods: Following ethics approval and patient consent, tissue samples were resected and prepared to create frozen slices for analysis. One slice was subjected to H&E staining while the remaining slices were photographed with Coherent Anti-Stokes Raman Scattering (CARS), Second-Harmonic Generation (SHG) microscopy, and Raman spectroscopy. Comparative observations and analyses at the same positions were carried out to explore the diagnostic ability of CARS, SHG, and Raman spectroscopy for adenomyosis.

Results: In adenomyotic tissue, we found two characteristic peaks at 1,155 and 1,519 cm^{-1} in the Raman spectrum, which were significantly different from normal tissue. The substances shown in the CARS spectrum were represented by peaks of 1,519 cm^{-1}. SHG microscopy showed a distribution of collagen at the focus of the adenomyosis.

Conclusion: This study represents a novel analysis of Raman microscopy, CARS, and SHG in the analysis of adenomyotic lesions. We found the diffraction spectrum useful in determining the focal boundary and the diagnosis of adenomyosis in the tested samples.

KEYWORDS

adenomyosis of uterus, disease diagnosis, Raman spectra, CARS, SHG

Introduction

Adenomyosis refers to the invasion of endometrial glands and stroma into the myometrium and the maintenance of functional changes such as periodic hyperplasia, exfoliation, and bleeding. The cause of the disease is unknown (1). It can lead to symptoms such as increased menstruation, prolonged menstruation, and progressive aggravated dysmenorrhea (2). A prior study by Di Donato and co-authors found that 21.8% of patients with endometriosis have adenomyosis. Patients with concurrent adenomyosis have been found to be older, and have a greater pain intensity and depth of infiltration of endometriosis (3). Furthermore, a diagnosis of adenomyosis in these patients has been shown to negatively impact postoperative pain following surgical treatment (4). Adenomyosis can be divided into two types: focal and diffuse. The uterus is uniformly enlarged in diffuse adenomyosis. Focal lesions, known as adenomyosis, grow locally and have no obvious boundary with surrounding tissue, which makes them difficult to resect during surgery. A population sample paper found that about 82% of patients with adenomyosis choose to undergo hysterectomy (5). However, total hysterectomy is obviously not feasible for women with reproductive needs. Uterine-sparing surgery is an alternative surgical treatment, but it is difficult to determine the scope and focus of adenomyosis as it is often mixed with the surrounding normal myometrium; it is therefore almost impossible to completely remove adenomyotic tissue while preserving the uterus (6). At present, the main methods of diagnosing adenomyosis are clinical symptoms and ultrasonography, and the gold standard of diagnosis is postoperative pathology, that is, H&E staining. However, H&E staining takes time and requires a pathologist. Identifying the lesion boundary to aid successful removal therefore represents a significant challenge for surgeons during uterine-sparing surgery (7).

Currently, utilizing engineering technology alongside medicine has proven popular. Among them, the application of optical microscopy in medicine is emerging. Coherent Anti-Stokes Raman scattering (CARS) is a third-order non-linear optical process based on the coherent excitation of molecular vibrations (8), which can obtain the molecular composition and distribution information of the sample to be tested according to the vibrational characteristics of the material molecules. Second-harmonic generation (SHG) microscopy has emerged as a powerful modality for imaging fibrillar collagen in a diverse range of tissues because it is highly sensitive to the collagen fibril/fiber structure (9).

Optical microscopy has been successfully applied to gastric cancer (8), colorectal cancer (9), human meningioma (10, 11), liver cancer (12), lung cancer (13), and other diseases. There have been a large number of studies on cervical cancer (14), ovarian cancer (15), endometrial carcinoma (16), and reproduction (17) in obstetrics and gynecology. Compared with the time-consuming traditional H&E staining method, the biggest advantage of optical microscopy lies in its convenience and efficiency. Notably, the prior application of Raman Microscopy in determining the lesion range in a cohort undergoing surgical treatment for brain cancer provided possible parallels to the identification and resection of tissue boundaries in adenomyosis patients (18). Our hypothesis therefore became: can the focus and scope of adenomyosis be determined through optical microscopy to meet the needs of women of childbearing age and to perform adenomyosis surgery with uterine preservation?

This study describes the first use of a Raman microscope, CARS, and SHG to study adenomyotic lesions.

Materials and methods

Sample preparation

This is a prospective study from February 2021 to March 2022. We randomly selected 10 patients (five normal and five with adenomyosis) who underwent surgery in the First Affiliated Hospital of Dalian Medical University. The adenomyosis patients were determined by preoperative ultrasound examination. After cutting off the uterus during the operation, we retain several pieces of tissue (in case of adenomyosis, parts of the adenomyosis lesion and the rest of normal muscle tissue will be retained) and place them in liquid nitrogen tanks for cold storage, in order to preserve the cell activity for the convenience of subsequent experiments. Patients with adenomyosis provided both adenomyotic and normal tissue samples while non-adenomyosis patients provided normal tissue samples. Finally, a total of 20 adenomyotic tissue samples and 20 normal samples (including five adenomyosis patients' normal samples) were included in this study. Adenomyosis samples were from patients who underwent total hysterectomy due to adenomyosis, and normal *in vitro* samples of the control group were from normal uterine muscle tissues of patients who underwent total hysterectomy owing to non-malignant diseases (to prevent tumor tissues from affecting the results), such as hysteromyoma and uterine prolapse. All patients signed the informed consent form under the informed consent of the research process after surgery, allowing us to conduct experiments on their *in vitro* tissues. This experiment was certified by the ethics Association of the First Affiliated Hospital of Dalian Medical University (IRB number: PJ-KS-KY-2022-257).

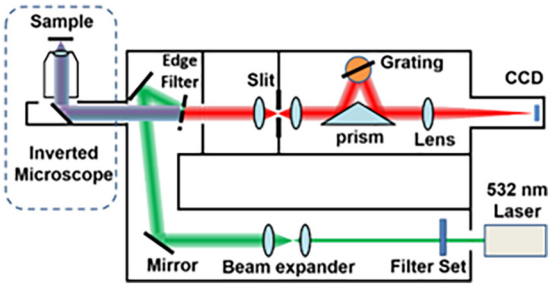

FIGURE 1
Layout of the optical system of the Raman micro-spectrometer.

After sample preparation, we took 40 fresh tissues (20 normal tissues and 20 adenomyosis tissues) about 1 × 1 cm in size from 10 uteruses (five normal tissues and five adenomyosis tissues), and made continuous frozen sections for each tissue. From each tissue, three 10 μm sections (slices) were cut. The three slices of the same tissue were numbered 1, 2, and 3. All "Slices 1" underwent H&E staining, and Slices 2 and 3 were directly observed under a non-linear optics microscope without any staining treatment.

H&E

Two experienced pathologists observed all H&E-stained slices to provide a diagnosis of either adenomyosis or normal tissue; their corresponding Slices 2 and 3 were imaged by CARS and Raman microscope, respectively.

Raman spectra

Raman spectra were obtained using a commercial Raman micro-spectrometer (Renishaw, InVia system) at 532 nm excitation wave number, which was focused onto the muscles using a 50× (NA = 0.75) objective for an integration time of 10 s. Cosmic ray was removed after acquiring each spectrum using the Renishaw WiRE 4.4 software. The experimental setup and its schematic illustration are shown in Figure 1.

Because the Raman microscope displays spectrum images of substances in a limited range, different substances display different Raman signals, so the carrier glass carrying tissue slices will inevitably display their own Raman signals. At this time, the glass is measured separately to display the Raman signal of the glass itself as a reference (red in Figure 2), so that the peak value of the glass and the characteristic peak value of adenomyosis can be distinguished.

We first identified the characteristic wave number range in the range of 500–3,000 cm^{-1}. As shown in Figure 2, the

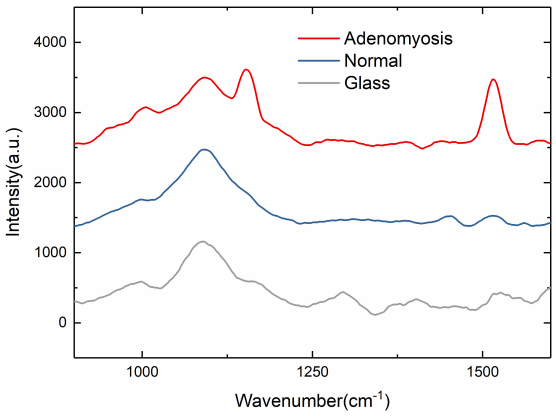

FIGURE 2
900–1,600 cm^{-1} Raman spectra of normal, adenomyosis, and glass (normal, adenomyosis, and glass from top to bottom).

characteristic wave number is about 1,200 and 1,500 cm^{-1}, so we set the wave number range at 900–1,600 cm^{-1} to facilitate the experiment.

Anti-stokes Raman scattering and second-harmonic generation

Figure 3 shows a schematic of the CARS system for non-linear optical imaging. Briefly, a mode-locked 80 fs Ti:sapphire laser (MaiTai, Spectra Physics, Santa Clara, USA) is tuned to 800 nm with pulse width at an 80 MHz repetition rate and divided into two parts by a polarization beam splitter. One beam works as the pump beam; the other beam is used to pump a photonic crystal fiber to produce the Stokes beam for CARS imaging. Two beams are combined at the dichroic mirror. The combined beams are sent into a multiphoton scanning microscope (Olympus, FV1200) and focused on the sample by an objective (10×, NA 0.4; UplanApo, Olympus, Tokyo, Japan). The average power of 75 mW is used for the pump and the probe beam. The CARS and SHG signals pass through a bandpass filter, respectively, before being detected by the PMT.

Results

All Slice 1 samples were viewed under the microscope before imaging for records (Figure 4A). Since Slices Nos. 1, 2, and 3 were cut continuously by a slicer, any differences can be ignored. We observed images with CARS and SHG at the same position of Slices 2. Examples of images are adenomyosis lesions imaged by CARS and SHG (Figure 4).

Figure 4A shows the contrast diagram of H&E staining. The lesions shown in the figure are the subject of this study. The "Y" structure pointed by the white arrow in the figure is the

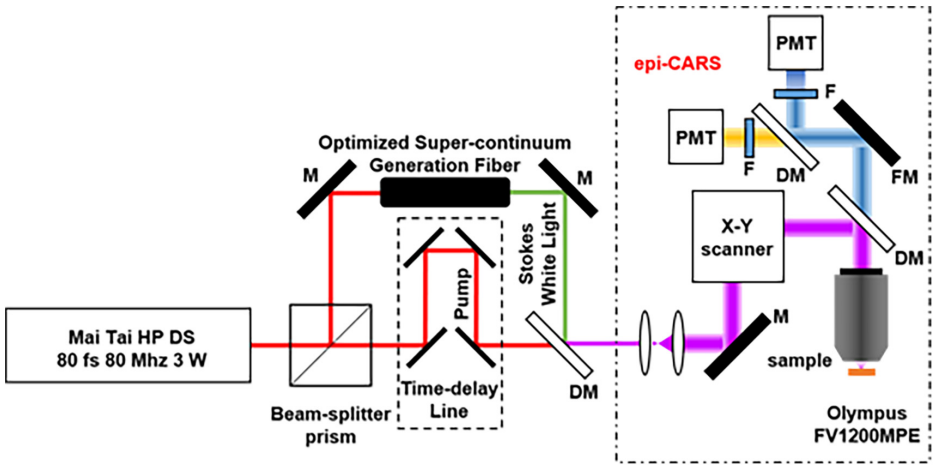

FIGURE 3
Optical path of the CARS system (M, mirror; DM, Dichroic Mirrors; F, Filter; FM, Flip Mirror; PMT, photomultiplier tube).

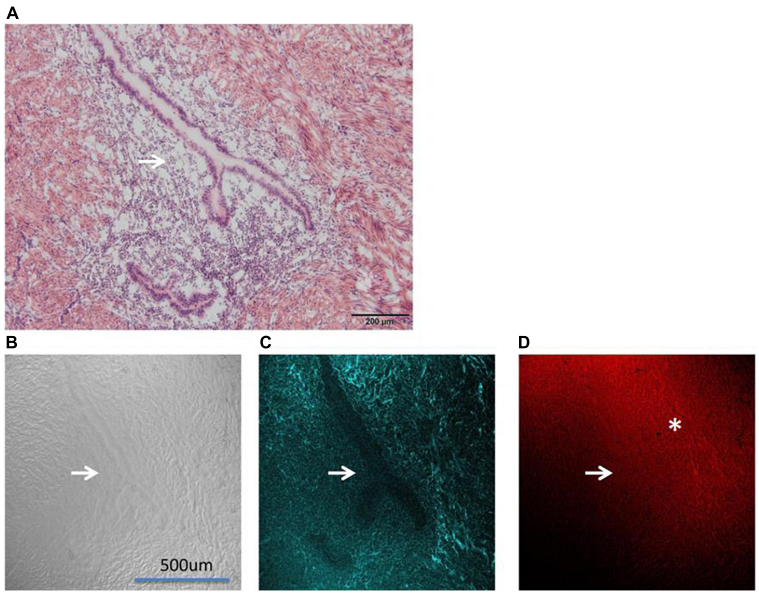

FIGURE 4
Shows the images of adenomyosis lesions under different microscopes (→: Ectopic uterine gland). **(A)** An H&E staining section. **(B)** DIC (differential interference contrast microscope) imaging. **(C)** SHG imaging. **(D)** CARS imaging (*: fibro-collagen proliferation).

ectopic endometrial structure in the myometrium, namely, the uterine gland. As shown in Figure 4A, the uterine gland is a single tube gland with branches at its end, mainly composed of secretory cells.

Figure 4B shows the differential interference contrast microscope (DIC imaging). DIC imaging is an image directly observed by the naked eye without staining. From the DIC imaging, we can see the sense of uneven layers in the image, vaguely seeing the ectopic uterine gland (the position indicated by the white arrow), but the peripheral structure is not clear.

In the Figure 4C CARS microscope presents the outline of the ectopic uterine gland perfectly. It can be seen that under the CARS microscope, the glandular part of the uterine gland is not imaged (the position indicated by the white arrow), and the image intensity of the surrounding interstitial part is relatively light, while the intensity of the surrounding muscle layer is relatively high. We used a 720 nm filter in CARS. The Raman wave number range corresponding to the 720 nm filter just includes our second characteristic peak of 1,519 cm^{-1}. Therefore, we determined that the imaging substance of CARS

was consistent with the representative substance of the second characteristic peak in Raman imaging.

Figure 4D shows the SHG image. Interestingly, SHG, unlike CARS imaging, showed stronger intensity in the uterine gland and its surrounding fibro-collagen proliferation section. The SHG microscope is widely used to image various fibrous collagens. 3D also clearly shows the distribution of collagen around the lesion. It can be seen that fibro-collagen proliferation exists around the uterine gland (*Marking section).

In the range of 900–1,600 cm^{-1}, the characteristic wave number of adenomyosis is more obvious and allows a clear distinction between the Raman spectrum curve of normal tissue and adenomyosis (Figure 2). In the following experiment, we used adenomyosis tissues from different patients, and the characteristic peaks appeared at 1,155 and 1,519 cm^{-1}.

Discussion

In this report, we used Raman, DIC, CARS, and SHG microscopes to directly image tissue sections without staining, and took HE staining images at the same location for comparison.

CARS microscopy, probing vibrations of molecular bonds for image contrast, and the high vibrational Raman cross sections of many hydrogen carbon bonds make the technique suitable for imaging polymers (19). CARS microscopy derives its contrast from intrinsic molecular vibrations in a sample; the CH group of membrane and cortical cytoskeleton proteins are the basis of CARS imaging (20). In the absence of staining (Figure 4C), CARS can clearly show the outline of the ectopic uterine gland and its boundary with surrounding tissue structure compare with Figure 4A, which has been stained with H&E. SHG visualizes highly ordered tissue structures, which are non-centrosymmetric like type I collagen fibers (21). As shown in Figure 4D, there is obvious fibro-collagen proliferation around the uterine gland, which is caused by bleeding of adenomyosis (22).

According to previous experiments, the characteristic curve of cervical cancer is concentrated at 720, 785, 1,095, 1,258, and 1,579 cm^{-1} (23–25). This is different from the representative peaks of adenomyosis (1,155 and 1,519 cm^{-1}) found in our study. Most studies regarding Raman microscopy in adenomyosis focus on the serological aspects of patients (26). During our literature search, only one such direct histological study was found. In this article (27), Wang et al. identified a peak different from that of normal tissue at 1,173 cm^{-1} in adenomyosis, believing the peak is induced by delta (C—O) shifts. Our initial peak was found to be 1,155 cm^{-1}. Considering Wang and co-authors used a light source of 785 nm compared to our own source at 532 nm, we consider that this finding is broadly consistent. However, our finding of the additional peak at 1,519 cm^{-1} in adenomyosis samples represents a novel finding. To understand our novel finding, we reviewed the existing literature to identify biological macromolecules and concurrent Raman wave numbers (Table 1). First, after an extensive literature search and integration, we created a corresponding table between Raman wave numbers and biological macromolecules (14, 16, 28–30). From Table 1, we can see that most representative substances with similar wave numbers are the same (however, there may be errors caused by different measurements). The corresponding substance of 1,516 cm^{-1} is amide II, considering some errors caused by different experimental conditions (temperature, tissue freshness, etc.) and instrument measurements, so our first hypothesis about the characteristic peak at 1,519 cm^{-1} was amide II. Of particular interest to our findings, two prior lung cancer studies using Raman microscopy found characteristic carotenoid Raman peaks at 1,152 and 1,518 cm^{-1} with the Raman peaks in lung cancer patients lower than those in normal subjects. The authors suggested these findings reflected C–C and conjugated C=C bond stretch (24, 31). In our study, characteristic peaks were found at 1,155 and 1,519 cm^{-1} in adenomyosis tissue, which is very similar to the characteristic peaks of carotenoids at 1,152 and 1,518 cm^{-1} in the previous two studies. Carotenoids represent the main source of Vitamin A in the body and provide anti-oxidation, immune regulation, anti-cancer, and anti-aging effects. Our findings representing similar peaks may support a possible relationship between carotenoids and adenomyosis; however, this remains speculative and requires additional investigation.

The outstanding advantages of Raman spectroscopy lie in its label-free nature and timeliness, which reduce the waiting time of intraoperative pathology and the burden upon pathologists at the surgery. Currently, Hand-Held Raman technology has been successfully applied to detect air components and diagnose plant diseases (32–34). There are also a large number of intraoperative boundary studies of brain tumors in medicine (18). Currently, there is no research regarding Hand-Held Raman technology on disease or surgery in obstetrics and gynecology. Our results suggest a possible further role for Hand-Held Raman microscopy in assisting the intraoperative diagnosis of adenomyosis and the localization of lesion boundaries to improve potential surgical outcomes in patients. Similarly, handheld SHG technology are also areas that have not been studied and discussed. The results of this study also found the potential utility in determining the location of adenomyosis lesions. SHG also confirmed the proliferation of fibro-collagen caused by bleeding around adenomyosis lesions. The application of these two microscopes in surgery will further help to determine and diagnose the location of adenomyosis lesions.

TABLE 1 Wave number of biomacromolecules.

Assignment	Raman shift (cm^{-1})
DNA	481, 784, 788, 826
DNA/RNA	1,231, 1,320
Saccharides	1,370
Monosaccharide	898
Disaccharide	898
Polysaccharide	477
Glycogen	933, 1,003, 1,025, 1,150
Amylaceum	540
Collagens	859, 1,032, 1,303, 1,309, 1,325, 1,332, 1,339, 1,445
Phosphatidylinositol	415, 519, 576
Phospholipid	1,085, 1,032, 1,078, 1,445, 1,745
Cholesterol	548
Cholesteryl ester	538, 614
Lipid	877, 968, 1,125, 1,057, 1,060, 1,095, 1,124, 1,275, 1,309, 1,369, 1,437, 1,447, 1,450, 1,452
Glycerol	630
Nuclein	1,299, 1,340, 1,578
Tyrosine	640, 642, 643, 821, 823, 830, 835, 849, 853, 855, 859, 1,170, 1,616
Methionine	695
Aspartate	1,700
Glutamate	1,700
Tryptophan	745, 752, 758, 880, 1,208, 1,365, 1,374, 1,376, 1,552, 1,560, 1,561, 1,616, 1,618, 1,618
Proline	814, 821, 853, 855, 880, 918, 928, 933, 935, 936, 1,043, 1,066, 1,447
Hydroxyproline	821, 853, 876, 1,588
Valine	928, 933, 935, 936, 1,066
Phenylalanine	1,000, 1,002, 1,003, 1,004, 1,030, 1,104, 1,582, 1,583, 1,588, 1,602
Cysteine	495–516
Protein	933, 951, 1,158, 1,369
Phosphorylated protein	968, 970
Pyrimidine ring	766
Uracil	780, 784
Cytosine	784, 1,175, 1,290, 1,506
Thymine	784
Guanine	1,175, 1,369
Adenine	721, 1,335
Porphyrin	1,369
C-C skeleton	928, 938, 1,130, 1,561
C-C stretching (collagen)	817
C-C stretching (phenylalanine)	1,339
C-H stretching (protein)	1,295
C-N stretching (protein)	1,053, 1,128
C-O stretching (protein)	1,053
C-O stretching (lipid)	1,723, 1,738, 1,792
Ribose vibration	867, 915
Antisymmetric vibration of phosphoric acid	1,185–300
Antisymmetric phosphate stretching vibration	1,230
Amide I	1,600, 1,601, 1,624, 1,637, 1,640, 1,645, 1,654, 1,655, 1,658, 1,660, 1,664, 1,670, 1,685, 1,697
Amide II	1,516, 1,570
Amide III	1,234, 1,236, 1,243, 1,246, 1,255, 1,275, 1,285, 1,302
β-Carotenoids	1,152, 1,518, 1,520

Conclusion

In this experiment, a Raman microscope, CARS, and SHG were used to study adenomyosis, which demonstrated the role of non-linear optics in diagnosing adenomyosis and distinguishing lesion boundaries. Moreover, the combination of CARS and SHG microscopes produces more extensive and complementary information.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Department of the First Affiliated Hospital of Dalian Medical University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

ZWS wrote the article, designed the experiment, and analyzed the results. YH, XW, ZH, and NJ were four experienced pathology teachers who contributed to our H&E staining and sectioning. ZYS, YW, and ZJS contributed to the revision of charts, literature retrieval, and experiments. This article had taken place under the guidance of two experienced tutors who were corresponding authors of this article, ZX and RL. All authors agree to be responsible for the content of this article and the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China-Liaoning Joint Fund (Grant No. 2019-BS-073) and the Scientific Research Fund of Liaoning Provincial Education Department (Grant No. LZ2019044).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Camboni A, Marbaix E. Ectopic endometrium: the pathologist's perspective. *Int J Mol Sci.* (2021) 22:10974.

2. Szubert M, Kozirog E, Wilczynski J. Adenomyosis as a risk factor for myometrial or endometrial neoplasms-review. *Int J Environ Res Public Health.* (2022) 19:2294.

3. Di Donato N, Montanari G, Benfenati A, Leonardi D, Bertoldo V, Monti G, et al. Prevalence of adenomyosis in women undergoing surgery for endometriosis. *Eur J Obstet Gynecol Reprod Biol.* (2014) 181:289–93.

4. Brosens I, Benagiano G. Poor results after surgery for rectovaginal endometriosis can be related to uterine adenomyosis. *Hum Reprod.* (2012) 27: 3360–1.

5. Yu O, Schulze-Rath R, Grafton J, Hansen K, Scholes D, Reed SD. Adenomyosis incidence, prevalence and treatment: United States population-based study 2006-2015. *Am J Obstet Gynecol.* (2020) 223:94 e1–e10. doi: 10.1016/j.ajog.2020.01.016

6. Stratopoulou CA, Donnez J, Dolmans MM. Conservative management of uterine adenomyosis: medical vs. surgical approach. *J Clin Med.* (2021) 10:4878.

7. Horng HC, Chen CH, Chen CY, Tsui KH, Liu WM, Wang PH. Uterine-sparing surgery for adenomyosis and/or adenomyoma. *Taiwan J Obstet Gynecol.* (2014) 53:3–7.

8. Allen CH, Hansson B, Raiche-Tanner O, Murugkar S. Coherent anti-Stokes Raman scattering imaging using silicon photomultipliers. *Opt Lett.* (2020) 45:2299–302. doi: 10.1364/OL.390050

9. Chen X, Nadiarynkh O, Plotnikov S, Campagnola PJ. Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. *Nat Protoc.* (2012) 7:654–69.

10. Zhang L, Zhou Y, Wu B, Zhang S, Zhu K, Liu CH, et al. Intraoperative detection of human meningioma using a handheld visible resonance Raman analyzer. *Lasers Med Sci.* (2022) 37:1311–9. doi: 10.1007/s10103-021-03390-2

11. Hollon T, Orringer DA. Label-free brain tumor imaging using Raman-Based methods. *J Neurooncol.* (2021) 151:393–402. doi: 10.1007/s11060-019-03380-z

12. Yan S, Cui S, Ke K, Zhao B, Liu X, Yue S, et al. Hyperspectral stimulated Raman scattering microscopy unravels aberrant accumulation of saturated fat in human liver cancer. *Anal Chem.* (2018) 90:6362–6. doi: 10.1021/acs.analchem.8b01312

13. Zhang Z, Yu W, Wang J, Luo D, Qiao X, Qin X, et al. Ultrasensitive Surface-Enhanced Raman scattering sensor of gaseous aldehydes as biomarkers of lung cancer on dendritic Ag nanocrystals. *Anal Chem.* (2017) 89:1416–20. doi: 10.1021/acs.analchem.6b05117

14. Karunakaran K, Saritha VN, Joseph MM, Nair JB, Saranya G, Raghu KG, et al. Diagnostic spectro-cytology revealing differential recognition of cervical cancer lesions by label-free surface enhanced Raman fingerprints and chemometrics. *Nanomedicine.* (2020) 29:102276. doi: 10.1016/j.nano.2020.102276

15. Honda K, Hishiki T, Yamamoto S, Yamamoto T, Miura N, Kubo A, et al. On-tissue polysulfide visualization by surface-enhanced Raman spectroscopy benefits

patients with ovarian cancer to predict Post-Operative chemosensitivity. *Redox Biol.* (2021) 41:101926.

16. Schiemer R, Furniss D, Phang S, Seddon AB, Atiomo W, Gajjar KB. Vibrational biospectroscopy: an alternative approach to endometrial cancer diagnosis and screening. *Int J Mol Sci.* (2022) 23:4859. doi: 10.3390/ijms23094859

17. O'Brien CM, Vargis E, Rudin A, Slaughter JC, Thomas G, Newton JM, et al. In vivo Raman spectroscopy for biochemical monitoring of the human cervix throughout pregnancy. *Am J Obstet Gynecol.* (2018) 218:528 e1–e18.

18. Karabeber H, Huang R, Iacono P, Samii JM, Pitter K, Holland EC, et al. Guiding brain tumor resection using surface-enhanced Raman scattering nanoparticles and a Hand-Held Raman scanner. *ACS Nano.* (2014) 8:9755–66. doi: 10.1021/nn503948b

19. Kee TW, Cicerone MT. Simple approach to one-laser, broadband coherent Anti-Stokes Raman scattering microscopy. *Opt Lett.* (2004) 29:2701–3. doi: 10.1364/ol.29.002701

20. Arkill KP, Moger J, Winlove CP. The structure and mechanical properties of collecting lymphatic vessels: an investigation using multimodal nonlinear microscopy. *J Anat.* (2010) 216:547–55. doi: 10.1111/j.1469-7580.2010.01215.x

21. Sehm T, Uckermann O, Galli R, Meinhardt M, Rickelt E, Krex D, et al. Label-free multiphoton microscopy as a tool to investigate alterations of cerebral aneurysms. *Sci Rep.* (2020) 10:12359. doi: 10.1038/s41598-020-69222-5

22. Wang S, Li B, Duan H, Wang Y, Shen X, Dong Q, et al. Abnormal expression of connective tissue growth factor and its correlation with fibrogenesis in adenomyosis. *Reprod Biomed Online.* (2021) 42:651–60. doi: 10.1016/j.rbmo.2020.11.002

23. Daniel A, Prakasarao A, Ganesan S. Near-infrared Raman spectroscopy for estimating biochemical changes associated with different pathological conditions of cervix. *Spectrochim Acta A Mol Biomol Spectrosc.* (2018) 190:409–16. doi: 10.1016/j.saa.2017.09.014

24. Huang Z, McWilliams A, Lui H, McLean DI, Lam S, Zeng H, et al. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. *Int J Cancer.* (2003) 107:1047–52.

25. Lyng FM, Faolain EO, Conroy J, Meade AD, Knief P, Duffy B, et al. Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool. *Exp Mol Pathol.* (2007) 82:121–9. doi: 10.1016/j.yexmp.2007.01.001

26. Parlatan U, Inanc MT, Ozgor BY, Oral E, Bastu E, Unlu MB, et al. Raman spectroscopy as a non-invasive diagnostic technique for endometriosis. *Sci Rep.* (2019) 9:19795. doi: 10.1038/s41598-019-56308-y

27. Liu G, Liu JH, Zhang L, Yu F, Sun SZ. [Raman spectroscopic study of uterine pathological tissue]. *Guang Pu Xue Yu Guang Pu Fen Xi.* (2005) 5:723–5.

28. Duraipandian S, Mo J, Zheng W, Huang Z. Near-infrared Raman spectroscopy for assessing biochemical changes of cervical tissue associated with precarcinogenic transformation. *Analyst.* (2014) 139:5379–86. doi: 10.1039/c4an00795f

29. Sitarz K, Czamara K, Bialecka J, Klimek M, Zawilinska B, Szostek S, et al. HPV infection significantly accelerates glycogen metabolism in cervical cells with large nuclei: Raman microscopic study with subcellular resolution. *Int J Mol Sci.* (2020) 21:2667. doi: 10.3390/ijms21082667

30. Wang J, Zheng CX, Ma CL, Zheng XX, Lv XY, Lv G-D, et al. Raman spectroscopic study of cervical precancerous lesions and cervical cancer. *Lasers Med Sci.* (2021) 36:1855–64.

31. Bakker Schut TC, Puppels GJ, Kraan YM, Greve J, van der Maas LL, Figdor CG. Intracellular carotenoid levels measured by Raman micro spectroscopy: comparison of lymphocytes from lung cancer patients and healthy individuals. *Int J Cancer.* (1997) 74:20–5. doi: 10.1002/(sici)1097-0215(19970220)74:1<20::aid-ijc4>3.0.co;2-2

32. Egging V, Nguyen J, Kurouski D. Detection and identification of fungal infections in intact wheat and sorghum grain using a hand-held Raman spectrometer. *Anal Chem.* (2018) 90:8616–21. doi: 10.1021/acs.analchem.8b01863

33. Farber C, Sanchez L, Kurouski D. Confirmatory non-invasive and non-destructive identification of poison ivy using a hand-held Raman spectrometer. *RSC Adv.* (2020) 10:21530–4. doi: 10.1039/d0ra03697h

34. Heleg-Shabtai V, Zaltsman A, Sharon M, Sharabi H, Nir I, Marder D, et al. Explosive vapour/particles detection using SERS substrates and a hand-held Raman detector. *RSC Adv.* (2021) 11:26029–36. doi: 10.1039/d1ra04637c

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fmed.2022.988927

OPEN ACCESS

EDITED BY
Victoria Bunik,
Lomonosov Moscow State University, Russia

REVIEWED BY
Mark O. Wielpütz,
Heidelberg University, Germany
Alexander Pfeil,
University Hospital Jena, Germany

*CORRESPONDENCE
Hubert S. Gabryś
hubert.gabrys@usz.ch
Stephanie Tanadini-Lang
stephanie.tanadini-lang@usz.ch

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 07 July 2022
ACCEPTED 31 October 2022
PUBLISHED 17 November 2022

CITATION
Gabryś HS, Gote-Schniering J, Brunner M, Bogowicz M, Blüthgen C, Frauenfelder T, Guckenberger M, Maurer B and Tanadini-Lang S (2022) Transferability of radiomic signatures from experimental to human interstitial lung disease. *Front. Med.* 9:988927. doi: 10.3389/fmed.2022.988927

Transferability of radiomic signatures from experimental to human interstitial lung disease

Hubert S. Gabryś[1]*, Janine Gote-Schniering[2,3], Matthias Brunner[3], Marta Bogowicz[1], Christian Blüthgen[4], Thomas Frauenfelder[4], Matthias Guckenberger[1], Britta Maurer[3] and Stephanie Tanadini-Lang[1]*

[1]Department of Radiation Oncology, University Hospital Zurich, Zurich, Switzerland, [2]Comprehensive Pneumology Center, Institute of Lung Health and Immunity, Helmholtz Zentrum München, Member of the German Center for Lung Research (DZL), Munich, Germany, [3]Department of Rheumatology and Immunology, University Hospital Bern, University of Bern, Bern, Switzerland, [4]Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland

Background: Interstitial lung disease (ILD) defines a group of parenchymal lung disorders, characterized by fibrosis as their common final pathophysiological stage. To improve diagnosis and treatment of ILD, there is a need for repetitive non-invasive characterization of lung tissue by quantitative parameters. In this study, we investigated whether CT image patterns found in mice with bleomycin induced lung fibrosis can be translated as prognostic factors to human patients diagnosed with ILD.

Methods: Bleomycin was used to induce lung fibrosis in mice (n_control = 36, n_experimental = 55). The patient cohort consisted of 98 systemic sclerosis (SSc) patients (n_ILD = 65). Radiomic features (n_histogram = 17, n_texture = 137) were extracted from microCT (mice) and HRCT (patients) images. Predictive performance of the models was evaluated with the area under the receiver-operating characteristic curve (AUC). First, predictive performance of individual features was examined and compared between murine and patient data sets. Second, multivariate models predicting ILD were trained on murine data and tested on patient data. Additionally, the models were reoptimized on patient data to reduce the influence of the domain shift on the performance scores.

Results: Predictive power of individual features in terms of AUC was highly correlated between mice and patients (r = 0.86). A model based only on mean image intensity in the lung scored AUC = 0.921 ± 0.048 in mice and AUC = 0.774 (CI95% 0.677-0.859) in patients. The best radiomic model based

on three radiomic features scored AUC = 0.994 ± 0.013 in mice and validated with AUC = 0.832 (CI95% 0.745-0.907) in patients. However, reoptimization of the model weights in the patient cohort allowed to increase the model's performance to AUC = 0.912 ± 0.058.

Conclusion: Radiomic signatures of experimental ILD derived from microCT scans translated to HRCT of humans with SSc-ILD. We showed that the experimental model of BLM-induced ILD is a promising system to test radiomic models for later application and validation in human cohorts.

KEYWORDS

radiomics, preclinical imaging, interstitial lung disease, lung fibrosis, systemic sclerosis, bleomycin

Introduction

Interstitial lung disease (ILD) defines a group of chronic, etiologically different parenchymal lung disorders, characterized by fibrosis as their common final pathophysiological stage. The prognosis of the most prevalent and severe subtypes, idiopathic pulmonary fibrosis (IPF) and ILD associated with the autoimmune disease systemic sclerosis (SSc), is as poor as that of untreated oncologic diseases (1, 2). Globally, non-malignant lung diseases including ILD rank third on the mortality scale (3).

Experimental models of fibrosing ILD are paramount for the identification of cellular and molecular key drivers of disease and as preclinical test systems for novel targeted drugs (4). The preferred and best characterized preclinical model of ILD is the murine model of bleomycin-induced lung fibrosis, which reflects important features of human ILD such as apoptosis of epithelial cells, influx of inflammatory cells into the interstitium, followed by activation of fibroblasts with increased deposition of extracellular matrix (ECM) proteins (5, 6).

Conventional endpoint measures of lung fibrosis involve histological and biochemical analyses, which, however, have certain disadvantages. To recapitulate the dynamic process of fibrosing ILD at multiple time points and to account for the high interindividual variability, large numbers of animals are required to reach significant statistical power (7). Additionally, lung biopsies are only rarely performed in human ILD (8, 9) and biopsy may not be representative for the whole lung pathology. Upcoming alternative outcome measures for translational ILD research include imaging methodologies. An integral part of the routine clinical management is medical imaging, particularly high-resolution computed tomography (HRCT), which allows non-invasive, highly sensitive, time- and spatially resolved visualization of the entire lung changes (10) and a correlative estimation of lung function (11). Similarly, in preclinical models of ILD, small animal microCT is increasingly recognized as a valuable assessment tool (4, 7). In the model of bleomycin-induced experimental ILD, the relative comparability of both imaging and molecular changes with human ILD (5, 12–15) support its suitability for translational ILD research.

The need for innovative, directly transferable, and readily applicable readouts in ILD have prompted the herein presented translational study on the potential value of the model of bleomycin-induced lung fibrosis as experimental "radiomic toolbox" for human ILD. Radiomics is a powerful strategy for in-depth analysis of pathologic tissue phenotypes by computational extraction of quantitative imaging features from medical images (16, 17). Radiomic features provide objective information on tissue shape, intensity, and texture on a molecular scale as demonstrated by studies on tumor biology showing correlation with tissue-based genomics and proteomics data (18–21). As image-derived tissue surrogates, their potential use as virtual biopsies could make radiomics analyzes an ideal tool for clinical decision support in ILD especially since radiomic features have also been shown to predict disease outcome and response to therapy (18, 19, 22–25). However, compared with oncology (18, 20–22), research into the potential of radiomics in non-malignant lung diseases is limited (26–30).

Nevertheless, the available literature on human lung pathologies, including chronic obstructive pulmonary disease, radiation-induced pneumonitis and connective tissue disease-related ILD showed that texture-based analysis of CT images can be superior compared to the visual or histogram-based measures for diagnosis (28, 31, 32). Few studies investigated the use of radiomics in experimental settings. Eresen et al. used MRI radiomics for prediction of response to vaccine therapy in a mouse model of pancreatic ductal adenocarcinoma (33, 34). Nunez et al. analyzed suitability of MRI radiomics for diagnosis of preclinical GL261 glioblastoma (35). Other researchers focused on radiomic-based prediction of liver metastases or liver fibrosis in mice (36, 37).

To date no study has shown the value of animal models in radiomics research. We are not aware of any studies reporting

transferability of radiomic patterns from experimental model to clinical setting. Establishing a link between preclinical and clinical radiomic patterns could enormously facilitate testing a vast range of hypotheses in an experimental setting. Such a link is currently missing. In this analysis, we evaluate if radiomic features and models can be translated from experimental to human ILD.

Materials and methods

Study design and data sets

Details of the study design and data sets are shown in Figure 1. In short, we investigated whether radiomic patterns indicative of ILD in mice were also present in human disease.

The preclinical model of bleomycin (BLM)-induced lung fibrosis was used to mimic human ILD. The experimental cohort consisted of 91 8-week-old female mice (C57BL/6J-rj, Janvier Labs). ILD was induced in 55 mice via intratracheal instillation of bleomycin (2 U/kg; Baxter 15,000 I.U.) as described in (6, 14, 38). The 36 control animals received equivalent volumes of 0.9% NaCl solution. Mice were randomized into the different experimental groups and instillation was performed blinded. Pulmonary micoCT scans were performed at different days (days 3, 7, 14, 21, 28, and 35) after bleomycin instillation to reflect different disease stages. Different mice were scanned at every time point as the animals were euthanized after image acquisition. Scanning mice at different time points after fibrosis induction did not serve a particular purpose in this work. Such design was chosen because this experimental data was also used in other studies which examined temporal aspect of fibrotic development.

A cohort of 98 SSc patients being followed at the Department of Rheumatology, University Hospital Zurich represented the validation data set. All included patients met the following criteria: diagnosis of SSc according to the Very Early Diagnosis of Systemic Sclerosis (VEDOSS) (39) or the 2013 American College of Rheumatology//European League against Rheumatism (ACR/EULAR) classification criteria (40), and availability of an HRCT scan. Patient characteristics are provided in Table 1.

The extent of lung fibrosis was defined as presence of reticular changes or honeycombing within whole lung volume (Figure 2). All visual analyses were performed by a senior radiologist (TF) using a standard picture archiving and communication system workstation (Impax, Version 6.5.5.1033; Agfa-Gevaert) and a high-definition liquid crystal display monitor (BARCO; Medical Imaging Systems).

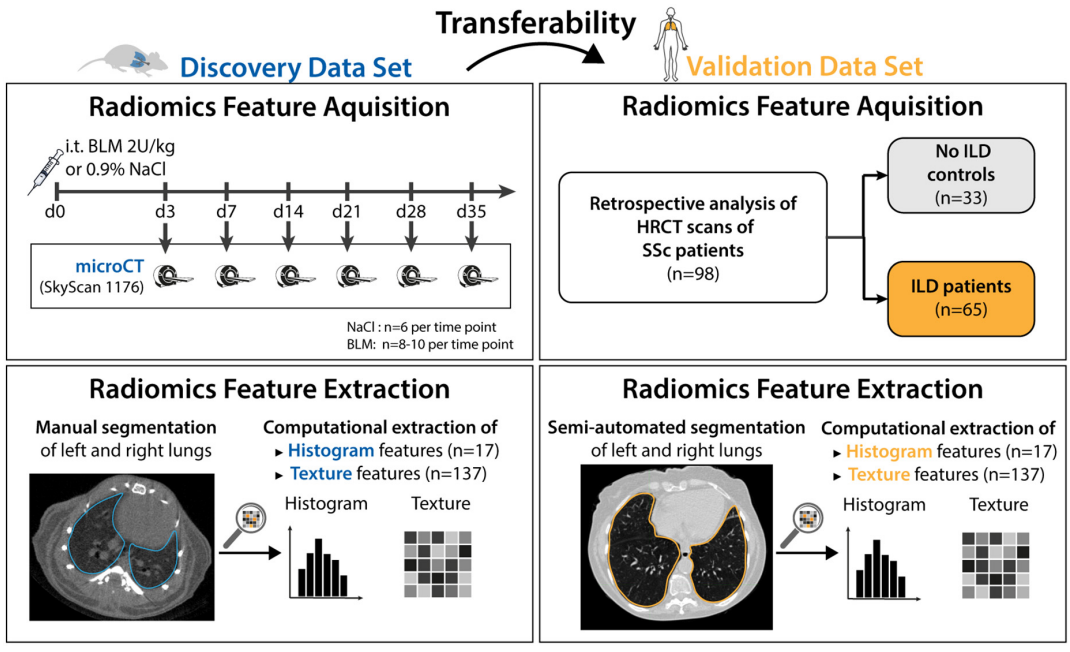

FIGURE 1
Study design. The mice data set (n = 91) was used to discover radiomic patterns predictive of ILD. The discovered patterns were tested in the human validation data set (n = 98). 55 mice were given Bleomycin to induce ILD, whereas 36 mice were given NaCl and served as the control group. The mice were euthanized at day 3, 7, 14, 21, 28, and 35 and scanned with a microCT scanner. Afterward, classification models were trained to predict occurrence of ILD based on images acquired from the scanner. The 98 patients from the validation data set were retrospectively collected. All patients were scanned with HRCT and graded according to the Goh scale of pulmonary fibrosis. The radiomic models built using mice data were tested in patients.

TABLE 1 Summary of patient's demographics and clinical baseline characteristics.

Characteristics	Zurich Cohort (n = 98)
Age (year)	60.0 ± 19.0
Sex	
Male	21 (21.4%)
Female	77 (78.6%)
Disease duration (year)*	5.0 ± 8.6
SSc subset (LeRoy 1988)	
Limited cutaneous SSc	41 (41.8%)
Diffuse cutaneous SSc	37 (37.8%)
No skin involvement	20 (20.4%)
Skin involvement	
Limited cutaneous	34 (34.7%)
Diffuse cutaneous	36 (36.7%)
No skin involvement	23 (23.5%)
Only sclerodactyly	5 (5.1%)
Autoantibodies	
Anti-centromere positive	26 (26.5%)
Anti-topoisomerase I positive	35 (35.7%)
Anti-RNA polymerase III positive	8 (8.2%)
Anti-PMScl positive	15 (15.3%)
FVC (% predicted)	91.0 ± 37.0
DLCO (% predicted)	70.0 ± 35.0
FEV1 (% predicted)	92.0 ± 27.0
Pulmonary hypertension†	18 (18.4%)
PAPsys (MmHg)	24.5 ± 9.8
6 min walk distance (m)	530.0 ± 172.5
SpO_2 before 6-MWT (%)	97.0 ± 1.0
SpO_2 after 6-MWT (%)	95.0 ± 6.8
Borg scale (unit)	3.0 ± 2.0
Extent of lung fibrosis on CT	
None	33 (33.7%)
Present	65 (66.3%)
Ground glass opacification	25 (25.5%)
Reticular changes	64 (65.3%)
Tractions	38 (38.8%)
Honeycombing	21 (21.4%)
Bullae	3 (3.1%)
Radiological subtype#	
NSIP	55 (56.1%)
UIP	9 (9.2%)
DIP	1 (1.0%)
Immunomodulatory therapy§	42 (42.9%)

Continuous variables are described as median ± interquartile range and categorical variables are present as absolute numbers with relative frequencies (percent).

*Disease duration of SSc was calculated as the difference between the date of baseline CT and the date of manifestation of the first non-Raynaud's symptom.

†Pulmonary hypertension was assessed by echocardiography or right heart catheterization.

#Radiological subtypes were only determined for SSc patients with ILD.

§Immunomodulatory therapy included prednisone, methotrexate, rituximab, cyclophosphamide, mycophenolate mofetil, hydroxychloroquine, tocilizumab, imatinib, azathioprine, adalimumab, leflunomid, cyclosporine.

PAPsys, systolic pulmonary artery pressure; FVC, forced vital capacity; FEV1, forced expiratory volume in 1 second; DLCO, diffusing capacity for carbon monoxide; 6-MWT, 6-min walk test; UIP, usual interstitial pneumonia; NSIP, non-specific interstitial pneumonia; DIP, diffuse interstitial pneumonia.

Imaging and extraction of radiomic features

Pulmonary microCT scans were acquired in free-breathing mice with prospective respiratory gating using Bruker SkyScan 1176. The following scan parameters were used: tube voltage 50 kV, tube current 500 μA, filter Al 0.5 mm, averaging (frames) 3, rotation step 0.7 degrees, sync with event 50 ms, X-ray tube rotation 360 degrees, resolution 35 μm, and slice thickness 35 μm. Images were reconstructed with NRecon reconstruction software (v.1.7.4.6; Bruker) using the built-in filtered back projection Feldkamp algorithm and applying misalignment compensation, ring artifact reduction, and a beam hardening correction of 10% to the images.

HRCT scans were acquired using Siemens scanners (SOMATOM Definition AS, SOMATOM Definition Flash, SOMATOM Force, SOMATOM Sensation 64, SOMATOM Sensation 16, Biograph 64, LightSpeed Pro 16, LightSpeed VCT). The scans were acquired in an inspiration (breath hold) mode. The median slice thickness was 1 mm (range 0.6-2 mm) and the median tube voltage was 120 kVp (range 80-150 kVp). The reconstruction kernels included B60f, B70f, and Bl64.

The contouring of whole lungs was performed manually in mice and semi-automatically in patients (region growing algorithm followed by manual correction) by two experienced examiners (JS and MB). Left and right lungs were contoured independently and then both contours were merged to generate a single contour including both lungs.

Feature extraction from CT images was performed with Z-Rad, an IBSI-compliant (41), in-house developed Python software. CT scans of mice and patients were interpolated to an isotropic resolution of 0.15 mm and 2.75 mm, respectively. The interpolation resolutions were chosen to achieve similar ratio of voxel size to average lung volume in mice and patients. The region of interest (ROI) for feature extraction was defined as the right and the left lung considered as a single organ. Only intensity values within the range from −1,000 HU to 200 HU were considered. We used a fixed bin size of 50 HU. The radiomic features describing image intensity (histogram, n = 17) and texture (n = 137) were extracted for each mouse and patient. The texture features were based on gray level co-occurrence matrix (GLCM, n = 26), gray level run length matrix (GLRLM, n = 16), gray level distance zone matrix (GLDZM, n = 16), gray level size zone matrix (GLSZM, n = 16), neighboring gray level dependence matrix (NGLDM, n = 16), and neighborhood gray tone difference matrix (NGTDM, n = 5) to capture wide variety of intensity patterns. Additionally, GLCM and GLRLM features were extracted with two different feature aggregation methods - with and without merging. In total, 154 features were extracted. The list of radiomic features is provided in the supplement.

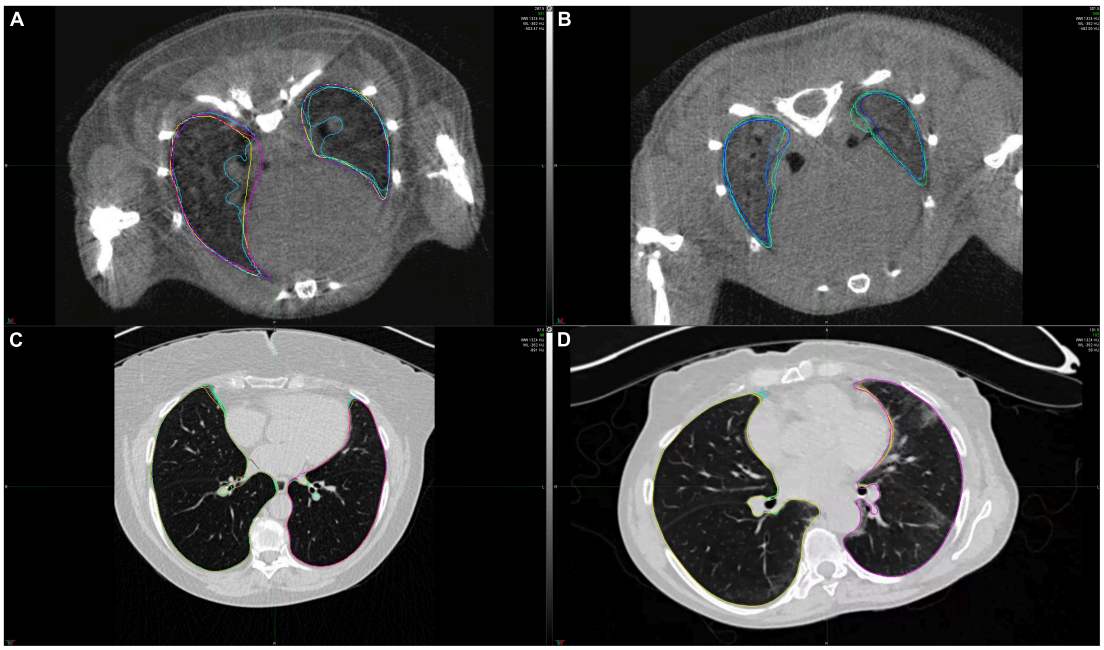

FIGURE 2
Example CT scans of healthy lungs and lungs affected with lung fibrosis. **(A)** microCT image of a healthy mice lung, **(B)** microCT image of a mice lung with lung fibrosis, **(C)** HRCT image of a healthy human lung, **(D)** HRCT image of a human lung with lung fibrosis. Lung contours marked in different colors show the extent of intra- and interobserver variability in lung segmentations for these two cases.

Statistical analysis

For every radiomic feature, robustness against intra- and interobserver variability was examined. This was realized with estimation of the corresponding intraclass correlation coefficients (ICC). Specifically, we used consistency of ICC (1, 3) according to the Shorut and Fleiss naming convention (42). Features with ICC $\geq$0.75 for intra- and interobserver settings in both mice and humans were considered stable and were retained. The rest of the features were excluded from further analysis.

Univariate predictive power of the radiomic features was evaluated by estimation of the area under the receiver operating characteristic curve (AUC). To facilitate comparison of the AUC values between mice and patient data sets, we adopted a convention that AUC is equal to the probability that a radiomic feature value of a randomly chosen patient from the positive group is greater than the value of a randomly chosen patient from the negative group. This allowed us to distinguish between features that were characterized by comparable predictive power but a different *direction* of the effect, for example, AUC = 0.3 in mice and AUC = 0.7 in patients. The linear association of the AUC scores between mice and patient groups has been evaluated with Pearson correlation coefficient.

Three model architectures were considered for evaluation of model transferability from mice to patients: (1) a model based on mean image intensity (MEAN), (2) a model based on first four moments of intensity distribution (mean, standard deviation, skewness, and kurtosis; MSSK), and (3) a machine learning model based on logistic regression (ML). While the first two models are based on predefined radiomic features, the machine learning model employed embedded feature selection methods. All models were built on the mice data and were validated in the patient data.

Feature selection and model tuning was realized within 4-times repeated 5-fold cross-validation. The first step of the feature selection procedure was dimensionality reduction by removing features that were highly linearly correlated (Pearson's r). The correlation threshold was one of tunable hyperparameters. The second step of feature selection was fitting a model and selection of most important features from this model which were then fed to the final classifier. In the case of a logistic regression model, the feature selection was realized with another logistic regression. In the case of, extra-trees model, most important features were extracted from a gradient tree-boosting model. The number of extracted features in both cases was one of tunable hyperparameters. For model tuning, we used 500 randomized hyperparameter samples. The optimized models were validated in patients. Additionally, the models were re-optimized in patients to evaluate transferability and predictive power of the discovered radiomic signatures rather than the models themselves. Furthermore, this allowed to reduce the influence of covariate shift between the data sets.

For visualization, statistical analysis, model building, and model testing, the following open-source Python packages were used: Matplotlib (43), NumPy & SciPy (44), Pandas (45), and scikit-learn (46).

Results

Influence of intra- and interobserver delineation variability on radiomic features

Intra- and interobserver delineation variability were evaluated separately in mice and patient data sets using 15 randomly selected cases per data set. Intraobserver variability was assessed based on delineations done by JS. Interobserver variability was assessed based on delineations provided by JS, CB, and MBr. **Figure 3** shows the proportion of the unstable features per feature class. In mice, 7 features from the initial set of 154 were considered unstable (ICC $<$ 0.75) and were excluded from the further analysis. In patients, all features were stable (ICC $\geq$ 0.75) so no further features were excluded.

Discriminative power of radiomic features is highly correlated between mice and patient data

The next steps in our analysis were the investigation of univariate discriminative power of radiomic features and the correlation of AUC scores between mice and patients. ICC analysis was performed to compare two feature aggregation methods of GLCM and GLRLM features. As both feature aggregation methods rendered highly correlated results (ICC_{GLCM} = 0.99, ICC_{GLRLM} = 0.83), only one feature aggregation per feature class method was kept for further analysis to reduce feature redundancy.

Univariate predictive power of radiomic features in terms of AUC is presented in **Figure 4A**. On average, features describing image intensity tended to perform better than texture-based features. Radiomic features were on average more predictive

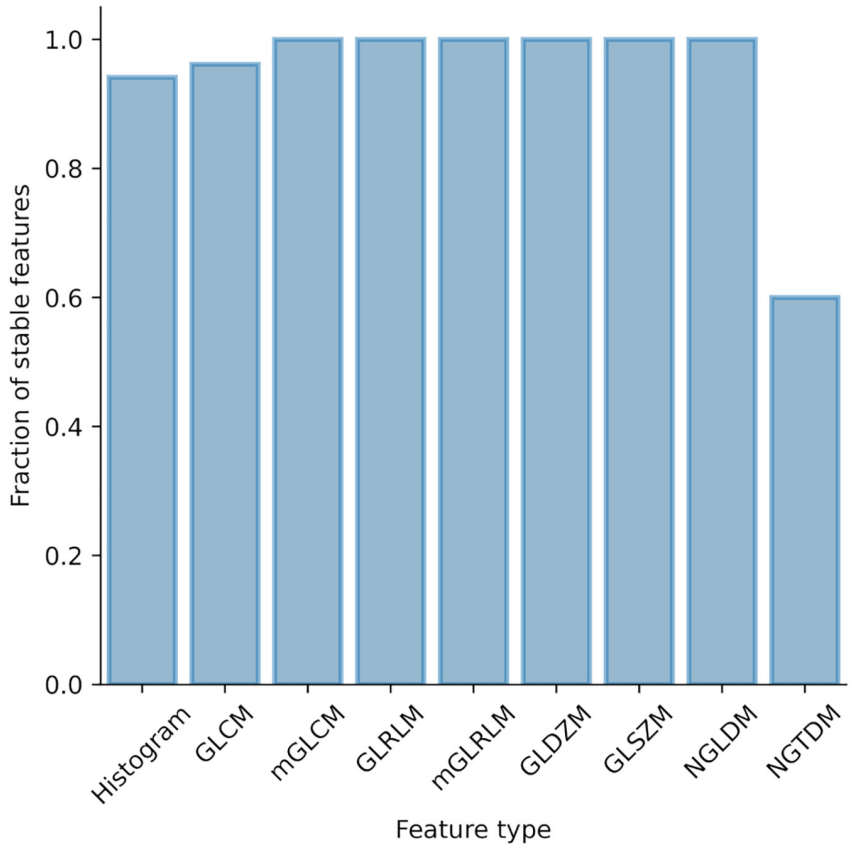

FIGURE 3
Influence of intra- and interobserver delineation variability on radiomic features stability. Proportion of unstable features stratified by feature type.

10.3389/fmed.2022.988927

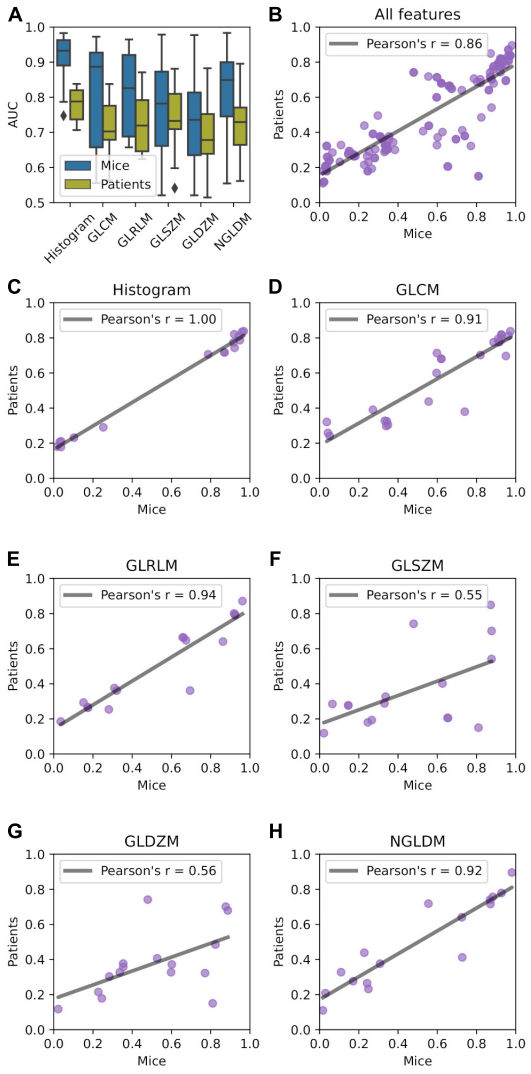

FIGURE 4
Relationship between predictive power of radiomic features in mice and patient data sets. **(A)** AUC distribution stratified by feature type (histogram, gray level co-occurrence matrix (GLCM, n = 26), gray level run length matrix (GLRLM, n = 16), gray level distance zone matrix (GLDZM, n = 16), gray level size zone matrix (GLSZM, n = 16), and neighboring gray level dependence matrix (NGLDM, n = 16). **(B–H)** Correlation of the AUC between mice and patient groups.

in mice than in patients. Most predictive features in mice achieved AUC = 0.988, whereas in patients AUC = 0.896. The complete list of feature predictive performance is provided in the supplement.

Univariate predictive power of the features was highly correlated between murine and patient groups (Figure 4B) with Pearson's r = 0.86. Very high correlation was observed for histogram-, GLCM-, GLRLM-, and NGLDM-based features (Figures 4C–E,H). GLSZM- and GLDZM-based features exhibited more variability (Pearson's r < 0.6; Figures 4F,G).

TABLE 2 Predictive performance in model tuning, testing, and re-optimization.

Model	AUC tuning (SD)	AUC testing (95% CI)	AUC re-opt. (SD)	TPR (95% CI)	TNR (95% CI)	PPV (95% CI)	NPV (95% CI)	LR + (95% CI)	LR- (95% CI)
MEAN	0.921 ± 0.048	0.774 (0.677, 0.859)	0.780 ± 0.110	0.54 (0.37, 0.71)	0.99 (0.97, 1.00)	0.96 (0.88, 1.00)	0.82 (0.73, 0.90)	53.69 (4.82, 597.44)	0.46 (0.32, 0.67)
MSSK	0.990 ± 0.017	0.815 (0.728, 0.890)	0.818 ± 0.092	0.65 (0.48, 0.82)	0.99 (0.97, 1.00)	0.97 (0.90, 1.00)	0.85 (0.77, 0.93)	64.35 (5.83, 710.65)	0.35 (0.22, 0.57)
ML	0.994 ± 0.013	0.832 (0.745, 0.907)	0.912 ± 0.058	0.78 (0.63, 0.92)	0.99 (0.97, 1.00)	0.97 (0.91, 1.00)	0.90 (0.83, 0.97)	76.92 (7.01, 844.09)	0.23 (0.12, 0.43

The uncertainty of the AUC estimates is provided as standard deviation of the cross-validation scores for model tuning and re-optimization and 95% confidence intervals for model testing. Additionally, true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR +), and negative likelihood ratio (LR-) were estimated based on the cutoff point maximizing Youden's J statistic (TPR + TNR-1). All 95% confidence intervals were estimated with bootstrap.

Radiomic patterns predictive of interstitial lung disease translate from experimental interstitial lung disease to patients

To analyze transferability of radiomic patterns and models from mice to patients we built and validated four classes of models: (1) a model based on mean image intensity (MEAN), (2) a model based on first four moments of intensity distribution (mean, standard deviation, skewness, and kurtosis; MSSK), and (3) a machine learning model based on logistic regression (ML). The models were trained on mice data and tested in patients. Additionally, the models were reoptimized in patients, that is, retrained using the features from the mouse models. The results and comparison of model performance is shown in Table 2.

All models achieved high diagnostic performance in mice. The baseline MEAN model scored AUC = 0.921 which left little room for improvement. Nevertheless, the MSSK and the ML models exceeded AUC = 0.990 resulting in almost perfect classification performance. Testing model performance in patients resulted in AUC scores varying from 0.754 (MEAN) to 0.832 (ML). Model re-optimization in patients allowed to improve the predictive performance of all models. ROC curves associated with model tuning, testing, and re-optimization together with the underlying features are presented in Figure 5. ROC curves show that re-optimization gave little improvement for the MEAN and the MSSK models as testing and re-optimization curves followed similar characteristics. On the other hand, machine learning models improved significantly in this process. The corresponding re-optimization ROC curves detached from the testing curves to position between tuning and testing curves.

Substantial differences in distribution of radiomic features included in the models in terms of location and dispersion are presented in Figure 6. Most of the features exhibit patterns of the same direction in both mice and patient data sets, that is, either rising or falling trend from healthy to ILD.

Discussion

In this analysis, we report that radiomic features and models can be translated from experimental to human ILD. Collectively, our data suggest that well characterized and

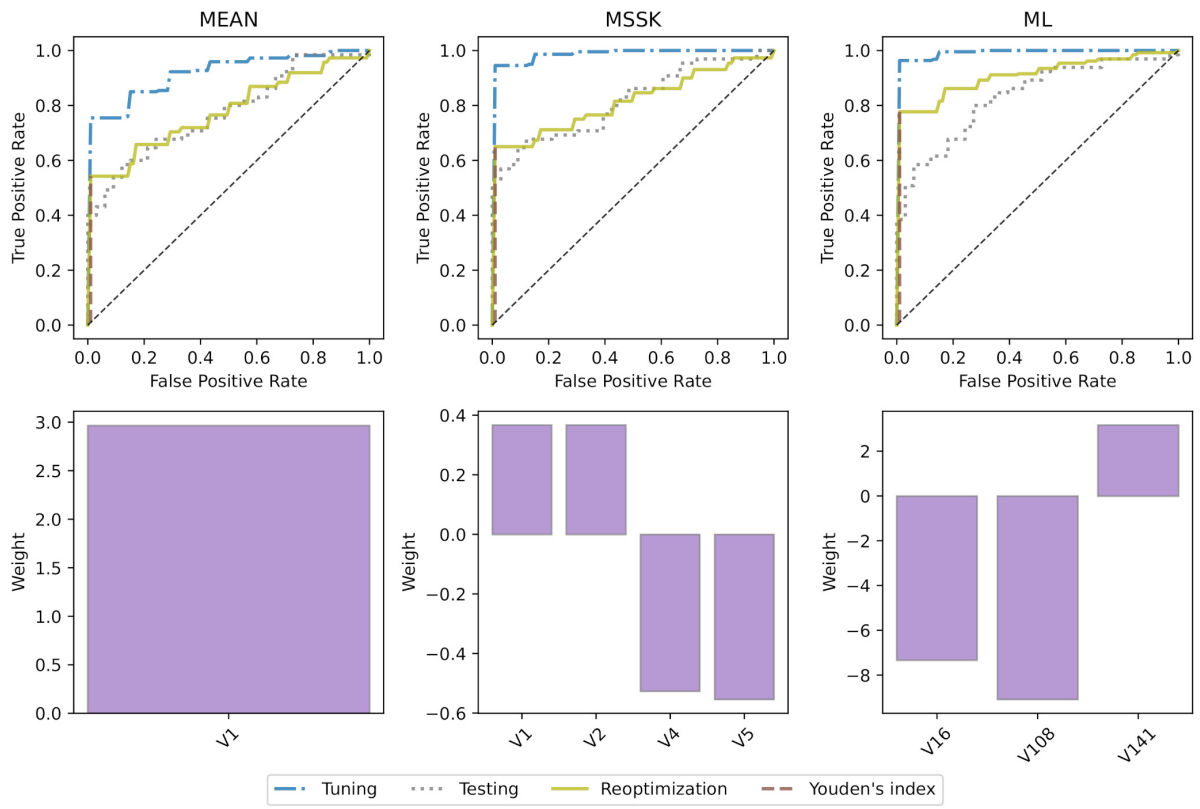

FIGURE 5
Model performance and underlying radiomic features. ROC curves and bar plots of the underlying features. V1 - mean (histogram), V2 - standard deviation (histogram), V4 - skewness (histogram), V5 - kurtosis (histogram), V16 - root mean square (histogram), V108 - gray level non-uniformity normalized (GLSZM), V141 - dependence count non-uniformity (NGLDM).

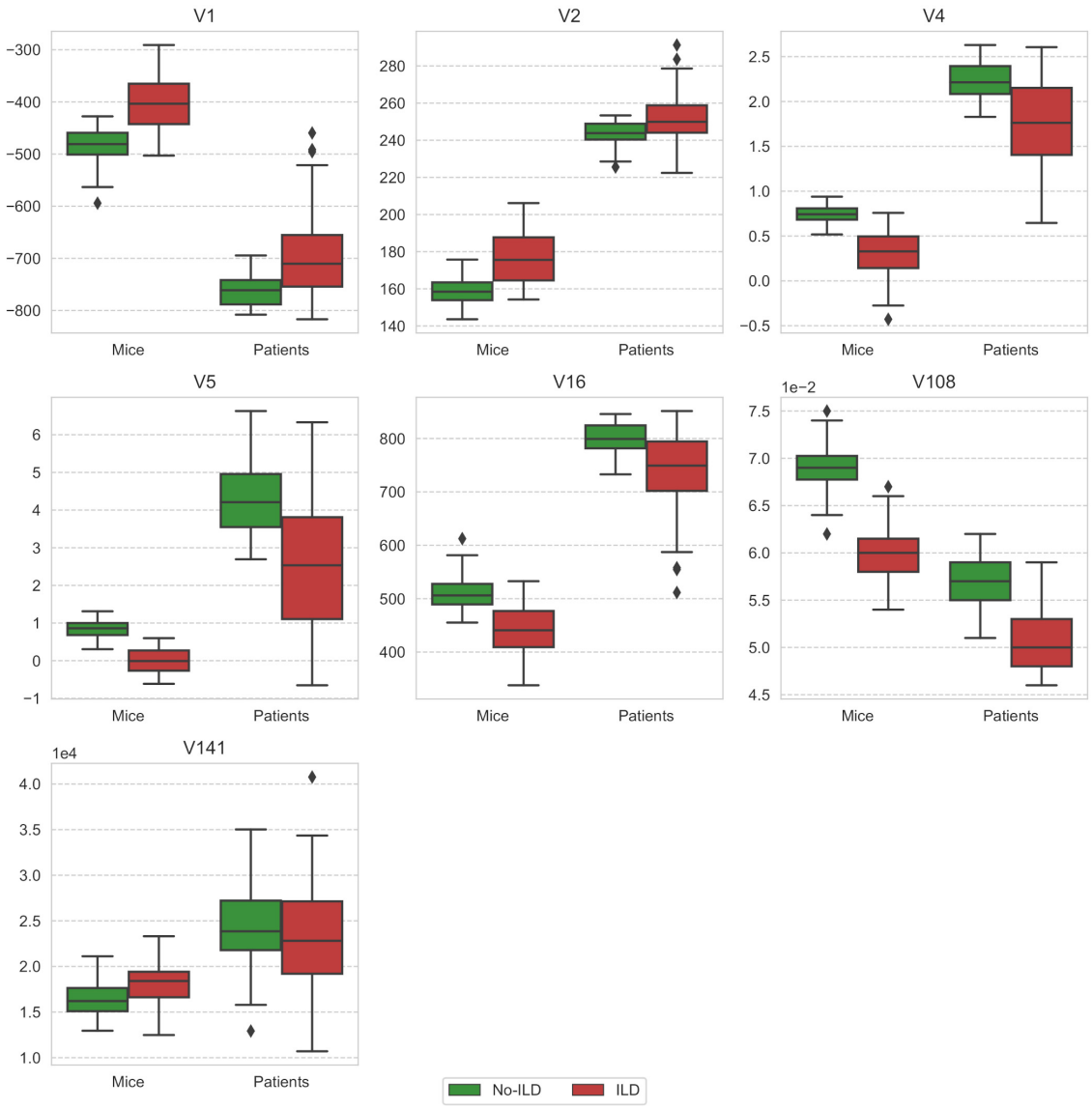

FIGURE 6
Comparison of feature distribution between mice and patient groups stratified by the ILD stage. V1 - mean (histogram), V2 - standard deviation (histogram), V4 - skewness (histogram), V5 - kurtosis (histogram), V16 - root mean square (histogram), V108 - gray level non-uniformity normalized (GLSZM), V141 - dependence count non-uniformity (NGLDM).

representative animal models could represent valuable systems for defined hypothesis testing in radiomics research, particularly for evaluating links with pathophysiology or studying responses to targeted therapies in rare diseases with low number of patients and limited access to tissue samples.

Radiomic features proved to be highly indicative of experimental- and SSc-ILD. Furthermore, we observed strong linear correlation in terms of discriminative power between features extracted from mice microCT scans and patient HRCT. We also showed that multivariate models of ILD translated well from mice to patient data sets. Nevertheless, we observed the differences between the data sets in terms of feature classes that were predictive. In mice, most of the feature groups contained features that reached similar maximum AUC scores. On the other hand, in patients we observed that even though histogram-based features achieved high discriminative power, some texture features were more predictive. This difference could be caused by inferior quality of microCT compared to HRCT. For this reason, the assessment of microCT done by our radiologist might have also been mainly led by first order characteristics rather than texture. Furthermore, the ILD manifestations can differ depending on the etiology. As a result, the observed differences may be caused by the limitation of the bleomycin-induced ILD being an imperfect model of SSc-ILD. In any case,

our results are in line with the available literature on human lung pathologies including chronic obstructive pulmonary disease, radiation-induced pneumonitis or connective tissue disease-related ILD, which showed that texture-based analysis of CT data can be superior compared to the visual or histogram-based measures for diagnosis (28, 31, 32).

Analysis of feature weights in the MEAN and the MSSK models showed that higher values of the mean and standard deviation of the image intensity and lower values of skewness and kurtosis correspond to larger risk of ILD. Effectively, this means that presence of ILD shifts the intensity distribution from a typical "healthy" positively skewed intensity distribution toward higher intensity values with a more symmetric distribution and thin tails. The best performing model (ML) relied on three radiomic features: the root mean square (histogram), gray level non-uniformity normalized (GLSZM), and dependence count non-uniformity (NGLDM). Significant improvement of machine learning models by re-optimization may suggest the existence of similar predictive radiomic patterns in training (mice) and test (patients) data sets in presence of domain shift between both groups.

The presented study has a few limitations. First, the differences in scanning parameters between microCT and HRCT cause a significant domain shift between experimental and patient data sets. Although, we were able to recover the predictive power of the analyzed multivariate models by re-optimization in the patient cohort, and by that confirm transferability of the underlying radiomic signatures, better calibration of the microCT scanner and selection of scanning parameters could potentially improve the transferability. Second, our study focused on CT-derived radiomics approaches, since HRCT scans are part of the routine work-up of ILD patients. Other imaging modalities such as nuclear imaging or MRI, although currently rarely performed in ILD (10), could be evaluated for radiomic analyses to assess whether they might provide additional or complementary information.

Conclusion

Radiomic signatures of experimental ILD derived from microCT scans translated as prognostic factors to HRCT of SSc-ILD. By this we showed that the well-established experimental model of BLM-induced ILD is a valuable system to test defined hypotheses in radiomics research for later validation in human cohorts.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the local ethics committees (approval numbers: pre-BASEC-EK-839 (KEK-no.-2016-01515), KEK-ZH-no. 2010-158/5, BASEC-no. 2018-02165, and BASEC-no. 2018-01873). The patients/participants provided their written informed consent to participate in this study. This animal study was reviewed and approved by the cantonal authorities and performed in compliance with the Swiss law of animal protection (ZH235-2018).

Author contributions

HG, JG-S, MG, BM, and ST-L contributed to the conception and design of the study. HG, JG-S, MBr, CB, and TF contributed to the acquisition and analysis of the data. HG, JG-S, BM, and ST-L contributed to the interpretation of data. HG and MBo contributed to the creation of software used in the study. All authors have drafted the work or substantively revised it.

Funding

This work was supported by the Forschungskredit PostDoc from University of Zurich (FK-19-046 to JG-S) and Swiss National Fund (SNF 310030_170159 to HG).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.988927/full#supplementary-material

References

1. Hutchinson JP, McKeever TM, Fogarty AW, Navaratnam V, Hubbard RB. Increasing global mortality from idiopathic pulmonary fibrosis in the twenty-first century. *Ann Am Thorac Soc.* (2014) 11:1176–85. doi: 10.1513/AnnalsATS.201404-145OC

2. Wallace B, Vummidi D, Khanna D. Management of connective tissue diseases associated interstitial lung disease: a review of the published literature. *Curr Opin Rheumatol.* (2016) 28:236–45.

3. John Gibson G, Loddenkemper R, Sibille Y, Lundbäck B. *The European Lung White Book: Respiratory Health and Disease in Europe*. Lausanne: European Respiratory Society (2013).

4. Carrington R, Jordan S, Pitchford SC, Page CP. Use of animal models in IPF research. *Pulm Pharmacol Ther.* (2018) 51:73–8.

5. Tashiro J, Rubio GA, Limper AH, Williams K, Elliot SJ, Ninou I, et al. Exploring animal models that resemble idiopathic pulmonary fibrosis. *Front Med.* (2017) 4:118. doi: 10.3389/fmed.2017.00118

6. Schniering J, Guo L, Brunner M, Schibli R, Ye S, Distler O, et al. Evaluation of Tc-rhAnnexin V-128 SPECT/CT as a diagnostic tool for early stages of interstitial lung disease associated with systemic sclerosis. *Arthritis Res Ther.* (2018) 20:183. doi: 10.1186/s13075-018-1681-1

7. Zhou Y, Chen H, Ambalavanan N, Liu G, Antony VB, Ding Q, et al. Noninvasive imaging of experimental lung fibrosis. *Am J Respir Cell Mol Biol.* (2015) 53:8–13.

8. Silver KC, Silver RM. Management of systemic-sclerosis-associated interstitial lung disease. *Rheum Dis Clin North Am.* (2015) 41:439–57.

9. Collins BF, Raghu G. Idiopathic pulmonary fibrosis: How should a confident diagnosis be made? In: Thillai M, Moller DR, Meyer KC editors. *Clinical Handbook of Interstitial Lung Disease*. (Boca Raton, FL: CRC Press) (2017). p. 135–48.

10. Hansell DM, Goldin JG, King TE, Lynch DA, Richeldi L, Wells AU. CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the fleischner society. *Lancet Respirat Med.* (2015) 3:483–96. doi: 10.1016/S2213-2600(15)00096-X

11. Wells AU, Desai SR, Rubens MB, Goh NSL, Cramer D, Nicholson AG, et al. Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography. *Am J Respir Crit Care Med.* (2003) 167:962–9. doi: 10.1164/rccm.2111053

12. Aichler M, Kunzke T, Buck A, Sun N, Ackermann M, Jonigk D, et al. Molecular similarities and differences from human pulmonary fibrosis and corresponding mouse model: MALDI imaging mass spectrometry in comparative medicine. *Lab Invest.* (2018) 98:141–9. doi: 10.1038/labinvest.2017.110

13. Schniering J, Benešová M, Brunner M, Haller S, Cohrs S, Frauenfelder T, et al. F-AzaFol for detection of folate receptor-β positive macrophages in experimental interstitial lung disease-a proof-of-concept study. *Front Immunol.* (2019) 10:2724. doi: 10.3389/fimmu.2019.02724

14. Schniering J, Benešová M, Brunner M, Haller S, Cohrs S, Frauenfelder T, et al. Visualisation of interstitial lung disease by molecular imaging of integrin αvβ3 and somatostatin receptor 2. *Ann Rheum Dis.* (2019) 78:218–27.

15. Schniering J, Gabrys H, Brunner M, Distler O, Guckenberger M, Bogowicz M, et al. Computed-tomography-based radiomics features for staging of interstitial lung disease – transferability from experimental to human lung fibrosis - a proof-of-concept studyImaging. *Eur Respir Soc.* (2019) 54:PA4806. doi: 10.1183/13993003.congress-2019.pa4806

16. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* (2012) 48:441–6.

17. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* (2017) 14:749–62. doi: 10.1038/nrclinonc.2017.141

18. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* (2014) 5:4006.

19. Lu H, Arshad M, Thornton A, Avesani G, Cunnea P, Curry E, et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. *Nat Commun.* (2019) 10:764. doi: 10.1038/s41467-019-08718-9

20. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol.* (2016) 6:71. doi: 10.3389/fonc.2016.00071

21. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife.* (2017) 6:e23421. doi: 10.7554/eLife.23421

22. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep.* (2015) 5:11044. doi: 10.1038/srep11044

23. Leijenaar RT, Bogowicz M, Jochems A, Hoebers FJ, Wesseling FW, Huang SH, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *Br J Radiol.* (2018) 91:20170498. doi: 10.1259/bjr.20170498

24. Aerts HJWL, Grossmann P, Tan Y, Oxnard GR, Rizvi N, Schwartz LH, et al. Defining a Radiomic response phenotype: a pilot study using targeted therapy in NSCLC. *Sci Rep.* (2016) 6:33860.

25. Schniering J, Maciukiewicz M, Gabrys HS, Brunner M, Blüthgen C, Meier C, et al. Computed tomography-based radiomics decodes prognostic and molecular differences in interstitial lung disease related to systemic sclerosis. *Eur Respir J.* (2021) 59:2004503. doi: 10.1183/13993003.04503-2020

26. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med.* (2018) 6:837–45. doi: 10.1016/S2213-2600(18)30286-8

27. Humphries SM, Yagihashi K, Huckleberry J, Rho B-H, Schroeder JD, Strand M, et al. Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. *Radiology.* (2017) 285:270–8. doi: 10.1148/radiol.2017161177

28. Kloth C, Blum AC, Thaiss WM, Preibsch H, Ditt H, Grimmer R, et al. Differences in texture analysis parameters between active alveolitis and lung fibrosis in chest CT of patients with systemic sclerosis: a feasibility study. *Acad Radiol.* (2017) 24:1596–603. doi: 10.1016/j.acra.2017.07.002

29. Kloth C, Henes J, Xenitidis T, Thaiss WM, Blum AC, Fritz J, et al. Chest CT texture analysis for response assessment in systemic sclerosis. *Eur J Radiol.* (2018) 101:50–8. doi: 10.1016/j.ejrad.2018.01.024

30. Lee SM, Seo JB, Oh SY, Kim TH, Song JW, Lee SM, et al. Prediction of survival by texture-based automated quantitative assessment of regional disease patterns on CT in idiopathic pulmonary fibrosis. *Eur Radiol.* (2018) 28:1293–300.

31. Sorensen L, Nielsen M, Lo P, Ashraf H, Pedersen JH, de Bruijne M. Texture-based analysis of COPD: a data-driven approach. *IEEE Trans Med Imaging.* (2012) 31:70–8. doi: 10.1109/TMI.2011.2164931

32. Cunliffe AR, Armato SG III, Straus C, Malik R, Al-Hallaq HA. Lung texture in serial thoracic CT scans: correlation with radiologist-defined severity of acute changes following radiation therapy. *Phys Med Biol.* (2014) 59:5387–98. doi: 10.1088/0031-9155/59/18/5387

33. Eresen A, Yang J, Shangguan J, Li Y, Hu S, Sun C, et al. MRI radiomics for early prediction of response to vaccine therapy in a transgenic mouse model of pancreatic ductal adenocarcinoma. *J Transl Med.* (2020) 18:61.

34. Eresen A, Yang J, Shangguan J, Benson AB, Yaghmai V, Zhang Z. Detection of immunotherapeutic response in a transgenic mouse model of pancreatic ductal adenocarcinoma using multiparametric MRI radiomics: a preliminary investigation. *Acad Radiol.* (2021) 28:e147–54. doi: 10.1016/j.acra.2020.04.026

35. Núñez LM, Romero E, Julià-Sapé M, Ledesma-Carbayo MJ, Santos A, Arús C, et al. Unraveling response to temozolomide in preclinical GL261 glioblastoma with MRI/MRSI using radiomics and signal source extraction. *Sci Rep.* (2020) 10:19699.

36. Becker AS, Schneider MA, Wurnig MC, Wagner M, Clavien PA, Boss A. Radiomics of liver MRI predict metastases in mice. *Eur Radiol Exp.* (2018) 2:11. doi: 10.1186/s41747-018-0044-7

37. Ni M, Wang L, Yu H, Wen X, Yang Y, Liu G, et al. Radiomics approaches for predicting liver fibrosis with nonenhanced T1 -weighted imaging: comparison of different radiomics models. *J Magn Reson Imaging.* (2021) 53:1080–9.

38. Schniering J, Borgna F, Siwowska K, Benešová M, Cohrs S, Hasler R, et al. In vivo labeling of plasma proteins for imaging of enhanced vascular permeability in the lungs. *Mol Pharm.* (2018) 15:4995–5004.

39. Minier T, Guiducci S, Bellando-Randone S, Bruni C, Lepri G, Czirják L, et al. EUSTAR co-workers, Preliminary analysis of the very early diagnosis of systemic sclerosis (VEDOSS) EUSTAR multicentre study: evidence for puffy fingers as a pivotal sign for suspicion of systemic sclerosis. *Ann Rheum Dis.* (2014) 73:2087–93.

40. van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Arthritis Rheum.* (2013) 65:2737–47.

41. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv* [Preprint] (2016). arXiv:1612.07003 [cs.CV].

42. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* (1979) 86:420–8.

43. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng*. (2007) 9:99–104.

44. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: A structure for efficient numerical computation. *Comput Sci Eng*. (2011) 13:22–30.

45. McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX (2010). p. 51–6.

46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. (2011) 12:2825–30.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fmed.2022.1083474

OPEN ACCESS

EDITED BY
Jingjing You,
The University of Sydney, Australia

REVIEWED BY
Cristoforo Pomara,
University of Catania, Italy
Cristina Mondello,
University of Messina, Italy

*CORRESPONDENCE
Jun-hong Sun
junhong.sun@sxmu.edu.cn

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 29 October 2022
ACCEPTED 19 December 2022
PUBLISHED 10 January 2023

CITATION
Du Q-x, Zhang S, Long F-h, Lu X-j, Wang L, Cao J, Jin Q-q, Ren K, Zhang J, Huang P and Sun J-h (2023) Combining with lab-on-chip technology and multi-organ fusion strategy to estimate post-mortem interval of rat.
Front. Med. 9:1083474.
doi: 10.3389/fmed.2022.1083474

Combining with lab-on-chip technology and multi-organ fusion strategy to estimate post-mortem interval of rat

Qiu-xiang Du[1,2], Shuai Zhang[2], Fei-hao Long[2], Xiao-jun Lu[3], Liang Wang[4], Jie Cao[2], Qian-qian Jin[2], Kang Ren[2], Ji Zhang[1], Ping Huang[1] and Jun-hong Sun[1,2]*

[1]Shanghai Key Laboratory of Forensic Medicine, Academy of Forensic Science, Shanghai, China, [2]School of Forensic Medicine, Shanxi Medical University, Jinzhong, Shanxi, China, [3]Criminal Investigation Detachment, Baotou Public Security Bureau, Baotou, Inner Mongolia, China, [4]National Center for Liver Cancer, Second Military Medical University, Shanghai, China

Background: The estimation of post-mortem interval (PMI) is one of the most important problems in forensic pathology all the time. Although many classical methods can be used to estimate time since death, accurate and rapid estimation of PMI is still a difficult task in forensic practice, so the estimation of PMI requires a faster, more accurate, and more convenient method.

Materials and methods: In this study, an experimental method, lab-on-chip, is used to analyze the characterizations of polypeptide fragments of the lung, liver, kidney, and skeletal muscle of rats at defined time points after death (0, 1, 2, 3, 5, 7, 9, 12, 15, 18, 21, 24, 27, and 30 days). Then, machine learning algorithms (base model: LR, SVM, RF, GBDT, and MLPC; ensemble model: stacking, soft voting, and soft-weighted voting) are applied to predict PMI with single organ. Multi-organ fusion strategy is designed to predict PMI based on multiple organs. Then, the ensemble pruning algorithm determines the best combination of multi-organ.

Results: The kidney is the best single organ for predicting the time of death, and its internal and external accuracy is 0.808 and 0.714, respectively. Multi-organ fusion strategy dramatically improves the performance of PMI estimation, and its internal and external accuracy is 0.962 and 0.893, respectively. Finally, the best organ combination determined by the ensemble pruning algorithm is all organs, such as lung, liver, kidney, and skeletal muscle.

Conclusion: Lab-on-chip is feasible to detect polypeptide fragments and multi-organ fusion is more accurate than single organ for PMI estimation.

KEYWORDS

forensic pathology, machine learning, multi-organ fusion, lab-on-chip, post-mortem interval

1. Introduction

Post-mortem interval (PMI), also called time since death, is the elapsed time between the death of an organism and the initiation of an official investigation (1). It is very important for the investigation of death in civil and criminal cases to accurately infer the time of death, such as civil investigation of life insurance fraud, identifying the victim and suspect, and accepting or rejecting the suspect's alibi (2). Traditional inference methods of PMI are usually based on corpse temperature (3) and early corpse phenomena such as livor mortis (4), rigor mortis (5), and post-mortem turbidity of cornea (6); it is difficult to precisely confirm the time since death, because these methods are rough, subjective, and empirical, as well as are greatly affected by environmental factors (7).

With the development of biomolecular technology, detection methods based on nucleic acid (1, 8, 9), metabolites (10, 11), and microorganisms (2, 12, 13) have been widely used in the past few decades. Some studies suggested that the genes, such as GAPDH2, ACTB2, 18S rRNA, miR-1, and miR-133a, are suitable indicators for estimating PMI (14–16). The level of the metabolite, which was detected by nuclear magnetism, mass spectrometry, and spectrograph, also provided a new direction for PMI inference at the tissues level (17–20). A further investigation into microorganisms of human and animal remains to study microbial community succession after death (21–24). In addition, with the development of imaging technology, post-mortem computed tomography (25), microCT (26), and visible and thermal 3D imaging (27) have also been used to infer the time since death. These technologies provide valuable ideas and methods for PMI estimation in forensic practice.

Protein is one of the biological macromolecules, an essential component of the organism, and participates in every cellular process. In recent years, proteins, in particular, have been evaluated for their potential to aid PMI delimitation. Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE)/western blotting (28, 29), immunohistochemistry (30, 31), and mass spectrometry (32, 33) were widely used to estimate the time since death. Although these approaches have shown some success and promise, there are certain limitations with these existing approaches, e.g., tedious operations, money-wasting, and slow. More importantly, there is no mature method to predict PMI accurately.

In the present study, a new experimental method, called lab-on-chip, is used to analyze protein and its degradation fragments, i.e., polypeptides. This method utilizes the Agilent 2,100 Bioanalyzer in combination with the protein LabChip kit, which simplifies the process of bioanalytical investigation and provides a system with standardized analysis handling and data processing (34). Although lab-on-chip cannot identify a polypeptide as a particular protein, the technology has been proven to be available for examining snake venom composition (35) and soybean cultivars in previous studies (36). It can perform molecular mass, migration time, peak height, peak area, relative concentration, and percentage of overall protein content and generate complete multi-peak spectrums of a sample. In addition, lab-on-chip is fast with minimal sample consumption, high throughput, and automatic quantitation (37), which means it is more appropriate for estimating PMI in practical work. At the same time, the abovementioned advantages also contribute to the united use of lab-on-chip and machine learning.

In the past decades, most studies have applied a single organ, such as the degradation of rat muscle proteins, used to estimate PMI by Zissler et al. (38). Although the two organs were used to estimate the time since death in the study by Mona Mohamed Abo El-Noor, the results of the heart and kidney were not analyzed jointly (39). In recent years, researchers from other fields have discovered that multi-organ fusion based on machine learning is more helpful to cancer diagnosis (40) and preclinical drugs than single organ (41). Hence, it is a beneficial trial that exploits multi-organ fusion and machine learning in estimating PMI. In the current study, lab-on-chip will analyze the polypeptide fragments in the lung, liver, kidney, and skeletal muscle of rat after death. We compare the performance of machine learning based on single and multiple organs to estimate the time since death and obtain the best prediction model based on multiple organs, which provides a new idea for forensic death time estimation.

2. Materials and methods

This study's workflow mainly involves the following (**Figure 1**). (1) Lab-on-chip analysis of the post-mortem degradation of polypeptides from the lung, liver, kidney, and skeletal muscle of rat at defined time points; (2) Base models (LR, SVM, RF, GBDT, and MLPC) and ensemble models (stacking, soft voting, and soft-weighted voting) evaluate the single organ's performances to predict PMI; and (3) The ensemble model based on a multi-organ fusion strategy evaluates multi-organ performances to predict PMI.

2.1. Equipment, reagents, and supplies

A two-place balance (AX223ZH/E, OHAUS, China), vortex finder (VXMNFS, OHAUS, China), thermocell mixing block (MSC-100, Aosheng, China), heraeus sepatech (2-16PK, Sartorius, Germany), climate chamber (RX2-260B, Ningbo, China), and Agilent 2,100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) were used.

Deionized water for protein extraction, a Agilent Protein 230 LabChip® kit (Agilent Technologies, CA, USA), and dithiothreitol (DTT, 1 M; Solarbio, Beijing, China) were used for the preparation of denaturant.

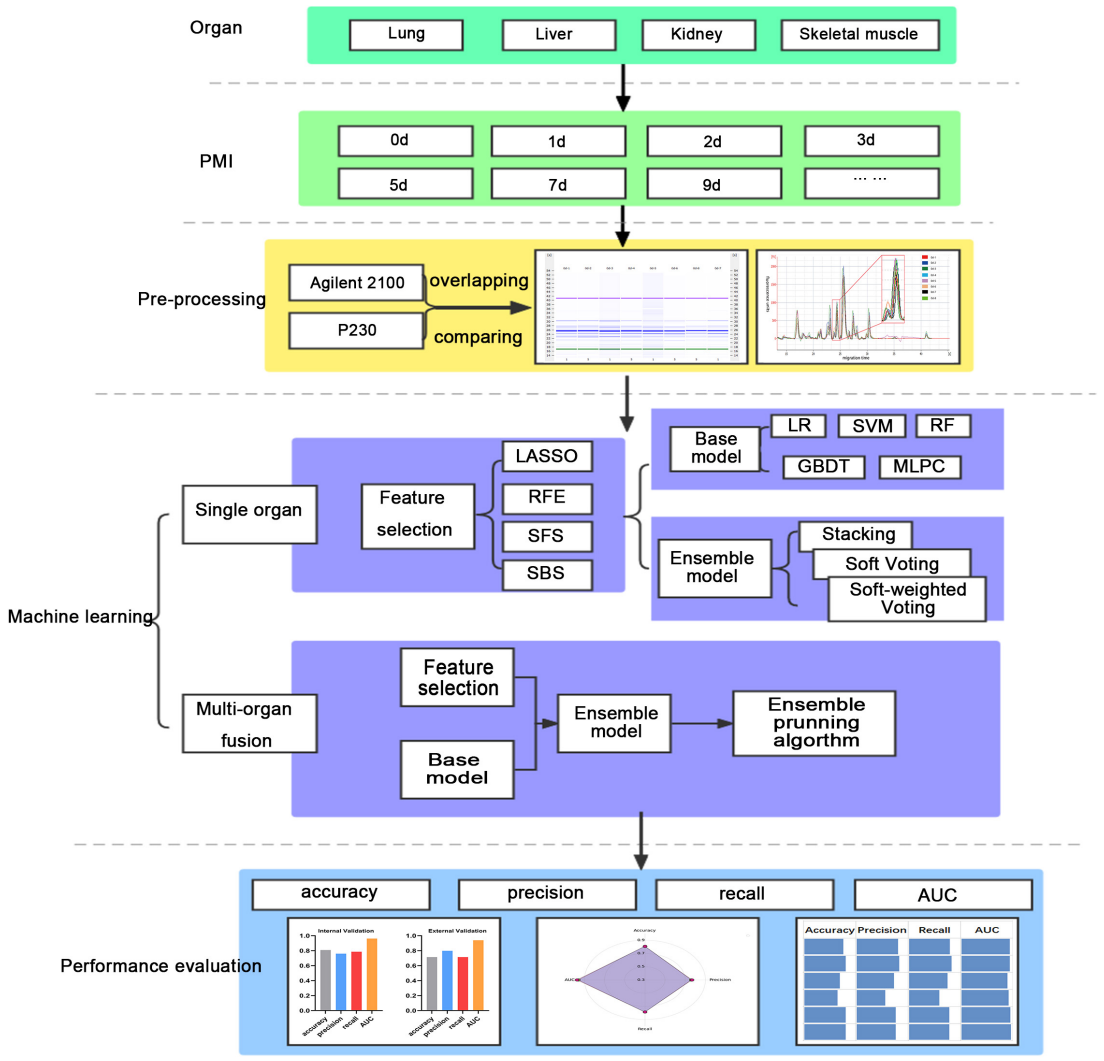

FIGURE 1
The workflow of this study.

2.2. Animal sample

This study was approved by the Institutional Animal Care and Use Committee of Shanxi Medical University. Animals received humane care in conformity with the principles in the Guide for the Care and Use of Laboratory Animals protocol, published by the Ministry of the People's Republic of China. This study was carried out in compliance with the ARRIVE guideline and evaluated and approved by the Institutional Animal Care and Use Committee of the Shanxi Medical University of China.

A total of 84 healthy male Sprague–Dawley rats, 10–12 weeks, weighing 200–230 g (provided by Animal Center of Shanxi Medical University) were housed in a cage with rat chow and water under a 12-h light–dark cycle at 22–25°C at a relative humidity of 40–60%. After 2 days, rats were sacrificed after pentobarbital anesthetization *via* cervical dislocation. The lung, liver, kidney, and right hind limb gastrocnemius muscle of each rat were harvested (200 mg ± 2 mg) at the fixed time points of 0, 1, 2, 3, 5, 7, 9, 12, 15, 18, 21, 24, 27, and 30 days (n = 6 rats) after sacrifice, and a total of 336 samples were placed in liquid nitrogen for quick freezing and stored at −80°C until analysis.

For external validation, 28 rats were taken according to the methods of the abovementioned experimental process. Each time point took two rats.

2.3. Water-soluble protein extraction and samples preparation

Analysis was performed according to the protocol provided by the manufacturer. A volume of 200 mg of the lung, liver, kidney, and skeletal muscle tissues were ground, added to

deionized water containing 1% phenylmethylsulfonyl fluoride (PMSF) according to the ratio of 1:3.5 (w/v), then incubated on ice for 60 min, and centrifuged at 12,000 × *g* (15 min, 4°C). A volume of 4 μl solution per sample was diluted by mixing with 2 μl of the sample buffer with a reducing agent (DTT). The diluted solution and ladder (Agilent) were heated for 5 min at 95°C and then diluted with 84 μl H_2O. Samples and ladder were loaded on the protein chip and measured immediately. To confirm the protein extraction process or the protein analysis process by lab-on-chip had avoided errors as much as possible, quality control samples were prepared.

2.4. Microfluidic LoaC electrophoresis

The protein profile of rat skeletal muscle using microfluidic capillary gel electrophoresis with laser-induced fluorescence (LIF) detection was carried out on the Bioanalyzer Agilent 2,100 using the Protein 230 Kit (Agilent Technologies, Waldbronn, Germany), which allows the separation of proteins from 14 to 230 kDa. According to the protocol, 4 μl of each tissue sample was mixed with 2 μl denaturing solution (35 mM dithiothreitol) in 0.5-ml tubes and denatured at 100°C for 5 min, incubated in ice for 2 min, and centrifuged for 15 s. Pure water was added to 100 μl, and samples were vortexed. Then, 6 μl of samples were added to each well of the chip. For the analysis, three biological replicates were used for each sample.

All reagents were provided with each LabChip kit, including the standard protein ladder containing different proteins with known concentrations and molecular weights that can be used for semi-quantitative analysis. The Agilent 2,100 Bioanalyzer separates and calculates the protein fragments based on the microfluidic capillary gel electrophoresis with LIF detection, where fluorescence intensities of proteins are measured. The migration times of polypeptide fragments were used to estimate the respective protein bands' molecular weights, and the height was calculated to semi-quantify each protein fragment's concentration. Data analysis performed with the Agilent 2,100 Expert software automatically determines molecular weight, concentration, and percentage of the sample's total individual proteins.

2.5. Confirmation of polypeptide fragments and data preprocessing

All protein electrophoresis chromatography analyses were performed by "comparison" and "overlap" operations in the software to calibrate, identify, and adjust peaks according to the lower and upper markers. The same polypeptide fragments of each organ can be marked as the same number according to the molecular mass of these peptides, from minor to major. Numerical data such as protein molecular mass, peak height, and migration time are outputted for subsequent analysis.

It is essential to confirm the polypeptide fragments, which could be used as an indicator to estimate the PMI. The present study acquired the raw data through Agilent 2,100 Expert software, and all CSV data were imported into MS Excel. Then, the polypeptide fragments detected in five out of six biological replicate samples were identified as meaningful indicators for estimating PMI. The deviation of migration times less than 2% was considered the same polypeptide fragment in different samples.

Then, the datasets of each organ with 84 rats have been randomly divided into two, namely, the training dataset, which was made up of 70% of the dataset, and the testing dataset, also named internal validation, which comprised the remaining 30%, and standardized. For the external validation of 28 rats, the same data preprocessing was applied as mentioned earlier.

2.6. Machine learning

2.6.1. Feature importance evaluated for machine learning

Feature selection, or feature ranking, reduces data processing time and memory requirements for machine-learning algorithms to deal with the essential predictors. In the present study, feature importance was evaluated through the least absolute shrinkage and selection operator (LASSO) (42), recursive feature elimination (RFE) (43), sequential forward selection (SFS), and sequential back selection (SBS) (44, 45).

2.6.2. Sub-model training and evaluation for PMI using different organs

Five machine learning algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (2), Gradient Boosting Decision Tree (GBDT), and Multilayer Perceptron Classier (MLPC), were implemented to predict PMI in the present study. The robustness and efficiency of 20 sub-models according to the four feature selection methods cross-match five machine learning algorithms are analyzed for each organ. The performance comparison analysis was performed by sequencing accuracy, precision, recall, and area under the ROC curve (AUC) of internal and external validation according to the order from good to wrong. And then, the ranking scores of all metrics were summed for each sub-model. Finally, the optimal classification model was determined by comparing the scores of 20 sub-models of each organ. It should be noted that the principle of this scoring method is to combine internal and external verification and comprehensive consideration of multiple evaluation indicators. Therefore, we believe that the model with the highest score has the highest comprehensive efficiency, which means that the model may not be the best in all indicators.

2.6.3. Ensemble model development and evaluation for PMI based on single organ

Ensemble learning can improve the classifier's performance by combining the trained sub-models contribution to solving the same classification problem in some studies (46). In the present study, there are three ensemble models, namely, stacking (47), soft voting (48), and soft-weighted voting (49), used to estimate the PMI based on the single organ. The accuracy, precision, recall, and AUC were calculated separately.

2.6.4. Multi-organ fusion strategy and ensemble pruning algorithm

A framework that is suitable for multi-organ fusion analysis is proposed in this study. First, each organ's best combinations of feature selection methods and sub-models were combined into a pipe. Four pipelines are used as four sub-models to complete each organ's feature selection and PMI prediction. Then, four parallel pipelines were performed to predict PMI by the abovementioned three ensemble models. In this step, the four organs are fused to predict PMI. Finally, the ensemble models based on multi-organ fusion were compared with the optimal sub-models and ensemble models based on single organ.

After getting the best model, the ensemble pruning algorithm was applied to ensure the best combination of an organ. The ensemble pruning algorithm is a technique where the model starts with all possible members being considered and removes members from the ensemble until no further improvement is observed. This could be performed in a greedy manner where members are removed one at a time and only if their removal results in a lift in the performance of the overall ensemble.

3. Results

3.1. Characterization of polypeptide fragments after death

A total of 45 polypeptide fragments were identified with different migration times in the lung, liver, kidney, and skeletal muscle samples (**Table 1**). These polypeptide fragments may be highly correlated with the PMI, and 21, 22, 19, and 23 polypeptide fragments were found in the lung, liver, kidney, and skeletal muscle tissues, respectively (**Figure 2A**). Among these polypeptide fragments, 4 polypeptide fragments were detected in four organs, 7 polypeptide fragments were present in three organs, and 14 polypeptide fragments were present in two organs (**Figure 2B**). There were three polypeptide fragments specific to the kidney and lung but seven to the liver and skeletal muscle.

After further analysis of the data, we found that the content of the abovementioned polypeptide fragments was highly homogeneous in the samples with the same PMI

TABLE 1 The polypeptide fragments in the lung, liver, kidney, and skeletal muscle samples.

Polypeptide	Molecular mass (−X ± SD)	Migration time (−X ± SD)	Organs[a]
1	14.25 ± 0.45	20.68 ± 0.08	Lu[b], Li[c], K[d], M[e]
2	15.53 ± 0.30	20.95 ± 0.05	M
3	17.61 ± 0.34	21.34 ± 0.06	M
4	19.65 ± 0.62	21.81 ± 0.13	K
5	25.58 ± 0.73	21.94 ± 0.15	Li
6	22.62 ± 0.83	22.36 ± 0.16	Li, K
7	23.84 ± 0.38	22.56 ± 0.07	M
8	24.54 ± 0.37	22.66 ± 0.06	Lu, Li
9	25.63 ± 0.69	22.93 ± 0.15	K, M
10	26.83 ± 0.35	23.12 ± 0.08	Lu, Li, M
11	29.43 ± 0.50	23.57 ± 0.14	Lu, K
12	31.68 ± 0.64	23.91 ± 0.10	K, M
13	32.92 ± 0.75	24.08 ± 0.12	Lu, Li, K
14	35.47 ± 0.91	24.40 ± 0.10	Lu, Li, M
15	39.65 ± 0.94	25.02 ± 0.18	Lu, Li, K, M
16	43.68 ± 0.82	25.59 ± 0.11	Lu, Li, K, M
17	45.15 ± 1.21	25.86 ± 0.17	Li
18	47.34 ± 0.91	26.19 ± 0.12	K
19	50.20 ± 1.11	26.46 ± 0.13	Lu, Li
20	51.97 ± 0.47	26.69 ± 0.09	K, M
21	53.06 ± 0.65	26.86 ± 0.08	Li
22	55.43 ± 0.43	27.15 ± 0.04	Li
23	57.44 ± 1.10	27.32 ± 0.16	Lu, K, M
24	58.43 ± 1.11	27.54 ± 0.18	Li, K, M
25	62.89 ± 0.70	28.08 ± 0.08	M
26	71.58 ± 1.57	28.80 ± 0.27	Lu, Li, K
27	74.83 ± 1.49	28.99 ± 0.16	Lu, M
28	78.23 ± 1.25	29.29 ± 0.07	M
29	82.13 ± 1.08	29.68 ± 0.09	Li
30	84.57 ± 0.77	29.88 ± 0.07	K
31	85.91 ± 0.98	29.90 ± 0.09	Lu, M
32	91.34 ± 1.04	30.30 ± 0.09	Lu
33	93.95 ± 1.91	30.59 ± 0.16	Lu, Li, K, M
34	104.86 ± 1.01	31.59 ± 0.08	Li
35	111.46 ± 5.41	32.09 ± 0.52	M
36	121.23 ± 1.82	32.98 ± 0.17	M
37	123.73 ± 2.40	33.23 ± 0.20	K, M
38	131.39 ± 1.01	33.90 ± 0.09	Li
39	135.25 ± 2.04	34.20 ± 0.17	Lu, Li

(Continued)

TABLE 1 (Continued)

Polypeptide	Molecular mass (−X ± SD)	Migration time (−X ± SD)	Organs[a]
40	141.40 ± 1.35	34.71 ± 0.15	Lu, M
41	144.59 ± 2.03	35.04 ± 0.18	Li, M
42	154.94 ± 1.46	35.78 ± 0.11	Lu, K
43	179.02 ± 1.05	37.28 ± 0.07	Lu, Li, K
44	217.22 ± 0.98	39.60 ± 0.07	Lu
45	225.04 ± 3.44	40.08 ± 0.21	Lu

[a]Organs with polypeptide fragments.
[b]Lu represents lung.
[c]Li represents liver.
[d]K represents kidney.
[e]M represents skeletal muscle.

(Figures 2C, D). The results showed no significant difference among the biological replicates, providing that the experimental operation was stable and reliable. In addition, the polypeptide fragments showed different peak heights at different PMIs (Figures 2E, F), which highly correlated with PMI.

To further clarify the correlation between peptide fragment content and PMI, the earlier data were clustered using TB tools. It can be found from the clustering heat map that the death time of this experiment could be divided into five different stages according to the content of polypeptide fragments in the lung. Specifically, 0 and 3 days, 1 and 2 days, 5, 18, and 21 days, 7, 12, and 15 days, and 9, 24, 27, and 30 days were divided together (Figure 3A). Similarly, the samples can be distinguished into 5, 4, and 5 different periods according to the content of polypeptide fragments in the liver, kidney, and skeletal muscle (Figures 3B–D).

3.2. Performance of sub-models based on different organs

3.2.1. Evaluating the sub-models by accuracy, precision, recall, and AUC

To compare the predictive accuracy of four different organs in inferring the PMI, a total of 80 combined results were generated by cross-combining of four feature selection methods (e.g., LASSO, RFE, SBS, and SFS) and five machine learning algorithms (e.g., LR, SVM, RF, GBDT, and MLPC) (Figure 4A).

The accuracy, precision, recall, and AUC of sub-models with four organs are summarized in Figures 4B–E. As is shown in Figure 4B, the internal validation accuracy ranges of the lung, liver, kidney, and skeletal muscle were 0.462 (RFE + GBDT and SFS + GBDT)–0.769 (SBS + RF and SFS + RF), 0.231 (LASSO + SVM)–0.692 (SFS + RF), 0.577 (SFS + GBDT)–0.808 (LASSO + RF and SFS + RF), and 0.346 (RFE + GBDT)–0.769 (RFE + RF, SFS + SVM, and SFS + RF), respectively. Their external verification accuracies were 0.286 (SFS + GBDT)–0.679 (LASSO + RF and LASSO + MLPC), 0.179 (LASSO + MLPC)–0.536 (SFS + RF), 0.429 (SFS + GBDT)–0.714 (LASSO + RF and SFS + RF), and 0.321 (SFS + GBDT)–0.679 (RFE + RF, SBS + RF, and SFS + RF), respectively. Similarly, the analysis of Figures 4C–E shows that the model with the kidney as the detection sample performs best in precision, recall, and AUC evaluation indexes.

The abovementioned results indicated that the liver is the worst, and the kidney is the best to predict PMI among the four organs. As for the feature selection methods, the four feature selection methods cannot clearly distinguish the advantages and disadvantages. These results further show that LASSO, RFE, and SFS help determine feature subsets, which means that feature selection methods are necessary for different organs. It is particularly interesting that RF, the best machine learning algorithm in all organs, has advantages over other machine learning algorithms in predicting PMI, as mentioned earlier, while GBDT performed worst in the lung, kidney, and skeletal muscle. The four organs' remaining indicators were similar results (Figures 4B–E).

3.2.2. Screening optimal model by the ranking principle

The ranking scores principle described in the "Sub-models training and evaluation for PMI using different organs" section was used to compare the sub-models of each organ comprehensively. As is shown in Table 2, the best model combination in the lung and liver is SFS + RF, with scores of 146 and 149, respectively. The optimal sub-model of the kidney is LASSO + RF, which has a score of 149. The best sub-model of skeletal muscle is RFE + RF, which has a score of 139.

We found that the kidney is more suitable than other organs to predict PMI, comparing the performance of the best models for each organ. In optimal sub-models of four organs, 0.808 and 0.714 are the highest internal and external validation accuracies based on LASSO-RF of the kidney (Figure 5E), respectively. In Figure 5F, the confusion matrix of external verification of the kidney showed that eight samples were misjudged, and many miscalculations in the prediction results of the kidney were found at 0–2 days and 12–18 days after death. Next, the internal validation of the lung and skeletal muscle is 0.769 based on SFS-RF. The former's accuracy of external validation is 0.607 lower than the latter, which is 0.679 (Figures 5A, G). The liver is the worst organ to predict PMI; the accuracy is 0.692 and 0.536 in internal and external validation using SFS-RF, which is the best classification model for the liver (Figure 5C). As shown in Figures 5B, D, H, there are 11, 13, and 9 samples of the lung, liver, and skeletal muscle, respectively, which were wrongly judged in their external verification.

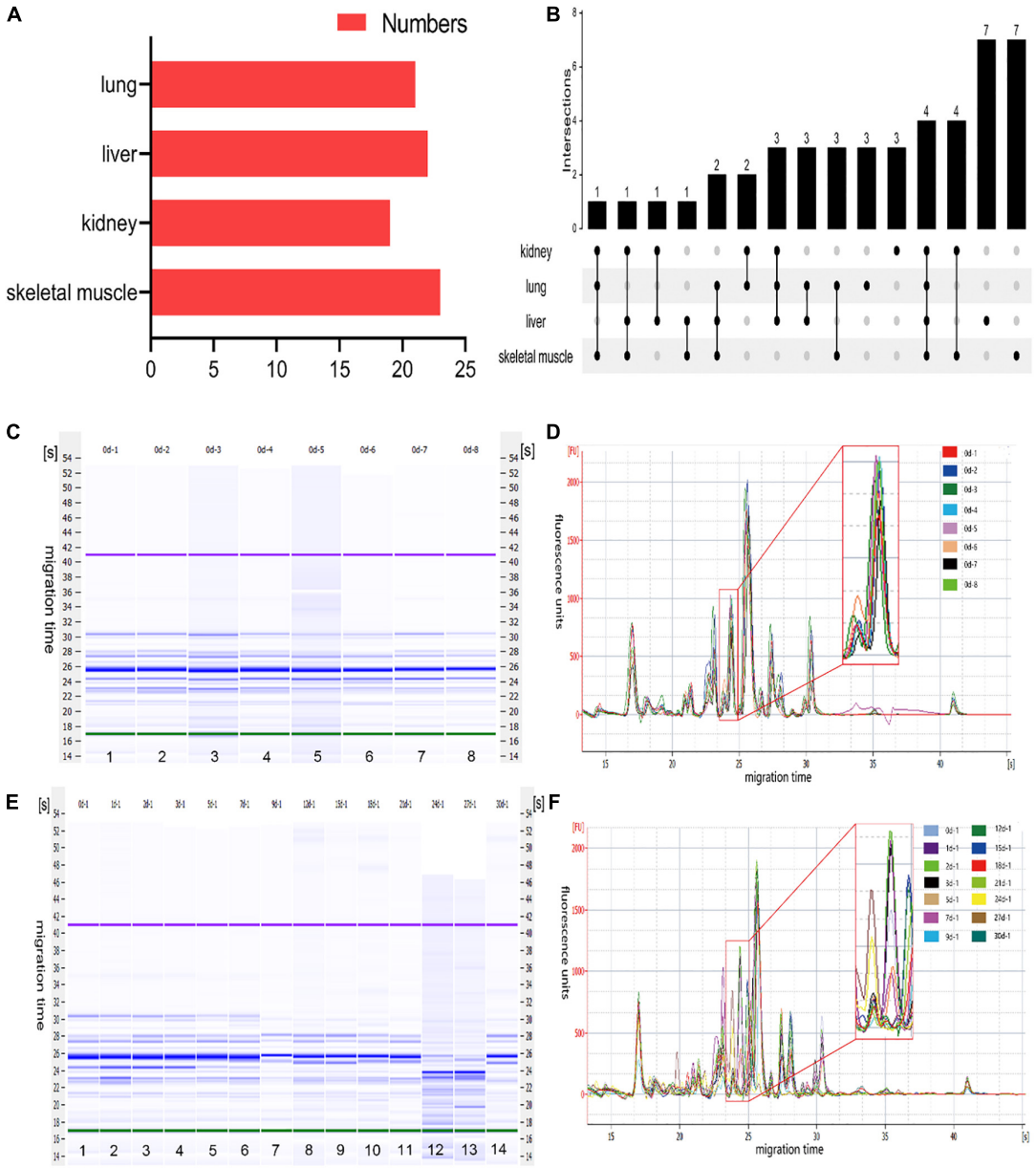

FIGURE 2
The characteristics of polypeptide fragments in different organs at different times after death. **(A)** The numbers of polypeptide fragments in different organs. **(B)** Co-expression analysis of polypeptide fragments in different organs. **(C)** The gel-like image of polypeptide fragments in skeletal muscle at the same time points after death. This figure is the simulated gel electrophoresis figure automatically given by Agilent 2,100 Bioanalyzer according to the molecular weight. Lanes 1–8 represent the gel diagram of eight skeletal muscle samples in 0 day after death. The migration time (s) is set on the side of the gel image. The purple bands at the top and the green bands at the bottom are the upper/lower ladder, which is the standard, respectively. The remaining blue bands are the detected protein fragments. The shade of the blue band represents the content of each protein fragment. **(D)** The electropherogram of skeletal muscle at the same time points after death in the microfluidic chip electrophoresis (LoaC) system. Multi-peak spectrums overlaid of different rats at the same time points after death, and there was no significant difference in peak height and number of peaks in the superposition of multi-peak spectra at the same time point after death of different rats. It is worth noting that there is a peptide peak around 24.5s in all samples at the same time point after death. **(E)** The gel-like image of polypeptide fragments in skeletal muscle at different time points after death. This figure is the simulated gel electrophoresis figure automatically given by Agilent 2,100 Bioanalyzer according to the molecular weight. Lanes 1–14 represent the gel diagram of 14 time points of skeletal muscle samples within 0–30 days after death. The migration time (s) is set on the side of the gel image. The purple bands at the top and the green bands at the bottom are the upper/lower ladder, which is the standard, respectively. The remaining blue bands are the detected protein fragments. The shade of the blue band represents the content of each protein fragment. **(F)** The electropherogram of skeletal muscle at different time points after death in the microfluidic chip electrophoresis (LoaC) system. The peak heights showed significant differences and a new peptide peak appears near 24.5 s by comparing multi-peak spectrums at different time points after death.

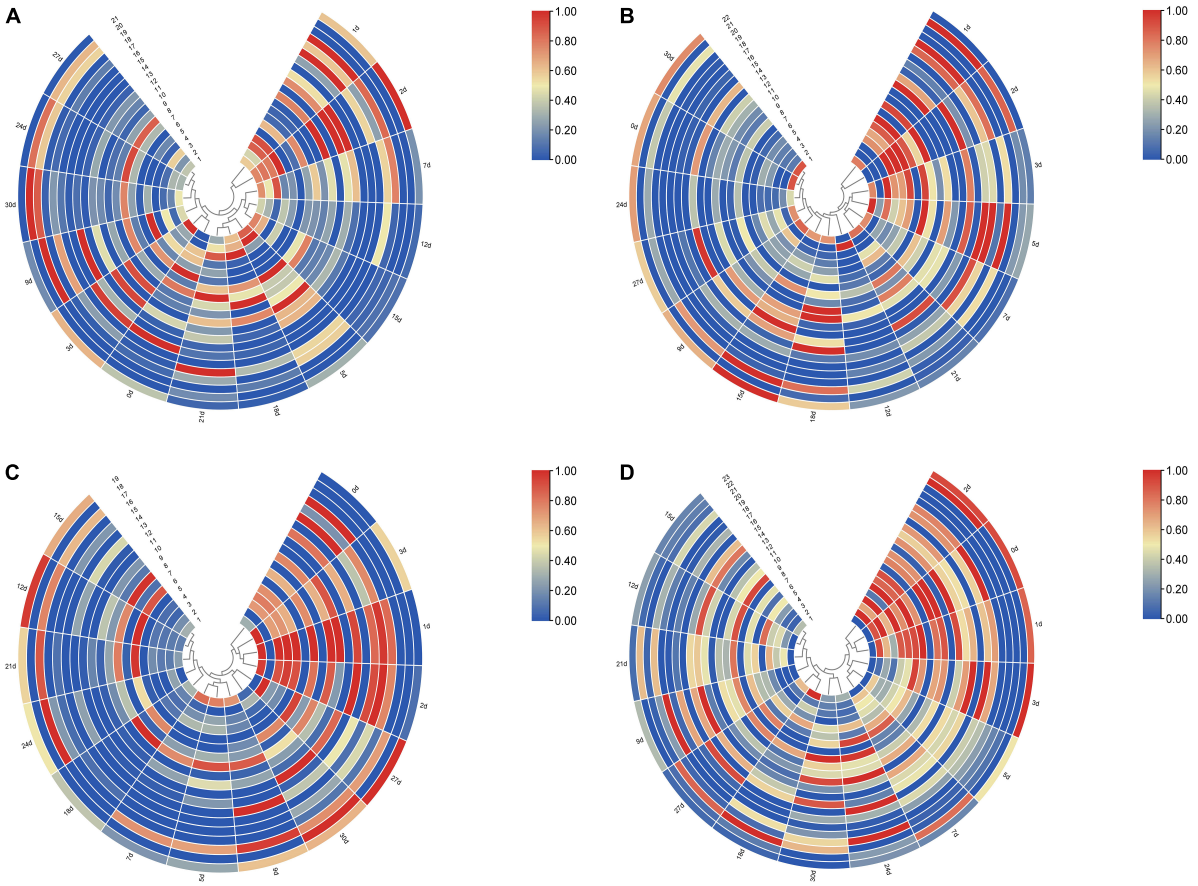

FIGURE 3
The clustering heat map based on the peak heights of polypeptide fragments in different organs. **(A)** Lung samples could be divided into five different stages, 0 and 3 days, 1 and 2 days, 5, 18, and 21 days, 7, 12, and 15 days, and 9, 24, 27, and 30 days were divided together, respectively. **(B)** Liver samples could be divided into five different stages, 1 and 2, 3, and 5 days, 7 and 21 days, 12 days, and 0, 9, 15, 18, 24, 27, and 30 days were divided together, respectively. **(C)** Kidney samples could be divided into four different stages, 0 to −3 days, 5, 7, and 9 days, 12, 15, 18, 21, and 24 days, and 27 and 30 days were divided together, respectively. **(D)** Skeletal muscle samples could be divided into five different stages, 0–2 days, 3, 5, and 7 days, 9, 12, 15, and 21 days, 18 and 27 days, and 24 and 30 days were divided together, respectively.

3.3. Performance of the single organ based on ensemble models

Considering that different prediction models have different prediction performances in four organs, this experiment will cross-combine the four feature selection methods and three ensemble models mentioned earlier to establish an ensemble model to improve the performance of PMI estimation in a single organ.

The performance of the ensemble models of four organs is shown in Figure 6A. In the validation of the lung, LASSO + soft-weighted voting generated the highest accuracy of 0.808 in the internal validation, while LASSO + soft voting generated the highest accuracy of 0.643 in the external validation. The best accuracy of internal validation based on the liver is 0.654, which was obtained by RFE + soft voting and RFE + soft-weighted voting. The accuracy of RFE + soft-weighted voting for external validation of the liver had reached 0.464. For kidneys, the accuracy for internal validation of LASSO + soft voting, LASSO + soft-weighted voting, SFS + soft voting, and SFS + soft-weighted voting was 0.808, while the optimal accuracy for external validation of LASSO + soft voting and RFE + stacking was 0.679. The highest accuracy for internal validation of skeletal muscle was 0.769, and the combined strategies were RFE + soft voting and RFE + soft-weighted voting, respectively. Furthermore, the external validation accuracy of SFS + soft voting and SFS + soft-weighted voting for skeletal muscle is 0.643. The details of the precision, recall, and AUC have similar results, as shown in Figure 6A.

According to the ranking principle described in the "Sub-models training and evaluation for PMI using different organs" section, the optimal ensemble model of each organ was screened in this experiment. The best ensemble model in the lung is LASSO + soft-weighted voting with ranking scores of 89, and

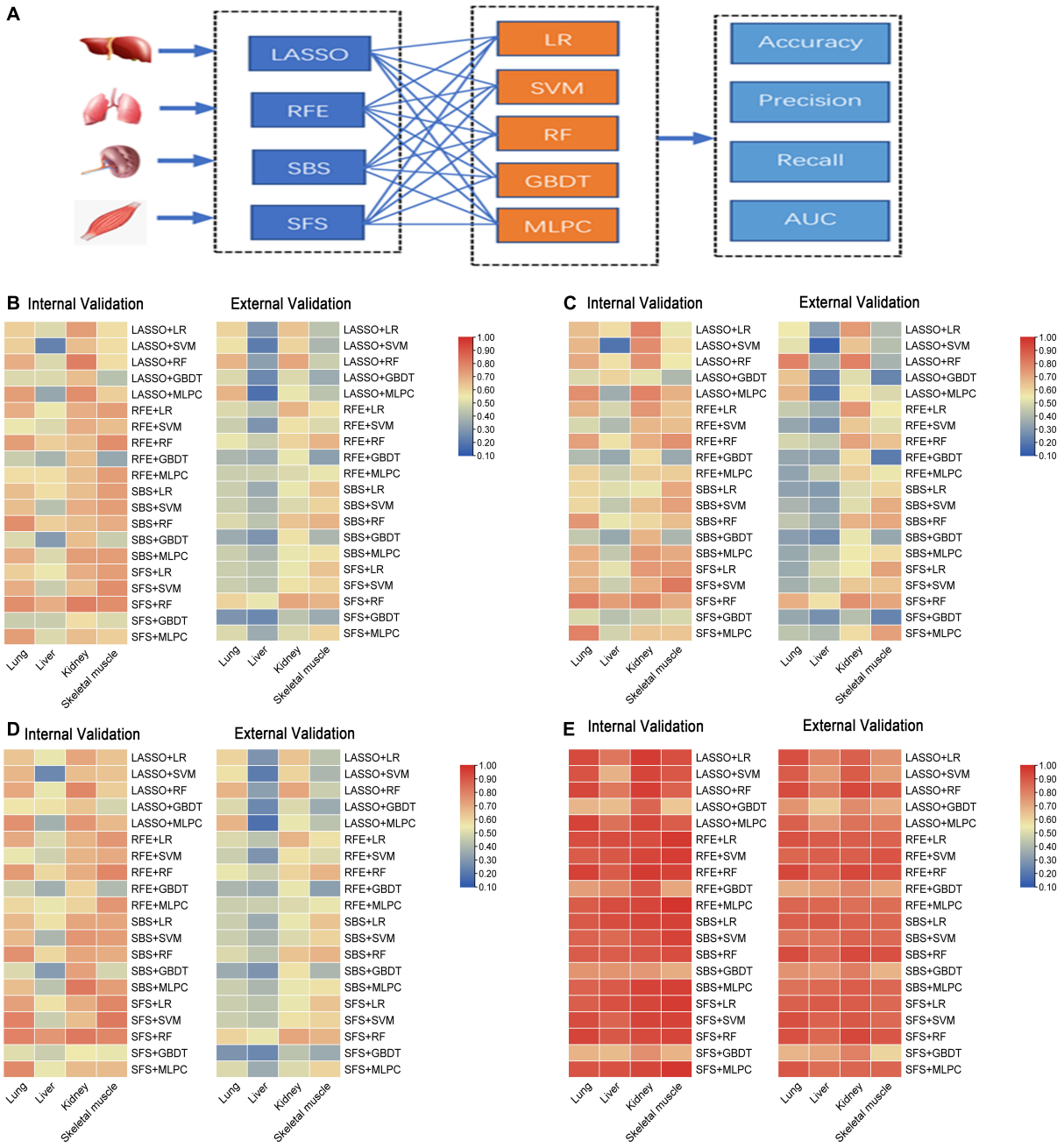

FIGURE 4
The performance of sub-models generated by cross-combination of four feature selection methods and five machine learning algorithms based on single organ. **(A)** Workflow of cross-combination of four feature selection methods and five machine learning algorithms to establish sub-models to predict PMI based on the lung, liver, kidney, and skeletal muscle. **(B)** The heat map on the left show accuracy of internal validation of sub-models based on lung, liver, kidney and skeletal muscle, and the heat map on the right shows the accuracy of external validation. **(C)** The heat map on the left show precision of internal validation of sub-models based on the lung, liver, kidney, and skeletal muscle, and the heat map on the right shows the precision of external validation. **(D)** The heat map on the left show recall of internal validation of sub-models based on the lung, liver, kidney, and skeletal muscle, and the heat map on the right shows the recall of external validation. **(E)** The heat map on the left shows AUC of internal validation of sub-models based on the lung, liver, kidney, and skeletal muscle, and the heat map on the right shows AUC of external validation.

the internal and external validation accuracies were 0.808 and 0.571, respectively (Table 3). Specifically, the optimal ensemble model of RFE + soft-weighted voting based on the liver was 89.5, and the internal and external validation accuracies were 0.654 and 0.464, respectively. The internal and external verification accuracies for the kidney are 0.808 and 0.679, respectively, based

TABLE 2 The scores of sub-models generated by cross-combination of four feature selection methods and five machine learning algorithms.

Model	Lung	Liver	Kidney	Skeletal muscle
LASSO + LR	107.5	76	128	50.5
LASSO + SVM	100	15	99.5	47.5
LASSO + RF	135.5	83	149	59
LASSO + GBDT	59.5	57.5	27	19.5
LASSO + MLPC	137	29	93.5	64.5
RFE + LR	104.5	113.5	110	107
RFE + SVM	48	70.5	92.5	93.5
RFE + RF	132	135.5	112	139
RFE + GBDT	18.5	48.5	46	14
RFE + MLPC	53	125	51.5	96
SBS + LR	64.5	109.5	65	115.5
SBS + SVM	60.5	71	62	111.5
SBS + RF	118	119	93.5	116.5
SBS + GBDT	20.5	32.5	65.5	25.5
SBS + MLPC	69	87.5	102	104.5
SFS + LR	71	130.5	83	130.5
SFS + SVM	99.5	96	91.5	122.5
SFS + RF	146	149	143	134.5
SFS + GBDT	19.5	37	11	22
SFS + MLPC	116	94	54.5	106.5

on LASSO + soft voting, which has the highest score of 76. The best ensemble model of skeletal muscle is SFS + soft-weighted voting, which scored 81, and the internal and external accuracy were 0.731 and 0.643, respectively.

In the present study, each organ's ensemble model was compared with the best sub-model of the same organ to determine whether the integrated model can improve the PMI prediction performance. Compared with SFS-RF, the best sub-model of the lung, although all metrics of internal validation are slightly improved, its external validation metrics significantly decreased according to LASSO + soft-weighted voting (**Figures 6B, C**). The RFE + soft-weighted voting model based on the liver predicts PMI with the most indicators lower than the best sub-model except for the AUC of internal and external validation (**Figures 6D, E**). Compared with the optimal kidney sub-model, the LASSO + soft voting model weakly improves the precision and AUC of internal validation (**Figures 6F, G**). By comparing SFS-soft-weighted voting with SFS-RF of skeletal muscle, the former only has feeble improvement in AUC of internal validation and precision of external validation (**Figures 6H, I**).

The abovementioned results indicated that the SFS + RF was the optimal model for predicting PMI based on the kidney. However, the single organ ensemble model could not effectively improve the PMI prediction performance. Therefore, in the multi-organ fusion based on ensemble model construction, the optimal sub-model performance will be compared with other models' performance in predicting PMI.

3.4. Performance of multi-organ fusion based on ensemble models

Since the single-organ ensemble strategy cannot improve the prediction efficiency of PMI, we further focus on the multi-organ integration strategy. **Figure 7A** shows the appropriate multi-organ fusion model establishment steps for estimating PMI. In brief, the best combinations of feature selection methods and sub-models in the lung, liver, kidney, and skeletal muscle were piped based on a multi-organ fusion strategy. Then, the ensemble model with the highest scores was selected by comparison. Finally, the ensemble pruning algorithm integrates multi-organ data based on the optimal model for PMI estimation.

By comparing the multi-organ integration model's internal and external verification accuracies, the soft voting fusion strategy has an absolute advantage with the internal and external verification accuracies of 0.962 and 0.893, respectively. In contrast, the staking model had the worst performance, and its internal and external validation accuracy is even lower than the single-organ optimal model based on the kidney, with only 0.692 and 0.679. The performance of soft-weighted voting was similar to that of soft voting, with internal and external validation accuracies of 0.923 and 0.893 (**Figures 7B, C**).

Although the AUC values of the internal and external validation of the three fusion strategies are all higher than 0.97, the confusion matrix results show that some samples are still misjudged according to the external validation (**Figures 7D–I**). The sample prediction error is mainly more than 15 days after death, indicating that if the prediction results show that the PMI exceeds 15 days, the prediction accuracy decreases and the credibility decreases.

The ensemble pruning algorithm showed that the optimal combination of multiple organs was four organs, i.e., lung, liver, kidney, and skeletal muscle, used in the present study to infer the PMI. Furthermore, soft voting and soft-weighted voting can significantly improve the prediction performance of PMI based on the multi-organ fusion strategy (**Table 4**).

3.5. Comparison of lab-on-chip and traditional protein detection methods

To further clarify the superiority of the analysis method and its application value in forensic practice, we summarize the main improvements of the proposed approach compared

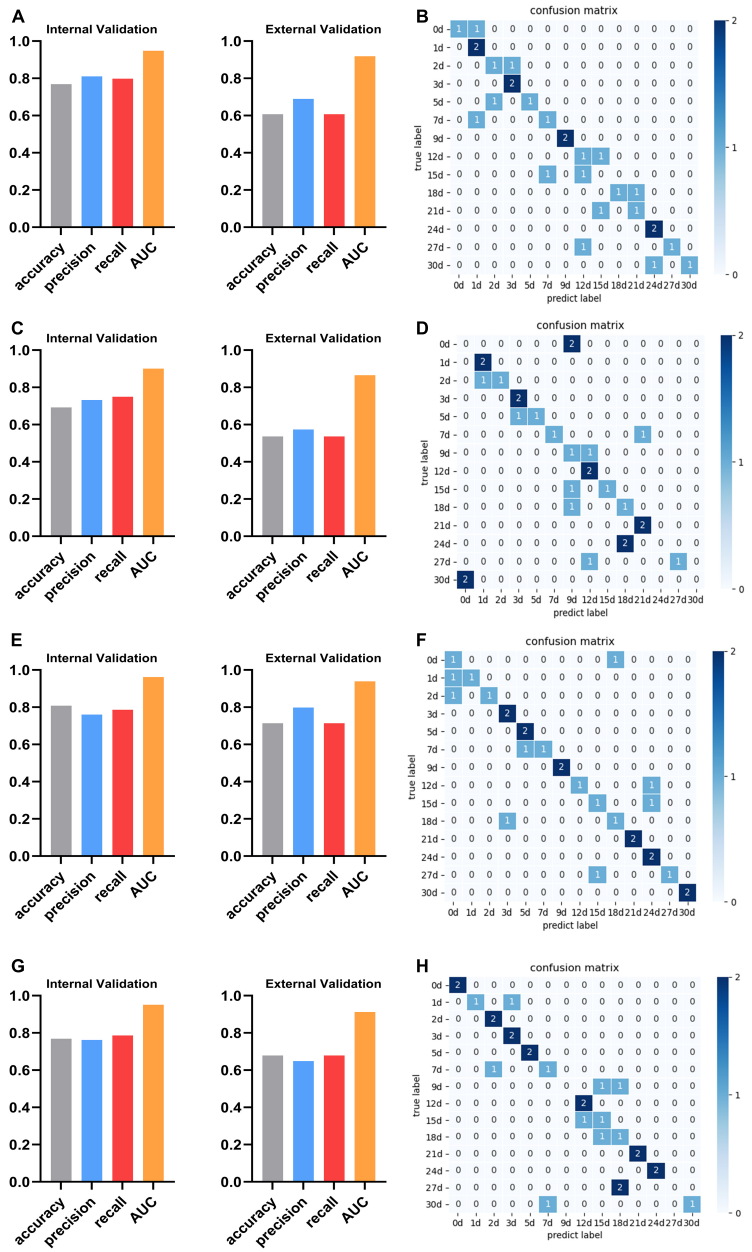

FIGURE 5
The performance and confusion matrix of the optimal sub-models with the lung, liver, kidney, and skeletal muscle. **(A)** The optimal sub-model of the lung is SFS + RF. The accuracy, precision, recall, and AUC of internal validation are 0.769, 0.810, 0.798, and 0.948, respectively. The accuracy, precision, recall, and AUC of external verification of this model are 0.607, 0.690, 0.607, and 0.919, respectively. **(B)** The confusion matrix of SFS + RF for the lung shows that the external validation samples were completely predicted correctly only at 1, 3, 9, and 24 days. The external validation predictions were wrong at 15 days after death. There was a misjudgment in the samples at other PMI. **(C)** The optimal sub-model of the liver is SFS + RF. The accuracy, precision, recall, and AUC are 0.692, 0.732, 0.750, and 0.900, respectively. The accuracy, precision, recall, and AUC of external verification are 0.536, 0.574, 0.536, and 0.865, respectively. **(D)** The confusion matrix of SFS + RF for the liver shows that the external validation samples of Liver were completely predicted correctly at 1, 3, 12, and 21 days. The external validation predictions were wrong at 0, 24, and 30 days after death. There was a misjudgment in the samples at other PMI. **(E)** The optimal sub-model of the kidney is LASSO + RF. The accuracy, precision, recall, and AUC are 0.808, 0.760, 0.786, and 0.962, respectively. The accuracy, precision, recall, and AUC of external verification are 0.714, 0.798, 0.714, and 0.939, respectively. **(F)** The confusion matrix of LASSO + RF for the kidney shows that the external validation samples of Kidney were completely predicted correctly at 3, 5, 9, 21, 24, and 30 days, and there was a misjudgment in the samples at other PMI. **(G)** The optimal sub-model of skeletal muscle is RFE + RF. The accuracy, precision, recall, and AUC are 0.769, 0.762, 0.786, and 0.951, respectively. The accuracy, precision, recall, and AUC of external verification of this model are 0.679, 0.649, 0.679, and 0.912, respectively. **(H)** The confusion matrix of RFE + RF for skeletal muscle shows that the external validation samples of skeletal muscle were completely predicted correctly at 0, 2, 3, 5, 12, 21, and 24 days, The external validation predictions were wrong at 9 and 27 days after death. There was a misjudgment in the samples at other PMI.

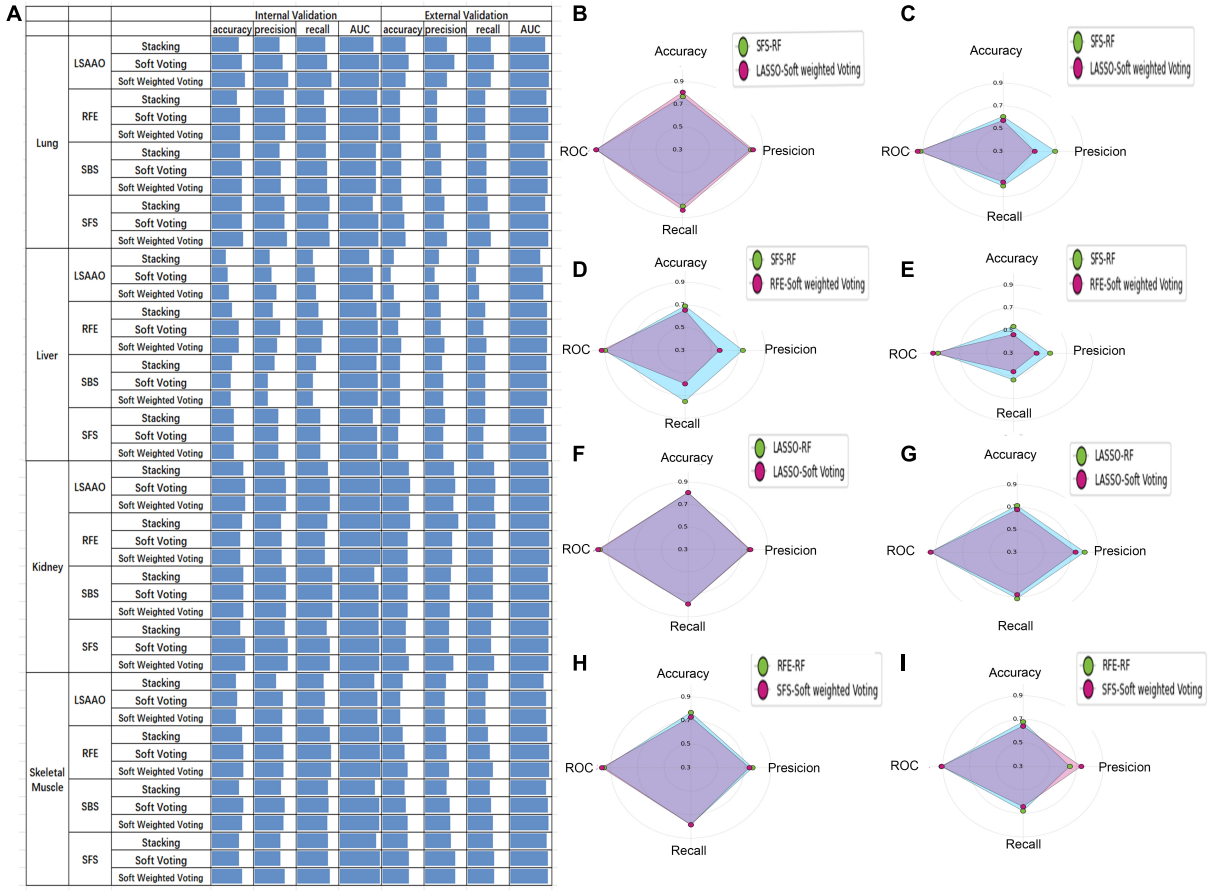

FIGURE 6
The prediction performance of ensemble model and best sub-model based on single organ. **(A)** Histogram performance of internal and external validation of four organs cross-combining of four feature selection methods and three ensemble models. **(B)** Radar map of SFS-RF and LSAAO-soft-weighted voting based on internal validation of lung. **(C)** Radar map of SFS-RF and LSAAO-soft-weighted voting based on external validation of lung. **(D)** Radar map of SFS-RF and RFE-soft-weighted voting based on internal validation of liver. **(E)** Radar map of SFS-RF and RFE-soft-weighted voting based on external validation of liver. **(F)** Radar map of LSAAO-RF and LSAAO-soft voting based on internal validation of kidney. **(G)** Radar map of LSAAO-RF and LSAAO-soft voting based on external validation of kidney. **(H)** Radar map of RFE-RF and SFS-soft-weighted voting based on internal validation of skeletal muscle. **(I)** Radar map of RFE-RF and SFS-soft-weighted voting based on external validation of skeletal muscle.

to the traditional methods. And the terms include whether the required instruments are expensive, whether the detection methods are cumbersome, and the length of analysis time. The results show that the present study's detection method and analysis strategy have good application prospects for estimating PMI (**Table 5**).

4. Discussion

Protein is one of the important components of an organism, so forensic pathologists have always used the analysis of protein degradation after death as an auspicious tool to determine PMI. Previous studies have shown that some specific proteins and their degradation products (e.g., desmin, cTnT, and calpain 1) could be used as markers for specific time intervals of post-mortem decomposition (50). In contrast, many protein detection methods have tested their applicability for predicting PMI. However, these technologies are complex, time-consuming, and expensive, but more importantly, the accuracy is not enough to infer PMI (51).

In the present study, the lab-on-chip combines Agilent 2,100 biological analyzer and the Protein 230 Plus LabChip kits, enabling the separation of polypeptides in the 14–230 kDa range. This technique allows the analysis of 10 samples in 30 min and avoids all the cumbersome post-electrophoresis procedures required for SDS-PAGE analysis, including staining, destaining, and storage, and does not need additional image analysis equipment. It is worth noting that the technology can directly display the results as gel-like images

TABLE 3 The scores of ensemble models based on four organs.

Model	Lung	Liver	Kidney	Skeletal muscle
LASSO + stacking	39.5	15.5	54.5	17
LASSO + soft voting	78.5	17	76	26
LASSO + soft weighted voting	89	30	71.5	21.5
RFE + stacking	24	56.5	60.5	45.5
RFE + soft voting	38.5	70	46	69.5
RFE + soft weighted voting	41	89.5	45	80.5
SBS + stacking	28	52	41.5	36.5
SBS + soft voting	42.5	51	49.5	68.5
SBS + soft weighted voting	45.5	54.5	49.5	67
SFS + stacking	50.5	65	23	46.5
SFS + soft voting	64	61.5	41.5	64.5
SFS + soft weighted voting	83	61.5	65.5	81

and electrophoretograms. It also can output the characteristics of each polypeptide peak as numerical data, such as molecular mass, peak height, and migration time. More importantly, the technology can simultaneously analyze multiple polypeptides or their degradation fragments of a sample. With this high-throughput advantage, this technology will help establish a human sample database and then realize the prediction of human samples with unknown PMI in the future.

The results of this study showed that the prediction accuracy of the kidney was the highest, followed by the lung and skeletal muscle, and that of the liver was the lowest when applying sub-models based on a single organ to predict PMI. The reason may be that the kidney, as a deep organ in the organism, is less affected by the outside world and less protease. The lung and skeletal muscles are greatly affected by the external environment because of gas exchange and relative superficial organs. The result of the liver was the lowest mainly because the detoxification organ of the organism contains many proteolytic enzymes.

According to the results mentioned in the "Performance of the single organ based on ensemble models" section, we found that the performance of ensemble models based on single organ is worse than that of the sub-model. When generating ensemble models, some fundamental principles should be considered. The first is diversity, which means the machine learning algorithm participating in ensemble learning should have enough diversity to obtain ideal prediction performance. The second is prediction performance, which means the individual machine learning algorithm should be as high as possible (52, 53). In the present research, we have used multiple models to ensure the diversity of algorithms. However, the disadvantage is that we have not deleted the worst-performing sub-models, such as GBDT, which may lead to the low accuracy of the integrated model.

In the current study, we designed a multi-organ fusion strategy combining multiple organs to predict PMI. The soft-voting and soft-weighted voting model based on multi-organ fusion strategy improved the predictive performance of internal and external verification. The results show that the soft-voting model drastically improved the accuracy of internal verification from 0.808 to 0.962 and the accuracy of external verification from 0.714 to 0.893. The reason may be that the essence of a multi-organ fusion strategy is to fuse and analyze multiple training datasets to fit different base models. It helps to integrate the characteristics of different organs better and increases the amount of data (53). Another possible reason is that we choose the optimal sub-model of the four organs in the multi-organ fusion strategy to have enough diversity to obtain ideal prediction performance (54).

Through this study, we also found significant differences in the predictive power of different ensemble models, which means it is necessary to compare and screen them. Compared with the Lu et al.'s study, they used the same four organs combined with mass spectrometry and multi-organ fusion to predict PMI, with an accuracy of 0.93 based on a stacking ensemble (55). However, the performance of the stacking ensemble was not satisfactory in our research. On the contrary, the accuracy of soft voting reached 0.96, which may be related to the different analytical techniques.

Ensemble pruning methods, called ensemble selection methods, aim to reduce ensemble models' complexity. These methods search for a subset of ensemble members that performs to some extent as well as the original ensemble (56). This method can reduce the size of the ensemble model, save training time, and improve accuracy and robustness (57). Hence, in our study, we also used the ensemble pruning algorithm to select the optimal subsets of base models for multi-organ fusion, which also means that we can determine the optimal multi-organ combination for the estimation of PMI. Finally, we obtained that the optimal organ combination is the lung, liver, kidney, and skeletal muscle for predicting time since death. This result after pruning also suggests that we should try to use more organs to find the best organ combination to infer future PMI.

In forensic medicine, estimating the PMI is influenced by many internal and external factors such as temperature and humidity, body weight, and disease. The limitations should be avoided in future studies, such as considering more influencing factors and increasing the number of human samples. Although the current experiment involves an idealized condition, we have proven a new analysis method, lab-on-chip combined with a machine learning algorithm, could use to predict the PMI.

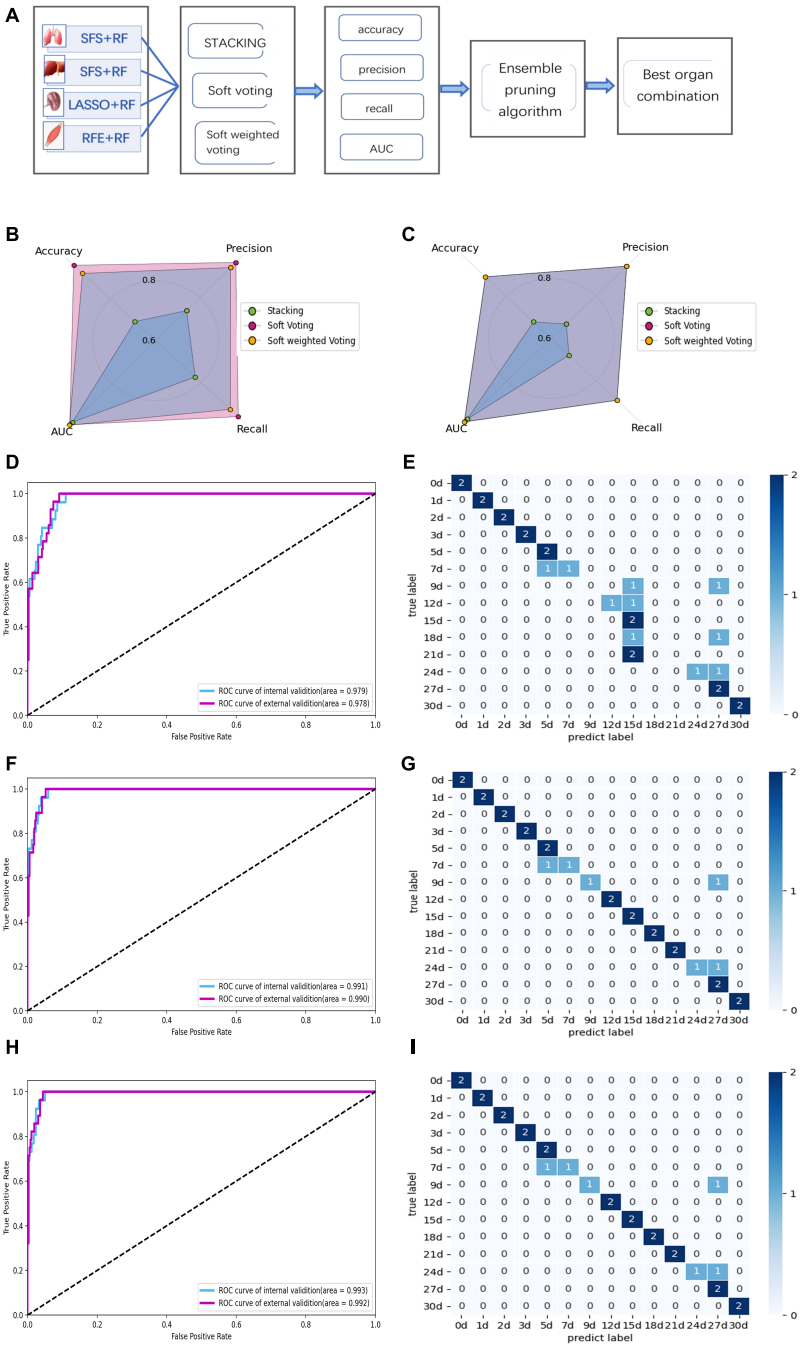

FIGURE 7
Performance of multi-organ fusion strategy to predict PMI. **(A)** Framework of multi-organ fusion strategy to predict PMI. **(B)** Accuracy, precision, recall, and AUC of internal validation for stacking are 0.692, 0.740, 0.774, and 0.979, respectively. Accuracy, precision, recall, and AUC of internal validation for soft voting are 0.962, 0.964, 0.964, and 0.991, respectively. Accuracy, precision, recall, and AUC of internal validation for soft-weighted voting are 0.923, 0.94, 0.929, and 0.993, respectively. **(C)** Accuracy, precision, recall, and AUC of external validation for stacking are 0.679, 0.668, 0.679, and 0.978, respectively. Accuracy, precision, recall and AUC of external validation for soft voting are 0.893, 0.94, 0.893, and 0.99, respectively. Accuracy, precision, recall, and AUC of external validation for soft-weighted voting are 0.893, 0.94, 0.893, and 0.992, respectively. **(D)** The ROC curve of internal and external validation for the stacking model based on multi-organ fusion strategy. **(E)** The confusion matrix of external validation for the stacking model, the mispredictions occurred 7 to 12 days and 18 to 24 days after death. **(F)** The ROC curve of internal and external validation for the soft voting model based on multi-organ fusion strategy. **(G)** The confusion matrix of external validation for the soft voting model. The external validation samples were predicted incorrectly at 7, 9, and 24 days of PMI. **(H)** The ROC curve of internal and external validation for the soft-weighted voting model based on multi-organ fusion strategy. **(I)** The confusion matrix of external validation for the soft-weighted voting model. The samples were mispredicted at 7, 9, and 24 days.

TABLE 4 The summary of all the optimal models is based on single organ sub-models, single organ ensemble models, and multi-organ fusion strategy.

Organ	Best model	Internal validation				External validation			
		ACC[a]	PRE[b]	REC[c]	AUC	ACC	PRE	REC	AUC
Lung	SFS + RF	0.769	0.81	0.798	0.948	0.607	0.690	0.607	0.919
	LASSO + soft weighted voting	0.808	0.827	0.833	0.949	0.571	0.536	0.571	0.941
Liver	SFS + RF	0.692	0.732	0.75	0.9	0.536	0.574	0.536	0.865
	RFE + soft weighted voting	0.654	0.56	0.595	0.924	0.464	0.474	0.464	0.901
Kidney	LASSO + RF	0.808	0.76	0.786	0.962	0.714	0.798	0.714	0.939
	LASSO + soft weighted voting	0.808	0.767	0.786	0.974	0.643	0.683	0.643	0.94
Skeletal muscle	RFE + RF	0.769	0.762	0.786	0.951	0.679	0.649	0.679	0.912
	SFS + soft weighted voting	0.731	0.738	0.786	0.966	0.643	0.735	0.643	0.91
Multi-organ fusion	Stacking	0.692	0.74	0.774	0.979	0.679	0.668	0.679	0.978
	Soft voting	0.962	0.964	0.964	0.991	0.893	0.94	0.893	0.99
	Soft weighted voting	0.923	0.94	0.929	0.993	0.893	0.94	0.893	0.992

[a]ACC represents accuracy.
[b]PRE represents precision.
[c]REC represents recall.

TABLE 5 Comparison of lab-on-chip and traditional protein detection methods.

	Lab-on-chip	Traditional methods		
		Western-blotting	ELISA	Protein mass spectrometry
Operations	Simplify	Complex	Complex	Complex
Sample consumption	Minimal	Major	Major	Minimal
Expenditure	Cheap	Cheap	Cheap	Expensive
Speed	Less than 30 min	Slow	Fast	Slow
Equipment	Only 2,100 Bioanalyzer	Variety	Few	Variety
Identify particular protein	No	Yes	Yes	Yes
Quantitation	Automatic	Semiquantitative	Semiquantitative	Automatic
High throughput	Yes	No	No	Yes
Data processing	Use machine learning	Manual analysis	Manual analysis	Use machine learning
Predict performance	Excellent	Poor	Poor	Good
Witnessed inspections	Yes	No	Yes	No

Furthermore, the multi-organ fusion strategy can significantly improve the performance of PMI prediction.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

This animal study was reviewed and approved by the Institutional Animal Care and Use Committee of Shanxi Medical University.

Author contributions

Q-XD, J-HS, and SZ conceived the idea and drafted the manuscript. F-HL, X-JL, and LW established animal model and lab-on-chip experiment. JC and J-HS conceived the study and performed some interpretation of the results. Q-QJ and KR performed data collection and mathematical model construction. JZ and PH interpreted the results and modify the manuscript. All authors read and approved the final version of the manuscript.

Funding

This study was supported by the Shanghai Key Laboratory of Forensic Medicine (Academy of Forensic Science) and the

special fund for the Science and Technology Innovation Teams of Shanxi Province.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Hunter M, Pozhitkov A, Noble P. Accurate predictions of postmortem interval using linear regression analyses of gene meter expression data. *Forensic Sci Int.* (2017) 275:90–101. doi: 10.1016/j.forsciint.2017.02.027

2. Metcalf J, Wegener Parfrey L, Gonzalez A, Lauber C, Knights D, Ackermann G, et al. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *Elife.* (2013) 2:e01104. doi: 10.7554/eLife.01104

3. Henssge C, Madea B, Gallenkemper E. Death time estimation in case work. II. Integration of different methods. *Forensic Sci Int.* (1988) 39:77–87. doi: 10.1016/0379-0738(88)90120-x

4. Kaatsch H, Stadler M, Nietert M. Photometric measurement of color changes in livor mortis as a function of pressure and time. Development of a computer-aided system for measuring pressure-induced blanching of livor mortis to estimate time of death. *Int J Legal Med.* (1993) 106:91–7. doi: 10.1007/bf01225047

5. Krompecher T. Experimental evaluation of rigor mortis. VIII. Estimation of time since death by repeated measurements of the intensity of rigor mortis on rats. *Forensic Sci Int.* (1994) 68:149–59. doi: 10.1016/0379-0738(94)90354-9

6. Canturk I, Celik S, Sahin M, Yagmur F, Kara S, Karabiber F. Investigation of opacity development in the human eye for estimation of the postmortem interval. *Biocybern Biomed Eng.* (2017) 37:559–65. doi: 10.1016/j.bbe.2017.02.001

7. Mathur A, Agrawal Y. An overview of methods used for estimation of time since death. *Aust J Forensic Sci.* (2011) 43:275–85. doi: 10.1080/00450618.2011.568970

8. Hansen J, Lesnikova I, Funder A, Banner J. DNA and RNA analysis of blood and muscle from bodies with variable postmortem intervals. *Forensic Sci Med Pathol.* (2014) 10:322–8. doi: 10.1007/s12024-014-9567-2

9. De Simone S, Giacani E, Bosco M, Vittorio S, Ferrara M, Bertozzi G, et al. The role of mirnas as new molecular biomarkers for dating the age of wound production: a systematic review. *Front Med.* (2021) 8:803067. doi: 10.3389/fmed.2021.803067

10. Jawor P, Ząbek A, Wojtowicz W, Król D, Stefaniak T, Młynarz P. Metabolomic studies as a tool for determining the post-mortem interval (PMI) in stillborn calves. *BMC Vet Res.* (2019) 15:189. doi: 10.1186/s12917-019-1935-4

11. Donaldson A, Lamont I. Estimation of post-mortem interval using biochemical markers. *Aust J Forensic Sci.* (2014) 46:8–26. doi: 10.1080/00450618.2013.784356

12. Hyde E, Haarmann D, Lynne A, Bucheli S, Petrosino J. The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. *PLoS One.* (2013) 8:e77733. doi: 10.1371/journal.pone.0077733

13. Wang, S, Chen W, Shang Y, Ren L, Zhang X, Guo Y, et al. High-throughput sequencing to evaluate the effects of methamphetamine on the succession of the bacterial community to estimate the postmortem interval. *Forensic Sci Res.* (2022) 1–12. doi: 10.1080/20961790.2022.2046368

14. Peng D, Lv M, Li Z, Tian H, Qu S, Jin B, et al. Postmortem interval determination using mrna markers and DNA normalization. *Int J Legal Med.* (2020) 134:149–57. doi: 10.1007/s00414-019-02199-7

15. Lv Y, Ma J, Pan H, Zeng Y, Tao L, Zhang H, et al. Estimation of the human postmortem interval using an established rat mathematical model and multi-RNA markers. *Forensic Sci Med Pathol.* (2017) 13:20–7. doi: 10.1007/s12024-016-9827-4

16. Lv Y, Ma K, Zhang H, He M, Zhang P, Shen Y, et al. A time course study demonstrating mRNA, microRNA, 18s rRNA, and U6 snRNA changes to estimate PMI in deceased rat's spleen. *J Forensic Sci.* (2014) 59:1286–94. doi: 10.1111/1556-4029.12447

17. Sato T, Zaitsu K, Tsuboi K, Nomura M, Kusano M, Shima N, et al. A preliminary study on postmortem interval estimation of suffocated rats by Gc-Ms/Ms-based plasma metabolic profiling. *Anal Bioanal Chem.* (2015) 407:3659–65. doi: 10.1007/s00216-015-8584-7

18. Dias A, Castro A, Melo P, Tarelho S, Domingues P, Franco JM. A fast method for Ghb-gluc quantitation in whole blood by Gc-Ms/Ms (Tqd) for forensic purposes. *J Pharm Biomed Anal.* (2018) 150:107–11. doi: 10.1016/j.jpba.2017.11.072

19. Wang Q, He H, Li B, Lin H, Zhang Y, Zhang J, et al. Uv-Vis and Atr-Ftir spectroscopic investigations of postmortem interval based on the changes in rabbit plasma. *PLoS One.* (2017) 12:e0182161. doi: 10.1371/journal.pone.0182161

20. Zhang J, Li B, Wang Q, Li C, Zhang Y, Lin H, et al. Characterization of postmortem biochemical changes in rabbit plasma using Atr-Ftir combined with chemometrics: a preliminary study. *Spectrochim Acta A Mol Biomol Spectrosc.* (2017) 173:733–9. doi: 10.1016/j.saa.2016.10.041

21. Javan G, Finley S, Can I, Wilkinson J, Hanson J, Tarone A. Human thanatomicrobiome succession and time since death. *Sci Rep.* (2016) 6:29598. doi: 10.1038/srep29598

22. Metcalf J, Xu Z, Weiss S, Lax S, Van Treuren W, Hyde E, et al. Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science.* (2016) 351:158–62. doi: 10.1126/science.aad2646

23. Finley S, Benbow M, Javan G. Microbial communities associated with human decomposition and their potential use as postmortem clocks. *Int J Legal Med.* (2015) 129:623–32. doi: 10.1007/s00414-014-1059-0

24. Pechal J, Schmidt C, Jordan H, Benbow ME. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. *Sci Rep.* (2018) 8:5724. doi: 10.1038/s41598-018-23989-w

25. Wells J, Lecheta M, Moura M, LaMotte L. An evaluation of sampling methods used to produce insect growth models for postmortem interval estimation. *Int J Legal Med.* (2015) 129:405–10. doi: 10.1007/s00414-014-1029-6

26. Schmidt V, Zelger P, Woess C, Pallua A, Arora R, Degenhart G, et al. Application of micro-computed tomography for the estimation of the post-mortem interval of human skeletal remains. *Biology.* (2022) 11:1105. doi: 10.3390/biology11081105

27. Wilk L, Edelman G, Roos M, Clerkx M, Dijkman I, Melgar J, et al. Individualised and non-contact post-mortem interval determination of human bodies using visible and thermal 3d imaging. *Nat Commun.* (2021) 12:5997. doi: 10.1038/s41467-021-26318-4

28. Pittner S, Ehrenfellner B, Zissler A, Racher V, Trutschnig W, Bathke A, et al. First application of a protein-based approach for time since death estimation. *Int J Legal Med.* (2017) 131:479–83. doi: 10.1007/s00414-016-1459-4

29. He J, Huang H, Qu D, Xue Y, Zhang K, Xie X, et al. Cxcl1 and Cxcr2 as potential markers for vital reactions in skin contusions. *Forensic Sci Med Pathol.* (2018) 14:174–9. doi: 10.1007/s12024-018-9969-7

30. Gauchotte G, Bochnakian A, Campoli P, Lardenois E, Brix M, Simon E, et al. Myeloperoxydase and Cd15 with glycophorin C double staining in the evaluation of skin wound vitality in forensic practice. *Front Med.* (2022) 9:910093. doi: 10.3389/fmed.2022.910093

31. Bertozzi G, Ferrara M, La Russa R, Pollice G, Gurgoglione G, Frisoni P, et al. Wound vitality in decomposed bodies: new frontiers through immunohistochemistry. *Front Med.* (2021) 8:802841. doi: 10.3389/fmed.2021.802841

32. Li C, Li Z, Tuo Y, Ma D, Shi Y, Zhang Q, et al. Maldi-Tof Ms as a novel tool for the estimation of postmortem interval in liver tissue samples. *Sci Rep.* (2017) 7:4887. doi: 10.1038/s41598-017-05216-0

33. Prieto-Bonete G, Pérez-Cárceles M, Maurandi-López A, Pérez-Martínez C, Luna A. Association between protein profile and postmortem interval in human bone remains. *J Proteomics.* (2019) 192:54–63. doi: 10.1016/j.jprot.2018.08.008

34. Mann A, Tighe B. Tear analysis and lens-tear interactions. Part I. Protein fingerprinting with microfluidic technology. *Contact Lens Anterior Eye.* (2007) 30:163–73. doi: 10.1016/j.clae.2007.03.006

35. Cristina R, Kocsis R, Tulcan C, Alexa E, Boldura O, Hulea C, et al. Protein structure of the venom in nine species of snake: from bio-compounds to possible healing agents. *Brazil J Med Biol Res.* (2020) 53:e9001. doi: 10.1590/1414-431x20199001

36. Blazek V, Caldwell R. Comparison of Sds gel capillary electrophoresis with microfluidic lab-on-a-chip technology to quantify relative amounts of 7s and 11s proteins from 20 soybean cultivars. *Int J Food Sci Technol.* (2009) 44:2127–34. doi: 10.1111/j.1365-2621.2009.02049.x

37. Schmut O, Horwath-Winter J, Zenker A, Trummer G. The effect of sample treatment on separation profiles of tear fluid proteins: qualitative and semi-quantitative protein determination by an automated analysis system. *Graefes Arch Clin Exp ophthalmol.* (2002) 240:900–5. doi: 10.1007/s00417-002-0537-0

38. Zissler A, Ehrenfellner B, Foditsch E, Monticelli F, Pittner S. Does altered protein metabolism interfere with postmortem degradation analysis for PMI estimation? *Int J Legal Med.* (2018) 132:1349–56. doi: 10.1007/s00414-018-1814-8

39. Abo El-Noor M, Elhosary N, Khedr N, El-Desouky K. Estimation of Early Postmortem Interval through biochemical and pathological changes in rat heart and kidney. *Am J Forensic Med Pathol.* (2016) 37:40–6. doi: 10.1097/paf.0000000000000214

40. Li L, Lin Y, Yu D, Liu Z, Gao Y, Qiao JPA. Multi-organ fusion and lightgbm based radiomics algorithm for high-risk esophageal varices prediction in cirrhotic patients. *IEEE Access.* (2021) 9:15041–52. doi: 10.1109/access.2021.3052776

41. Kozawa S, Sagawa F, Endo S, De Almeida G, Mitsuishi Y, Sato T. Predicting human clinical outcomes using mouse multi-organ transcriptome. *iScience* (2020) 23:100791. doi: 10.1016/j.isci.2019.100791

42. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Ser B Stat Methodol.* (2011) 73:273–82. doi: 10.1111/j.1467-9868.2011.00771.x

43. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* (2002) 46:389–422.

44. Aha D, Bankert R. A comparative evaluation of sequential feature selection algorithms: learning from data. In: Fisher D, Lenz H editors. *Learning from Data. Lecture Notes in Statistics*. New York, NY: Springer (1996).

45. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Res.* (2001) 11:1878–87. doi: 10.1101/gr.190001

46. Cai Z, Xu D, Zhang Q, Zhang J, Ngai S, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol Biosyst.* (2015) 11:791–800. doi: 10.1039/c4mb00659c

47. Wolpert D. Stacked Generalization. *Neural Netw.* (1992) 5:241–59.

48. Lin X, Yacoub S, Burns J, Simske S. Performance analysis of pattern classifier combination by plurality voting. *Pattern Recogn Lett.* (2003) 24:1959–69.

49. Zhang Y, Zhang H, Cai J, Yang BB. A weighted voting classifier based on differential evolution. *Abstr Appl Anal.* (2014) 2014:6. doi: 10.1155/2014/376950

50. Pittner S, Ehrenfellner B, Monticelli F, Zissler A, Sänger A, Stoiber W, et al. Postmortem muscle protein degradation in humans as a tool for PMI delimitation. *Int J Legal Med.* (2016) 130:1547–55. doi: 10.1007/s00414-016-1349-9

51. Zissler A, Stoiber W, Steinbacher P, Geissenberger J, Monticelli F, Pittner S. Postmortem protein degradation as a tool to estimate the PMI: a systematic review. *Diagnostics.* (2020) 10:1014. doi: 10.3390/diagnostics10121014

52. Deng H, Runger G, Tuv E, Vladimir M. A time series forest for classification and feature extraction. *Inform Sci.* (2013) 239:142–53. doi: 10.1016/j.ins.2013.02.030

53. Sagi O, Rokach L. Ensemble learning: a survey. *Wiley Interdiscip Rev Data Min Knowl Discov.* (2018) 8:e1249. doi: 10.1002/widm.1249

54. Lin S, Chen S. Parameter determination and feature selection for C4.5 algorithm using scatter search approach. *Soft Comput.* (2012) 16:63–75. doi: 10.1007/s00500-011-0734-z

55. Lu, X, Li J, Wei X, Li N, Dang L, An G, et al. A novel method for determining postmortem interval based on the metabolomics of multiple organs combined with ensemble learning techniques. *Int J Legal Med.* (2022). doi: 10.1007/s00414-022-02844-8

56. Bian Y, Wang Y, Yao Y, Chen H. Ensemble pruning based on objection maximization with a general distributed framework. *IEEE Trans Neural Netw Learn Syst.* (2020) 31:3766–74. doi: 10.1109/tnnls.2019.2945116

57. Martínez-Muñoz G, Hernández-Lobato D, Suárez A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Trans Pattern Anal Mach Intell.* (2009) 31:245–59. doi: 10.1109/tpami.2008.78

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 08 February 2023
DOI 10.3389/fmed.2023.999312

OPEN ACCESS

EDITED BY
Jingjing You,
The University of Sydney, Australia

REVIEWED BY
Taobo Hu,
Peking University People's Hospital, China
Enrico Capobianco,
Jackson Laboratory, United States

*CORRESPONDENCE
Liang Zeng
✉ zlxx03@126.com
Ming Li
✉ liming@csu.edu.cn
Xiyun Deng
✉ dengxiyunmed@hunnu.edu.cn

SPECIALTY SECTION
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

RECEIVED 20 July 2022
ACCEPTED 16 January 2023
PUBLISHED 08 February 2023

CITATION
Wang K, Zheng C, Xue L, Deng D, Zeng L, Li M and Deng X (2023) A bibliometric analysis of 16,826 triple-negative breast cancer publications using multiple machine learning algorithms: Progress in the past 17 years. *Front. Med.* 10:999312. doi: 10.3389/fmed.2023.999312

A bibliometric analysis of 16,826 triple-negative breast cancer publications using multiple machine learning algorithms: Progress in the past 17 years

Kangtao Wang[1], Chanjuan Zheng[2], Lian Xue[2], Dexin Deng[3], Liang Zeng[4]*, Ming Li[5]* and Xiyun Deng[2]*

[1]Department of General Surgery, The Xiangya Hospital, Central South University, Changsha, Hunan, China, [2]Key Laboratory of Model Animals and Stem Cell Biology in Hunan, Department of Pathophysiology, School of Medicine, Hunan Normal University, Changsha, Hunan, China, [3]Xiangya School of Medicine, Central South University, Changsha, Hunan, China, [4]Department of Pathology, Guangzhou Women and Children's Medical Center, Guangdong Provincial Clinical Research Center for Child Health, Guangzhou, China, [5]Department of Immunology, College of Basic Medical Sciences, Central South University, Changsha, Hunan, China

Background: Triple-negative breast cancer (TNBC) is proposed at the beginning of this century, which is still the most challenging breast cancer subtype due to its aggressive behavior, including early relapse, metastatic spread, and poor survival. This study uses machine learning methods to explore the current research status and deficiencies from a macro perspective on TNBC publications.

Methods: PubMed publications under "triple-negative breast cancer" were searched and downloaded between January 2005 and 2022. R and Python extracted MeSH terms, geographic information, and other abstracts from metadata. The Latent Dirichlet Allocation (LDA) algorithm was applied to identify specific research topics. The Louvain algorithm established a topic network, identifying the topic's relationship.

Results: A total of 16,826 publications were identified, with an average annual growth rate of 74.7%. Ninety-eight countries and regions in the world participated in TNBC research. Molecular pathogenesis and medication are most studied in TNBC research. The publications mainly focused on three aspects: Therapeutic target research, Prognostic research, and Mechanism research. The algorithm and citation suggested that TNBC research is based on technology that advances TNBC subtyping, new drug development, and clinical trials.

Conclusion: This study quantitatively analyzes the current status of TNBC research from a macro perspective and will aid in redirecting basic and clinical research toward a better outcome for TNBC. Therapeutic target research and Nanoparticle research are the present research focus. There may be a lack of research on TNBC from a patient perspective, health economics, and end-of-life care perspectives. The research direction of TNBC may require the intervention of new technologies.

KEYWORDS

machine learning, bibliometric analysis, Latent Dirichlet Allocation, triple-negative breast cancer, Nanoparticle research

Highlights

- All Triple-negative breast cancer (TNBC) publications in the PubMed database from 2005 to 2021 were included in the analysis.
- Triple-negative breast cancer research mainly focused on three aspects: Therapeutic target research, Prognostic research, and Mechanism research.
- Therapeutic target research and Nanoparticle research are the present research focus.
- The Latent Dirichlet Allocation (LDA) algorithm we built is a convenient tool that can help researchers discover changes in research focus from medical text big data.

1. Background

Breast cancer currently accounts for 30% of newly diagnosed malignant tumors in women and causes 15% of women to die from cancer (1). For the first time, Perou described the intrinsic molecular subtypes of breast cancer and described Triple-negative breast cancer (TNBC) in 2000 using complementary DNA microarray technology (2). Furthermore, TNBC is the most aggressive subtype of breast cancer, accounting for about 10–20% of breast cancer cases (3, 4). TNBC is still unsatisfactory in diagnosis and treatment.

Bibliometrics is a quantitative analysis method of academic publications, which can discover the progress of discipline research from a macro perspective and provide support for future research directions (5). TNBC-related literature information analysis is scarce. Teles et al. (6) conducted a bibliometric study of 1,932 publications in 2018 to study nanomedicine research's global trend on TNBC. However, the inclusion criteria of this study are too broad, and the analysis methods are insufficient to analyze the *status quo* of the TNBC study. Unfortunately, bibliometric studies on TNBC remain insufficient due to the lack of practical language analysis tools to integrate metatext data.

Natural Language Processing (NLP) is a computing technology used to analyze human language, a part of machine learning (7). Various algorithms have been successfully applied to deal with medical information (8). Latent Dirichlet Allocation (LDA) is bibliometrics's most classical topic modeling method to present many unstructured texts and information (9, 10). LDA can perform topic analysis on texts (5). We recently constructed LDA and NLP methods to analyze more than 23,000 rectal cancer-related publications between 1994 and 2018. We have found the research deficiencies in the last 25 years and predicted the future research focus (11). Therefore, through the use of mature LDA methods and machine learning techniques to discover the current research from a macro perspective, at the same time discover the missing research topics in the past, and predict potential research breakthroughs in the future.

We analyzed all past TNBC publications indexed by PubMed under Triple-negative breast cancer in the present study. We improved our algorithm based on previous research and conducted a more detailed analysis of all TNBC publications with more visual expression to highlight current research focus in TNBC, research deficiencies, and specific areas with future opportunities.

Abbreviations: TNBC, triple-negative breast cancer; NLP, natural language processing; LDA, Latent Dirichlet Allocation.

2. Materials and methods

2.1. Research design

The study design was based on the basic rules of bibliometrics, as shown in **Figure 1** for a flowchart (12, 13). The study used a two-stage structured approach to bibliometric analysis and visual assessment of published scientific literature. Provide an understanding based on the data and the researcher's professional background. The PubMed database[1] is a biomedical specialty database that provides multiple search strategies and is a free, publicly available database. For this research, the PubMed database, which contains an application programming interface (API) that can export abstracts, was used, and publications containing abstracts were downloaded for analysis.

2.2. Inclusive and exclusive criteria

Table 1 shows the steps to obtain full TNBC-related publications in the PubMed database. All publications under Triple Negative Breast Cancer were downloaded between January 1, 2005, and January 1, 2022. There are 17,562 publications. Missing data, conference abstracts, conference proceedings, book reviews, and news items were excluded, and 17,338 publications were ultimately included in the bibliometric analysis (**Figure 1A**). Details of inclusion and exclusion are shown in **Table 2**. After excluding non-English publications and incomplete abstracts, the final 16,826 publications were analyzed by the LDA algorithm to obtain the focus changes and their relevance of research topics in publications in this field. The whole record of search results is downloaded in XML format *via* R's easyPubMed package. Data extracted from R[2] and Python[3], including publication year, abstract, study types, geographic information, and Medical Subject Headings (MeSH) terms, were obtained.

2.3. LDA and algorithms and analytical methods

Latent Dirichlet Allocation was used to identify more specific research topics in each article. Python was used to model the topics by analyzing the abstracts of all indexed articles in the record. Topics were set at 50. The criteria for selecting the number of topics were perplexity, redundancy, and legibility. Based on the algorithmic calculation of topic probability, we finally determined the topic to which each article belongs. Next, we manually checked the names of each glossary based on the abstract. Finally, we used the Louvain algorithm and Gephi to perform cluster analysis to establish a topic network to determine the relationship between topics (14). We identified the two topics with the highest attribution probability in each publication, counted the number of simultaneous occurrences in each document, and established links between topics.

All the original data were uploaded and publicly available, including all retrieval methods, algorithm codes, and raw literature data in this article (**Figure 1A**). The literature search and download

1 https://pubmed.ncbi.nlm.nih.gov/

2 https://www.r-project.org/, version:4.1.1

3 https://www.python.org/, version 3.7.1

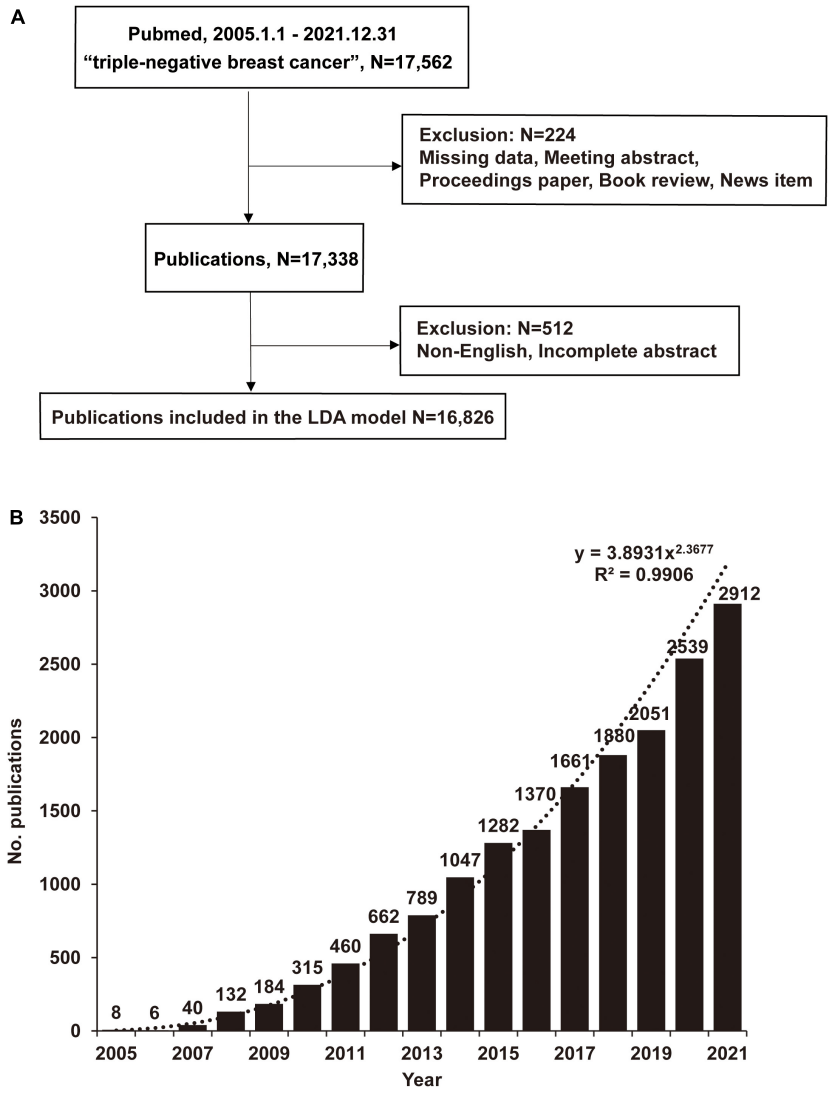

FIGURE 1
The number of publications on triple-negative breast cancer (TNBC) has increased rapidly in recent 17 years. **(A)** Using the search terms "triple-negative breast cancer" in the PubMed database, download publications through the R pubquery package. Missing data or when the publication was a meeting abstract, proceedings paper, a correction, a book review, or a news item were manually excluded, and finally, 17,338 publications were included in the general analysis. Latent Dirichlet Allocation (LDA) analyzed 16,826 publications. **(B)** Publications analyzed by LDA, Python. Data were visualized using Excel. The number of publications is shown yearly, and $y = 3.8931x^{2.3677}$ ($R^2 = 0.9906$) is the fitted function.

code can be obtained on R by easyPubMed package[4]. The R code is publicly available on GitHub[5]. We have uploaded relevant Python code on GitHub[6], Zenodo[7] and LDA code (Supplementary LDA coding-updated). The network visualization in this article is carried out using the software package Gephi[8]. This study used publicly published data and did not need approval by the relevant institutional review board or ethics committee. A step-by-step instruction is provided in the **Supplementary material** to facilitate the reader to understand further the research details (**Supplementary information 1**).

4 https://cran.r-project.org/web/packages/easyPubMed/index.html
5 https://github.com/christopherBelter/pubmedXML
6 https://github.com/mxdwangdali11/guid-to-Bibliometric-LDA-Analysis
7 https://doi.org/10.5281/zenodo.7461925
8 https://gephi.org/, version 0.9.2

3. Results

3.1. The number of publications in TNBC research increases every year

We identified and analyzed 16,826 publications from January 2005 to 2022 (**Figure 1B**). The annual growth trend aligns with the fitting curve $y = 3.8931x^{2.3677}$ ($R^2 = 0.9906$). An average of 1,019 publications are published each year, with an average annual growth rate of 74.7%. It is expected that 3,650 publications will be published in 2022. Among all publications, 1,646 journals have publications on TNBC. We identified the ten most popular journals that published 3,118 publications, accounting for 18.0% of all publications (**Supplementary Table 1**). Therefore, emphasizing posts from these key journals helps us keep up with the latest trends. *Breast*

Cancer Research and Treatment, *PLoS One*, and *Scientific Reports* are the top three journals with 690, 427, and 331 publications.

3.2. The proportion of clinical trials in TNBC publications has increased every year

To explore the research fields of TNBC, we first divided the publications into nine categories according to the fields provided by the database from 2010 in cancer research and set them as 100 per cent (Figure 2). We found that clinical trials and multicenter studies accounted for 25% of publications. The proportions of reviews and meta-analyses increased from 35% in 2011 to 50% in 2021. Since high-quality meta-analysis is generally considered a clinically guiding study, it is reasonable to expect that the publication of TNBC meta-analysis will increase. Many clinical trials of TNBC have been improved and will continue to improve its clinical practice.

3.3. The United States and China have the highest number of publications in the field of TNBC

To further understand the global TNBC research situation, we analyzed the geographic information by research institutions. We found that 98 countries or regions worldwide have publications on TNBC (Figure 3A). The top 10 countries' publications accounted

TABLE 1 Triple-negative breast cancer (TNBC) publications assortment steps.

Exploration steps	Query on PubMed	Description
1	Triple negative breast cancer	("triple negative breast neoplasms"[MeSH Terms] OR ("triple"[All Fields] AND "negative"[All Fields] AND "breast"[All Fields] AND "neoplasms"[All Fields]) OR "triple negative breast neoplasms"[All Fields])
2	Data duration	(2005:2021[pdat])

TABLE 2 Inclusive and exclusive criteria.

Parameter of selection of a publication	Inclusion criterion	Exclusion criterion	Rationale for inclusion–exclusion
Language	English	Other languages	The working language of the LDA algorithm is English. Other languages are not recognized
Publication date	2005–2021	Publications before 2005 and after 2021	Not included in the 2022 publication as it has not been fully published
Publication type	All	Missing data, meeting abstract, proceeding paper, book review, news item	As the LDA algorithm is unsupervised machine learning, the analysis must include abstract as the text editor. In addition to incomplete content, try to include research articles and reviews.
Funding sponsor	All	No exclusion	This parameter does not affect the selection criterion
Affiliation/organization	All	No exclusion	This parameter does not affect the selection criterion
Funding	All	No exclusion	This parameter does not affect the selection criterion
Country	All	No exclusion	Publication from each country has its significance

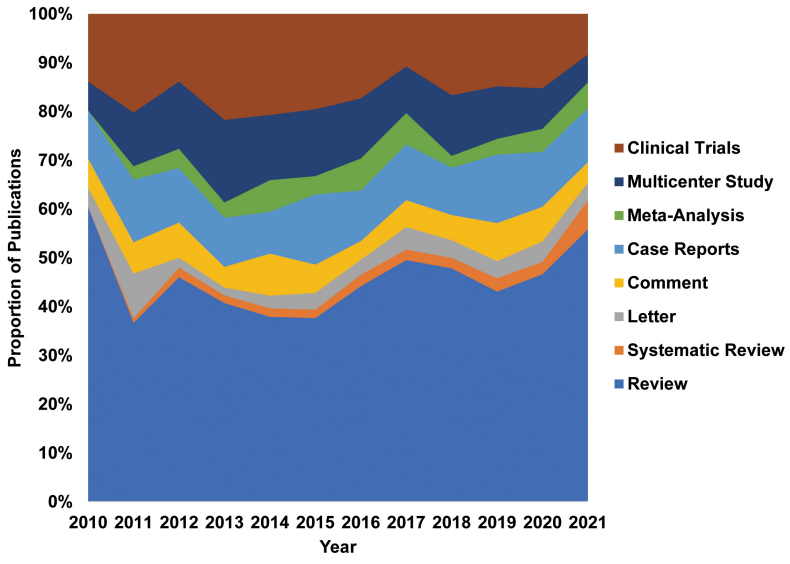

FIGURE 2
Clinical trials and multicenter studies have a large proportion of research. We divide publications into eight categories according to the types provided in the database. Data were shown by percentage.

for 78.2%, indicating a pronounced head effect. Moreover, more than half of the publications were derived from the United States, China, Korea, and Italy, accounting for 25.0%, 21.8%, 5.4%, and 4.9% of all publications, respectively (Figure 3B). This phenomenon reminds us that the vast majority of the global population has participated in TNBC research, especially in the northern hemisphere.

3.4. Molecular pathogenesis and medication are most studied in TNBC research

MeSH terms can represent the research content of the publications. A total of 6,288 MeSH terms appeared 248,250 times in all 16,826 publications, indicating that the studies covered multiple aspects (Supplementary Table 2). The top 10 cited MeSH terms are listed in Figure 4. Both pathology and metabolism have appeared more than 7,000 times, suggesting that the research on TNBC focused on exploring its molecular pathogenesis. In addition, 5 of the top 10 cited MeSH terms are directly related to medication research. Therefore, we infer that pathogenic mechanism and medication research will continue to focus on TNBC research in the foreseeable future.

3.5. LDA results: TNBC research focus on therapeutic target research, prognostic research, and mechanism research

The topic network analyzed by LDA and Louvain algorithm highlights the areas where interrelated topic clusters appear

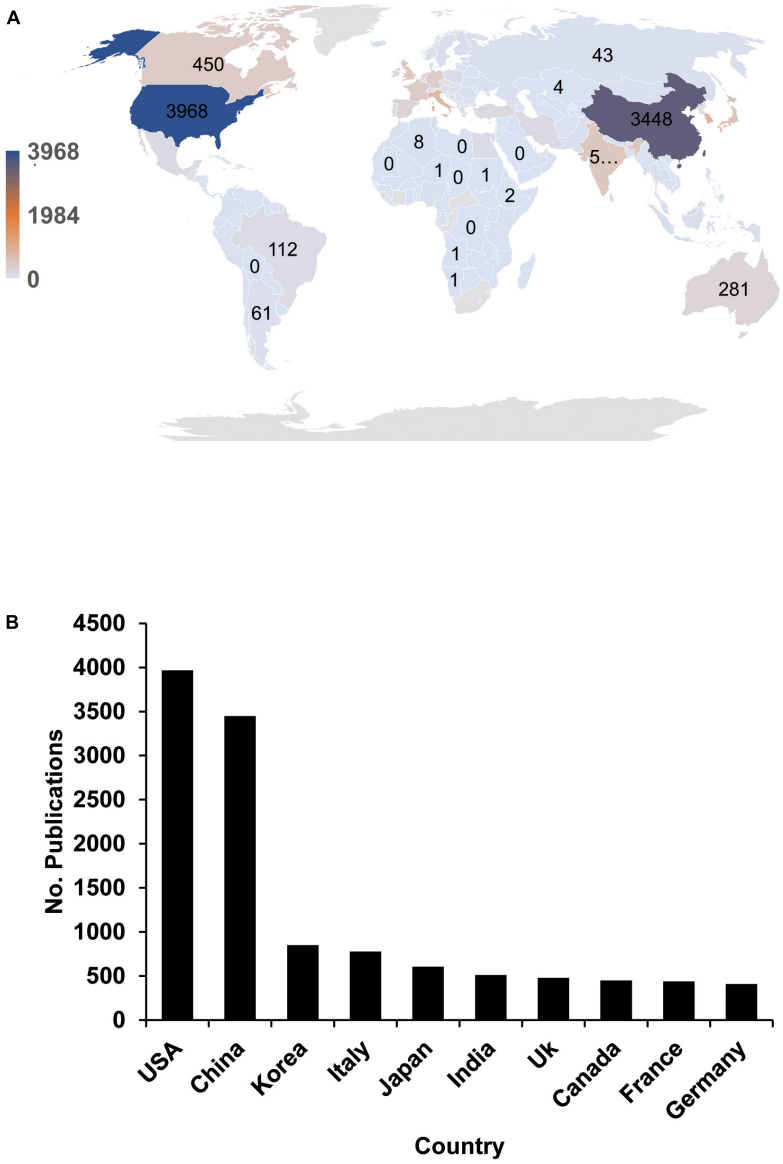

FIGURE 3
Global triple-negative breast cancer (TNBC) research differs significantly between regions. **(A)** The global distribution of TNBC publications in the recent 17 years is shown. We extracted the country information based on the first publication's affiliation. **(B)** Top 10 countries with the highest publication numbers in TNBC research.

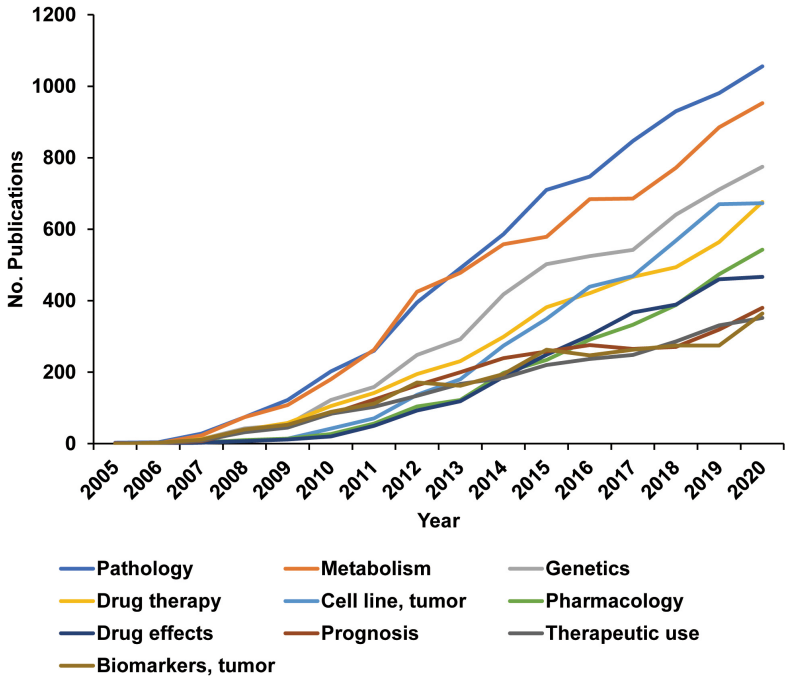

FIGURE 4
Molecular pathogenesis and medication are most studied in triple-negative breast cancer (TNBC) research. Each publication contains several Medical Subject Headings terms to describe the research content roughly. R was used to analyze the themes of the publications through Medical Subject Headings terms. The figure shows the most researched topics in the last 16 years.

simultaneously and provides remarkable insights into the relationships between the essential topics of interest. We divided publications into 50 topics. The results of the LDA analysis suggest that all TNBC-related studies are mainly focused on three clusters, i.e., Therapeutic target research, Prognostic research, and Mechanism research (Figure 5). However, few studies on hospice care, patient perspective, surgical treatment of metastasis, and economics are available.

The Therapeutic target research cluster contains 3,465 publications. The research focuses on Therapeutic target research, Protein expression, and Chemotherapy research. This cluster is particularly close to the other two clusters, indicating that the relationship between essential clinical integration and TNBC basic research is very close. We also found that clinical trials can quickly transform basic research into clinical practice to improve patient prognosis.

In the Prognostic research cluster, Survival related research and Demography research are the most studied topics. There are 1,275 publications on Prognostic research, which account for the most significant proportion and are closely related to the other two topics, indicating that prognostic research is the research focus. Interestingly, we found that Demography research and Methylation research are highly connected, weighing 359. We further analyzed and found that TNBC methylation differs significantly among races with different genetic backgrounds, and long-term survival studies are lacking.

In the Mechanism research cluster, we found that Apoptosis research, Growth factors study, and Nanoparticle research are the three most researched topics. In addition, The research cluster contains 21 topics, accounting for up to 42%, covering everything from basic medical research to clinical research.

3.6. LDA results: Therapeutic target research and Nanoparticle research are the research focus

To understand the changes in research focus, we visualized the LDA results and generated a heat map showing the changes in all 50 research topics of TNBC obtained by the LDA algorithm (Figure 6). The number of publications on therapeutic target research and nanoparticle research has increased dramatically, with 15.4% and 15.7%. These results indicate these two are research focus in the future.

3.7. LDA and citation analysis results: TNBC research is based on technology that advances TNBC subtyping, new drug development, and clinical trials

Highly cited publications often represent the emergence of outstanding contributions, leading knowledge, or examples in the field. Attention was paid to the citations of publications within the TNBC field. All publications with a total of 490,599 citations, among which the top ten publications with the highest internal citations are listed in Table 3, the publication with the highest internal citations, 1,293, and the total citations of these 10 publications are 21,550. These publications focus on three categories, clinical characteristics of extensive population studies (15–17), clinical trials of new medications (18–21), and subtyping studies of TNBC (22–24). They represent researchers focused on discovering new molecular targets and developing multiple therapies such as Atezolizumab and

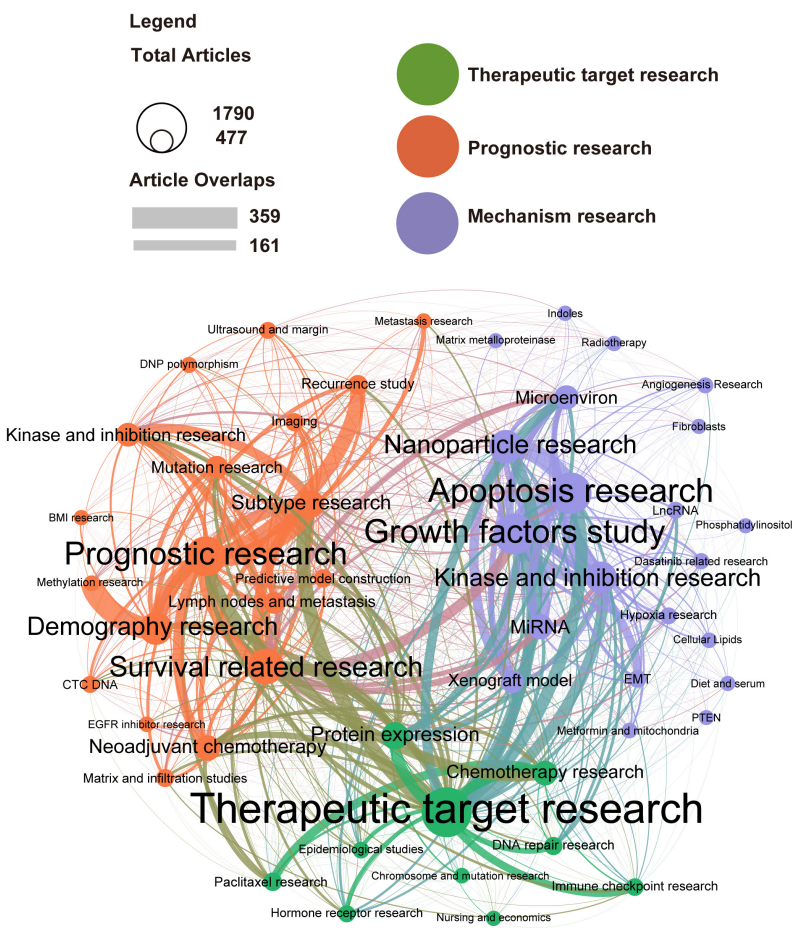

FIGURE 5
Latent Dirichlet Allocation (LDA) identified that the triple-negative breast cancer (TNBC) research is focused on three areas Therapeutic target research, Prognostic research, and Mechanism research. Topic cluster network studied by Latent Dirichlet Allocation: inter-and intra-relationships. Therapeutic target research (green), Prognostic research (orange), and Mechanism research (purple) are three major clusters in TNBC research. The circle size represents the number of publications on each topic; the line's thickness represents the weight of the connection between each topic.

Nab-Paclitaxel for treatment. Therefore, under the guidance of this research model, similar studies in the future can get more citations. On the other hand, combined with the steady increase of MeSH terms year by year, the lack of drastic changes suggests that TNBC research presents a stable and mature research model, that is, new drug development based on TNBC typing, target drug development, and clinical trials.

4. Discussion

We analyzed 16,826 publications in the field of TNBC from 2005 to 2022 using machine learning and NLP. Furthermore, we visualize and analyze the results from a macro perspective. Over the past 17 years, we found that TNBC-related publications have increased from none to 16,826 in 2021, with more extensive research content. TNBC research focuses on Therapeutic target research, Prognostic research, and Mechanism research. Research topics have changed over the years, and the current research focus is expected to be Therapeutic target research and Nanoparticle research, according to our LDA results.

Bibliometrics is a compelling analysis method to obtain information from massive texts quantitatively, and there are very few bibliometrics analyses on TNBC such as VOSviewer, Bibliographic Items Co-occurrence Matrix Builder (BICOMB), and CiteSpace. However, with the development of the publishing industry, these tools have difficulty applying to massive publication analysis due to their architecture, insufficient computer memory, and sharing protocols. Therefore, our research uses the LDA algorithm based on Python, an unsupervised topic model. Furthermore, our topic model is based on the publication's abstract, not on the keywords. It is easy to use with negligible memory consumption and can analyze massive publications.

We found that Therapeutic target research has always been research-focused because TNBC lacks effective therapeutic targets and has high heterogeneity (24, 25). Our research found that this part contains a variety of attempts, DNA repair research, immune checkpoint research, and protein expression. We only found 137 publications related to immune checkpoint research, and immunotherapy research is not closely related to the prognosis and mechanism research of TNBC. Several clinical studies are being carried out, including IMpassion130, KEYNOTE-355, and Impassion 131 (26–28). Some positive results can reduce the risk of death by up to 35%. However, more important is the research on the underlying mechanism and the exploration of various influencing factors, especially

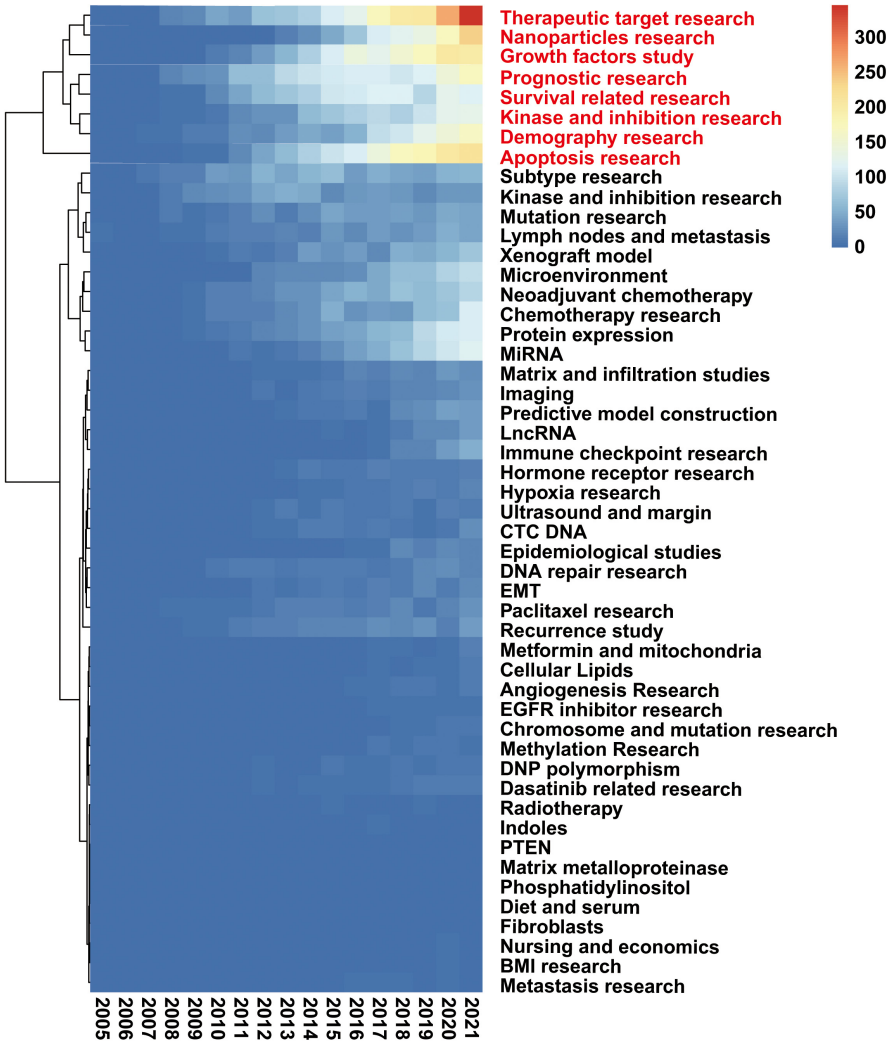

FIGURE 6
Therapeutic target research and Nanoparticles research are research focus. Heatmap presents the change of 50 research topics of triple-negative breast cancer (TNBC). Latent Dirichlet Allocation (LDA) generated all data. The topics marked in red are the research focus. The lighter the color in the figure, the more publications.

the extracellular matrix, hypoxia, and immune cell infiltration (29). In addition, immune checkpoint research has just started for five years, according to our results, and several medications have already been applied in the clinic. This research trend will continue, and immunotherapy will become a safe and effective treatment option.

The research scope of the TNBC mechanism is pervasive, covering the immune microenvironment and subtypes of TNBC. The successful subtyping provides a solid theoretical basis for the precision therapy of TNBC (30). Gene sequencing technology allows us to fully understand the mutation rate of TNBC, which is about 1.68 bp/Mb (31). Mutations occur in genes in multiple key signaling pathways such as PI3K/Akt/mTOR pathway, RAS/RAF/MEK pathway, JAK/STAT pathway, DNA repair pathway, and cell cycle checkpoint (32–34). Therefore, various treatments targeting the signal pathways are currently undergoing clinical trials. Some inhibitors have been used as potential medications for TNBC treatment, including PI3K, MEK, PARP, EGFR, VEGF, and AR inhibitors (32).

Triple-negative breast cancer subtyping has always been the focus of research. There is no unified standard based on the TNBC genome and cell heterogeneity. The first classification was based on Lehmann's gene expression analysis of breast cancer and constructed a "triple negative classification" and six subclassifications (24). In 2016, Lehmann's further research found that immunomodulatory (IM) patients are more likely to benefit from checkpoint inhibitor therapy (35). With the advancement of technology, such as the emergence of single-cell RNA sequencing, spatial transcriptomics, and radionics, and the further expansion of data volume, new technologies have provided new insights into the typing of TNBC and proposed guidance for treatment. Xie's research established a new prognostic model through the comprehensive analysis of multiple cell death patterns on more than 1,000 breast cancer patients, which can predict the clinical prognosis and drug sensitivity after TNBC surgery (36). In addition to technological progress, an in-depth understanding of the oncological course, mechanism of occurrence and development, and algorithm advances will provide a more detailed classification of TNBC.

TABLE 3 Top 10 publications of triple-negative breast cancer (TNBC) based on internal citations and Latent Dirichlet Allocation (LDA) results.

Reference title DOI	References	Internal citation	Total citation	LDA results
J Clin Invest. 2011 Jul; 121 (7): 2750-67 https://doi.org/10.1172/jci45014	Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies (24)	1,293	3,205	Protein expression
Clin Cancer Res. 2007 Aug 1; 13 (15 Pt 1): 4429-34 https://doi.org/10.1158/1078-0432.ccr-06-3045	Triple-negative breast cancer: clinical features and patterns of recurrence (23)	1,220	3,025	Subtype research
N Engl J Med. 2010 Nov 11; 363 (20): 1938-48 https://doi.org/10.1056/nejmra1001389	Triple-negative breast cancer (21)	1,062	2,501	Therapeutic target research
J Clin Oncol. 2008 Mar 10; 26 (8): 1275-81 https://doi.org/10.1200/jco.2007.14.4147	Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer (20)	700	1,909	Prognostic research
Cancer. 2007 May 1; 109 (9): 1721-8 https://doi.org/10.1002/cncr.22618	Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry (17)	550	1,483	Demography research
Clin Cancer Res. 2007 Apr 15; 13 (8): 2329-34 https://doi.org/10.1158/1078-0432.ccr-06-1109	The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes (16)	515	1,472	Subtype research
Nat Rev Clin Oncol. 2016 Nov; 13 (11): 674-690 https://doi.org/10.1038/nrclinonc.2016.66	Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease (15)	485	1,280	Therapeutic target research
N Engl J Med. 2018 Nov 29; 379 (22): 2108-2121 https://doi.org/10.1056/nejmoa1809615	Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer (19)	358	2,064	Immune checkpoint research
Lancet. 2014 Jul 12;384 (9938): 164-72 https://doi.org/10.1016/s0140-6736(13)62422-8	Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis (18)	335	2,113	Neoadjuvant chemotherapy
Ann Oncol. 2011 Aug; 22 (8): 1736-47 https://doi.org/10.1093/annonc/mdr304	Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011 (22)	311	2,498	Therapeutic target research

On the other hand, studies on operations and radiotherapy were rarely reported, especially for re-operations related to local-regional recurrence risk or distant metastasis. Many studies suggest that surgery is essential in treating distant metastases of cancers, such as colorectal cancer (37). In addition, many studies on other cancers, including pancreatic and colorectal cancer, demonstrated that the tumor microenvironment, especially the extracellular matrix, has been found to play an essential role in cancer metastasis, local recurrence, and chemotherapeutic drug resistance (38, 39). Many potential drugs are used due to their ability to target the extracellular matrix, such as PEGPH20 (an enzyme that targets matrix hyaluronic acid), pegilodecakin (a PEGylated IL-10) (40, 41). However, the study on extracellular matrix in TNBC is insufficient so far.

Although the research on TNBC has made significant progress in many aspects, the present research also found some research deficiencies on TNBC. There is a lack of research on TNBC from patients' perspectives, health economics, and hospice care. Although, at present, the 5 years overall survival rate of most tumors has been dramatically improved, helping tumor patients with psychological issues re-enter society will become a new important research topic (42). TNBC patients are more likely to relapse and metastasize than other breast cancer subtypes, resulting in more significant mental and economic pressure on patients and their families. Studies on patients with more prolonged survival can better understand TNBC and even other long-term survival tumors (43). In the future, we will face more challenges for patients with a long survival period of 5–10 years (44).

There are some limitations in the present study. Besides PubMed, several other databases, including Scopus, Web of Science, and Embase, could be used for bibliometric research. Although PubMed contains the highest quality peer-reviewed research and excludes irrelevant, non-peer-reviewed publications, the literature will provide detailed and comprehensive knowledge if other databases are explored simultaneously. Secondly, we considered that all publications publish more positive research results. Negative results and clinical participants' perspectives are naturally more difficult to be published. With the development of complete medical record texts, publication databases, and improved algorithms, it is reasonable for machine learning to play a more active auxiliary role in future clinical practice. The data presented in this study will hopefully help scientists understand the current status of TNBC research and design more relevant basic and clinical research projects.

5. Conclusion

We analyzed 16,826 TNBC publications through the NLP Method. TNBC research shows insufficiencies, especially in long-term survival-related research, and a lack of research from patients' perspectives. The publications mainly focused on three aspects: Therapeutic target research, Prognostic research, and Mechanism research. The research direction of TNBC may require the intervention of new technologies.

Data availability statement

The original contributions presented in this study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

KW initiated the project, analyzed the data, constructed analytical methods, and wrote the primary manuscript draft. XD initiated and supervised all aspects of the project and wrote the primary manuscript draft. CZ performed statistical analyses and contributed to the manuscript writing. DD helped interpret results and contributed to the statistical analyses. LZ contributed to the manuscript's revision in terms of writing and interpretation. ML contributed to the interpreting results and supervising statistical analyses. All authors contributed to the manuscript writing and read and approved the final version of the manuscript.

Funding

This work was supported by funds from China Scholarship Council in the form of a scholarship to KW (202006370023), Guangzhou Institute of Pediatrics/Guangzhou Women and Children's Medical Center to LZ (4001013-04 and 5001-4001008), and the National Natural Science Foundation of China (to ML 30771122 and to XD 82173374 and 81872167).

Acknowledgments

We would like to express our gratitude to Wen Yan, who supported the study by programming.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.999312/full#supplementary-material

References

1. Banerjee S, Tian T, Wei Z, Shih N, Feldman MD, Peck KN, et al. Distinct microbial signatures associated with different breast cancer types. *Front Microbiol.* (2018) 9:951. doi: 10.3389/fmicb.2018.00951

2. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature.* (2000) 406:747–52. doi: 10.1038/35021093

3. Pareja F, Reis-Filho JS. Triple-negative breast cancers - a panoply of cancer types. *Nat Rev Clin Oncol.* (2018) 15:347–8. doi: 10.1038/s41571-018-0001-7

4. Yi H, Wu M, Zhang Q, Lu L, Yao H, Chen S, et al. Reversal of HER2 negativity: an unexpected role for lovastatin in triple-negative breast cancer stem cells. *J Cancer.* (2020) 11:3713–6. doi: 10.7150/jca.39265

5. Tran BX, Latkin CA, Sharafeldin N, Nguyen K, Vu GT, Tam WWS, et al. Characterizing artificial intelligence applications in cancer research: a latent dirichlet allocation analysis. *JMIR Med Inform.* (2019) 7:e14401. doi: 10.2196/14401

6. Teles RHG, Moralles HF, Cominetti MR. Global trends in nanomedicine research on triple negative breast cancer: a bibliometric analysis. *Int J Nanomedicine.* (2018) 13:2321–36. doi: 10.2147/IJN.S164355

7. Buchlak QD, Esmaili N, Leveque JC, Farrokhi F, Bennett C, Piccardi M, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg Rev.* (2020) 43:1235–53. doi: 10.1007/s10143-019-01163-8

8. Jun I, Rich SN, Chen Z, Bian J, Prosperi M. Challenges in replicating secondary analysis of electronic health records data with multiple computable phenotypes: A case study on methicillin-resistant staphylococcus aureus bacteremia infections. *Int J Med Inform.* (2021) 153:104531. doi: 10.1016/j.ijmedinf.2021.104531

9. Feng C, Wu Y, Gao L, Guo X, Wang Z, Xing B. Publication landscape analysis on gliomas: how much has been done in the past 25 years? *Front Oncol.* (2019) 9:1463. doi: 10.3389/fonc.2019.01463

10. Li C, Liu Z, Shi R. A bibliometric analysis of 14,822 researches on myocardial reperfusion injury by machine learning. *Int J Environ Res Public Health.* (2021) 18:8231. doi: 10.3390/ijerph18158231

11. Wang K, Feng C, Li M, Pei Q, Li Y, Zhu H, et al. A bibliometric analysis of 23,492 publications on rectal cancer by machine learning: basic medical research is needed. *Therap Adv Gastroenterol.* (2020) 13:1756284820934594. doi: 10.1177/1756284820934594

12. Kumar R, Rani S, Awadh MA. Exploring the application sphere of the internet of things in industry 4.0: a review, bibliometric and content analysis. *Sensors.* (2022) 22:4276. doi: 10.3390/s22114276

13. Kumar R, Goel P. Exploring the domain of interpretive structural modelling (ism) for sustainable future panorama: a bibliometric and content analysis. *Arch Comput Methods Eng.* (2022) 29:2781–810. doi: 10.1007/s11831-021-09675-7

14. Traag VA. Faster unfolding of communities: speeding up the louvain algorithm. *Phys Rev E Stat Nonlin Soft Matter Phys.* (2015) 92:032801. doi: 10.1103/PhysRevE.92.032801

15. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol.* (2016) 13:674–90. doi: 10.1038/nrclinonc.2016.66

16. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin Cancer Res.* (2007) 13:2329–34. doi: 10.1158/1078-0432.CCR-06-1109

17. Bauer KR, Brown M, Cress RD, Parise CA, Caggiano V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer registry. *Cancer.* (2007) 109:1721–8. doi: 10.1002/cncr.22618

18. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet.* (2014) 384:164–72. doi: 10.1016/S0140-6736(13)62422-8

19. Schmid P, Adams S, Rugo HS, Schneeweiss A, Barrios CH, Iwata H, et al. Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *N Engl J Med.* (2018) 379:2108–21. doi: 10.1056/NEJMoa1809615

20. Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol.* (2008) 26:1275–81. doi: 10.1200/JCO.2007.14.4147

21. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. *N Engl J Med.* (2010) 363:1938–48. doi: 10.1056/NEJMra1001389

22. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thurlimann B, Senn HJ, et al. Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the st. gallen international expert consensus on the primary therapy of early breast cancer 2011. *Ann Oncol.* (2011) 22:1736–47. doi: 10.1093/annonc/mdr304

23. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res.* (2007) 13:4429–34. doi: 10.1158/1078-0432.CCR-06-3045

24. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* (2011) 121:2750–67. doi: 10.1172/JCI45014

25. Deng X, Faqing T, Rosol TJ. *Triple-Negative Breast Cancer*. Singapore: World Scientific (2020). p. 21–70. doi: 10.1142/11199

26. Cortes J, Cescon DW, Rugo HS, Nowecki Z, Im SA, Yusof MM, et al. Pembrolizumab plus chemotherapy versus placebo plus chemotherapy for previously untreated locally recurrent inoperable or metastatic triple-negative breast cancer (KEYNOTE-355): a randomised, placebo-controlled, double-blind, phase 3 clinical trial. *Lancet.* (2020) 396:1817–28. doi: 10.1200/JCO.2020.38.15_suppl.1000

27. Miles D, Gligorov J, Andre F, Cameron D, Schneeweiss A, Barrios C, et al. Primary results from IMpassion131, a double-blind, placebo-controlled, randomised phase III trial of first-line paclitaxel with or without atezolizumab for unresectable locally advanced/metastatic triple-negative breast cancer. *Ann Oncol.* (2021) 32:994–1004. doi: 10.1016/j.annonc.2020.08.2243

28. Emens LA, Adams S, Barrios CH, Dieras V, Iwata H, Loi S, et al. First-line atezolizumab plus nab-paclitaxel for unresectable, locally advanced, or metastatic triple-negative breast cancer: IMpassion130 final overall survival analysis. *Ann Oncol.* (2021) 32:983–93. doi: 10.1016/j.annonc.2021.05.355

29. Bou-Dargham MJ, Draughon S, Cantrell V, Khamis ZI, Sang QA. Advancements in human breast cancer targeted therapy and immunotherapy. *J Cancer.* (2021) 12:6949–63. doi: 10.7150/jca.64205

30. Lee YM, Oh MH, Go JH, Han K, Choi SY. Molecular subtypes of triple-negative breast cancer: understanding of subtype categories and clinical implication. *Genes Genomics.* (2020) 42:1381–7. doi: 10.1007/s13258-020-01014-7

31. Mittendorf EA, Philips AV, Meric-Bernstam F, Qiao N, Wu Y, Harrington S, et al. PD-L1 expression in triple-negative breast cancer. *Cancer Immunol Res.* (2014) 2:361–70. doi: 10.1158/2326-6066.CIR-13-0127

32. Islam R, Lam KW. Recent progress in small molecule agents for the targeted therapy of triple-negative breast cancer. *Eur J Med Chem.* (2020) 207:112812. doi: 10.1016/j.ejmech.2020.112812

33. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature.* (2012) 486:395–9.

34. Vanhaesebroeck B, Guillermet-Guibert J, Graupera M, Bilanges B. The emerging mechanisms of isoform-specific PI3K signalling. *Nat Rev Mol Cell Biol.* (2010) 11:329–41. doi: 10.1038/nrm2882

35. Lehmann BD, Jovanovic B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS One.* (2016) 11:e0157368. doi: 10.1371/journal.pone.0157368

36. Zou Y, Xie J, Zheng S, Liu W, Tang Y, Tian W, et al. Leveraging diverse cell-death patterns to predict the prognosis and drug sensitivity of triple-negative breast cancer patients after surgery. *Int J Surg.* (2022) 107:106936. doi: 10.1016/j.ijsu.2022.106936

37. Dijkstra M, Nieuwenhuizen S, Puijk RS, Timmer FEF, Geboers B, Schouten EAC, et al. Primary tumor sidedness, ras and braf mutations and msi status as prognostic factors in patients with colorectal liver metastases treated with surgery and thermal ablation: results from the amsterdam colorectal liver met registry (AmCORE). *Biomedicines.* (2021) 9:962. doi: 10.3390/biomedicines9080962

38. Gu Z, Du Y, Zhao X, Wang C. Tumor microenvironment and metabolic remodeling in gemcitabine-based chemoresistance of pancreatic cancer. *Cancer Lett.* (2021) 52:98–108. doi: 10.1016/j.canlet.2021.08.029

39. Song X, Xie D, Tan F, Zhou Y, Li Y, Zhou Z, et al. Intravascular emboli relates to immunosuppressive tumor microenvironment and predicts prognosis in stage III colorectal cancer. *Aging.* (2021) 13:20609–28. doi: 10.18632/aging.203451

40. Gourd E. PEGPH20 for metastatic pancreatic ductal adenocarcinoma. *Lancet Oncol.* (2018) 19:e81. doi: 10.1016/S1470-2045(17)30953-1

41. Hecht JR, Lonardi S, Bendell J, Sim HW, Macarulla T, Lopez CD, et al. Randomized phase iii study of folfox alone or with pegilodecakin as second-line therapy in patients with metastatic pancreatic cancer that progressed after gemcitabine (SEQUOIA). *J Clin Oncol.* (2021) 39:1108–18. doi: 10.1200/JCO.20.02232

42. Watkins CC, Kanu IK, Hamilton JB, Kozachik SL, Gaston-Johansson F. Differences in coping among African American women with breast cancer and triple-negative breast cancer. *Oncol Nurs Forum.* (2017) 44:689–702. doi: 10.1188/17.ONF.689-702

43. Mediratta K, El-Sahli S, D'Costa V, Wang L. Current progresses and challenges of immunotherapy in triple-negative breast cancer. *Cancers.* (2020) 12:3529. doi: 10.3390/cancers12123529

44. Ertas G, Basal FB, Ucer AR, Benzer E, Altundag MB, Demirci U, et al. Clinical features of metaplastic breast carcinoma: A single-center experience. *J Cancer Res Ther.* (2020) 16:1229–34.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 25 May 2023
DOI 10.3389/fmed.2023.1070499

OPEN ACCESS

EDITED BY
Jingjing You,
The University of Sydney, Australia

REVIEWED BY
Billy Peralta,
Andres Bello University, Chile
Abdel-Hameed Al-Mistarehi,
Johns Hopkins Medicine, United States

*CORRESPONDENCE
Guillermo Droppelmann
✉ guillermo.droppelmann@meds.cl

RECEIVED 14 October 2022
ACCEPTED 09 May 2023
PUBLISHED 25 May 2023

CITATION
Saavedra JP, Droppelmann G, García N, Jorquera C and Feijoo F (2023) High-accuracy detection of supraspinatus fatty infiltration in shoulder MRI using convolutional neural network algorithms.
Front. Med. 10:1070499.
doi: 10.3389/fmed.2023.1070499

High-accuracy detection of supraspinatus fatty infiltration in shoulder MRI using convolutional neural network algorithms

Juan Pablo Saavedra[1], Guillermo Droppelmann[2,3,4]*, Nicolás García[2], Carlos Jorquera[5] and Felipe Feijoo[1]

[1]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, [2]Research Center on Medicine, Exercise, Sport and Health, MEDS Clinic, Santiago, Chile, [3]Health Sciences PhD Program, Universidad Católica de Murcia UCAM, Murcia, Spain, [4]Principles and Practice of Clinical Research (PPCR), Harvard T. H. Chan School of Public Health, Boston, MA, United States, [5]Facultad de Ciencias, Escuela de Nutrición y Dietética, Universidad Mayor, Santiago, Chile

Background: The supraspinatus muscle fatty infiltration (SMFI) is a crucial MRI shoulder finding to determine the patient's prognosis. Clinicians have used the Goutallier classification to diagnose it. Deep learning algorithms have been demonstrated to have higher accuracy than traditional methods.

Aim: To train convolutional neural network models to categorize the SMFI as a binary diagnosis based on Goutallier's classification using shoulder MRIs.

Methods: A retrospective study was performed. MRI and medical records from patients with SMFI diagnosis from January 1st, 2019, to September 20th, 2020, were selected. 900 T2-weighted, Y-view shoulder MRIs were evaluated. The supraspinatus fossa was automatically cropped using segmentation masks. A balancing technique was implemented. Five binary classification classes were developed into two as follows, A: 0, 1 v/s 3, 4; B: 0, 1 v/s 2, 3, 4; C: 0, 1 v/s 2; D: 0, 1, 2, v/s 3, 4; E: 2 v/s 3, 4. The VGG-19, ResNet-50, and Inception-v3 architectures were trained as backbone classifiers. An average of three 10-fold cross-validation processes were developed to evaluate model performance. AU-ROC, sensitivity, and specificity with 95% confidence intervals were used.

Results: Overall, 606 shoulders MRIs were analyzed. The Goutallier distribution was presented as follows: 0=403; 1=114; 2=51; 3=24; 4=14. Case A, VGG-19 model demonstrated an AU-ROC of 0.991±0.003 (accuracy, 0.973±0.006; sensitivity, 0.947±0.039; specificity, 0.975±0.006). B, VGG-19, 0.961±0.013 (0.925±0.010; 0.847±0.041; 0.939±0.011). C, VGG-19, 0.935±0.022 (0.900±0.015; 0.750±0.078; 0.914±0.014). D, VGG-19, 0.977±0.007 (0.942±0.012; 0.925±0.056; 0.942±0.013). E, VGG-19, 0.861±0.050 (0.779±0.054; 0.706±0.088; 0.831±0.061).

Conclusion: Convolutional neural network models demonstrated high accuracy in MRIs SMFI diagnosis.

KEYWORDS

classification, deep learning, fatty infiltration, MRI, supraspinatus

Introduction

Rotator cuff tears (RCTs) are among the most critical musculoskeletal conditions of the shoulder (1). This prevalence affects worldwide (2), resulting in direct and indirect economic burdens for patients and healthcare systems (3). Furthermore, this progressive degenerative condition (4) affects both sexes, and its incidence in the general population increases with age (5).

Image medical analysis plays a significant role in diagnosis and the optimal detection of the tear magnitude, allowing therapeutic planning resolutions, including physical therapy and surgical repair (6). Many imaging techniques have been developed for the detection of RCTs. Magnetic resonance imaging (MRI) presents the highest diagnostic value (sensitivity and specificity) for detecting any lesion (7, 8), especially for evaluating the integrity of the rotator cuff in tear size. Another essential radiological aspect of assessing the MRI shoulder is atrophy and fatty infiltration. Patients with a low stage of fatty infiltration have significantly better outcomes than those with a severe condition, since patients who present a re-tear are the most affected (9, 10).

For this reason, to determine the magnitude the SMFI, Goutallier et al. proposed a classification with five stages ranging from 0 to 4 (11). However, the original proposal has been adapted with MRI by Fuchs et al. (12) using three stages, combining stages zero and one as normal, two as moderate, and three with four as severe fatty infiltration. In the MRI adaptation of the classification, there has been controversy regarding the ideal technique for grading (13).

One of the most significant challenges in image diagnosis is reducing the variability between observers in assessing rotator cuff muscle quality on MRI (14). Recent studies have implemented the use of Artificial Intelligence (AI), Machine Learning (ML), and particularly Deep Learning (DL) techniques to improve the accuracy of diagnosis, helping radiologists with the interpretation of imaging data (15). This process has been facilitated by developing AI and ML tools and incorporating these into the diagnostic support of medical images (16). Also, as it is common to have small datasets in medical imaging, transfer learning using well-trained non-medical ImageNet datasets has shown promising results for medical image analysis in recent years. Some of the most used DL architectures in medical imaging analysis (17) include Inception-v3 (18), ResNet-50 (19), and VGG-19 (20).

Random forest (RF) and DL techniques, such as convolutional neural network (CNN), have been used to identify the segmentation of rotator cuff muscles on MRI (21). Also, automatic algorithms have been implemented to detect supraspinatus muscle atrophy (22), and detection of supraspinatus tears on MRI (23). However, such algorithms have not yet been implemented to detect this structure's fatty infiltration level. Incorporating these artificial intelligence tools would improve diagnostic precision and patient prognosis. Kim (22) demonstrated CNNs' ability to segment the supraspinatus muscle and supraspinatus fossa to calculate their ratio in an MRI dataset. Similarly, Ro and collaborators (24) developed a model that analyzes the muscle proportion in the supraspinatus fossa and quantifies fatty infiltration in MRI through Otsu thresholding (25). The Otsu thresholding is used to create pixel clusters from grayscale images and optimizes the pixel intensity value to establish foreground and background. In this case, the foreground would be fat, and the background would be muscle. This method is highly influenced by the difference in pixel intensity due to fatty infiltration level. This was addressed by computing a standard deviation for every Goutallier level. Using this method, Otsu thresholding showed 0.06; 4.68; 20.10; 42.86; and 55.76 for grades 0, 1, 2, 3, and 4, respectively. Finally, in the context of RCT and fatty infiltration imaging analysis, Taghizadeh (26) developed a convolutional neural network model to automatically quantify and characterize the degeneration of rotator cuff muscles from CT images. The backbone of this model is the U-Net architecture, which can segment muscle fossa into a pre-morbid state. Most convolutional neural network models have been used to segment regions of interest, including supraspinatus, infraspinatus, and subscapular muscles. Since Goutallier's grade scale is a qualitative method and diagnoses are highly influenced by clinicians' and experts' intuitive judgment, literature has claimed that classification of Goutallier's grade via DL methods is not an easy task (24).

To assess this hypothesis, this study aims to build a DL architecture to classify patients as "risky" or "not risky" based on the Goutallier's supraspinatus fatty infiltration classification from shoulder MRI to help clinicians and medical staff in decision-making. Results demonstrate that DL models provide high accuracy and classification accuracy (discriminatory capacity) for Goutallier's supraspinatus fatty infiltration levels.

Materials and methods

Study design

This study was designed as a retrospective and one site study. It was written following the Strengthening the Reporting of Observation studies in Epidemiology (STROBE) guideline. All patients record were obtained from a MRI exam at MEDS Clinic in Santiago, Región Metropolitana, Chile. This study started on September 25th, 2020.

Datasets characteristics

The dataset used in this work comprises MRI and medical records from patients with an SMFI diagnosis who underwent examinations from January 1st, 2019, to September 20th, 2020. MRI images were saved in DICOM format, a widely used file format in medical imaging contexts. This format can save images, patient information, and study characteristics in one file. Each MRI image in the data set is obtained from a shoulder T2-weighted Y-view. The patient data were anonymized before being analyzed descriptively.

The initial dataset contained 900 MRI studies. But 669 images had valid annotations. Then, a musculoskeletal radiologist labeled the images based on Goutallier's fatty infiltration level. Two labeled images were excluded due to missing label records, and one was excluded because it was not conclusive for fatty infiltration analysis. After this process, 666 images were selected to perform manual segmentation. Sixty images had pixel configuration errors, and thus no segmentation could be done. The final dataset consists of 606 images, Figure 1.

To perform the labeling process, we developed a simple Python software, Figure 2, that reads a folder with all the images to be annotated and then shows the MRI image one at a time. The radiologist selects the diagnosis for that MRI image. The program creates a two-field JSON file with the decision made for the

professional for each image. One field is the image ID, and the other is the label record selected by the radiologist. These labels are our study's ground truth.

Statistical analysis

Dataset was analyzed and statistical tests were computed. For the analysis, python (with libraries such scipy) were used. Normality tests were performed. Statistical differences between groups were computed using the Mann–Whitney U test or t-test. A value of p of 0.05 was used to measure statistical significance. Descriptive analysis over the age of the patients was also performed and presented as mean and standard deviation (m ± sd). Percentages and frequencies are presented as statistical description for categorical.

Models' performances were computed and compared using accuracy, sensitivity, specificity, and AU-ROC. A binary classifier outputs one of two possible values for a given input, 0 or 1. For every input there is an actual expected output, which is also 0 or 1. Table 1, also known as confusion matrix, shows the four possible outcome situations.

We computed accuracy, sensitivity, specificity as follows:

- Accuracy: (TN + TP)/(TN + FP + FN + TP)
- Sensitivity (True positive rate): TP/(TP + FN)
- Specificity: TN/(TN + FP).

Area under the receiver operator curve or (AU-ROC) is a measure of the performance of the classifier regardless the threshold defined to translate probability scores to class decision. The horizontal axis corresponds to recall, or sensitivity, and the vertical axis corresponds to the precision, computed as TP/(TP + FP). As both axes are limited to 1, the maximum value of the area under the curve inside the square is 1, therefore, the closer to 1 the better the classifier. A random classifier will have an AU-ROC equal to 0.5.

In the case of the model performance, 95% confidence interval over the mean for the metrics, such as accuracy, sensitivity, specificity, and AU-ROC.

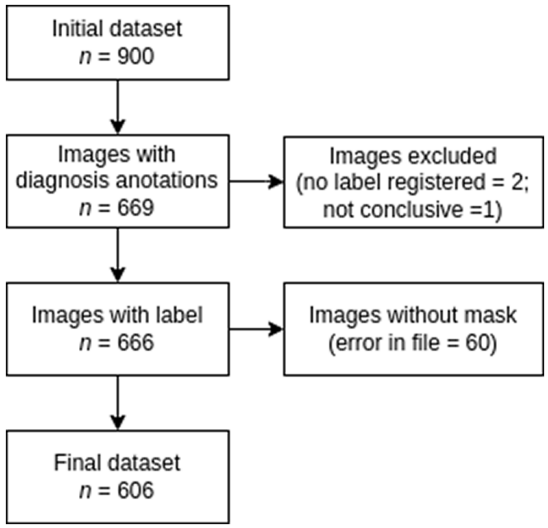

FIGURE 1
Flowchart for dataset selection.

Data preparation

The data preparation consisted of two main steps. First, the correct labeling of each image and the manual segmentation of the region of

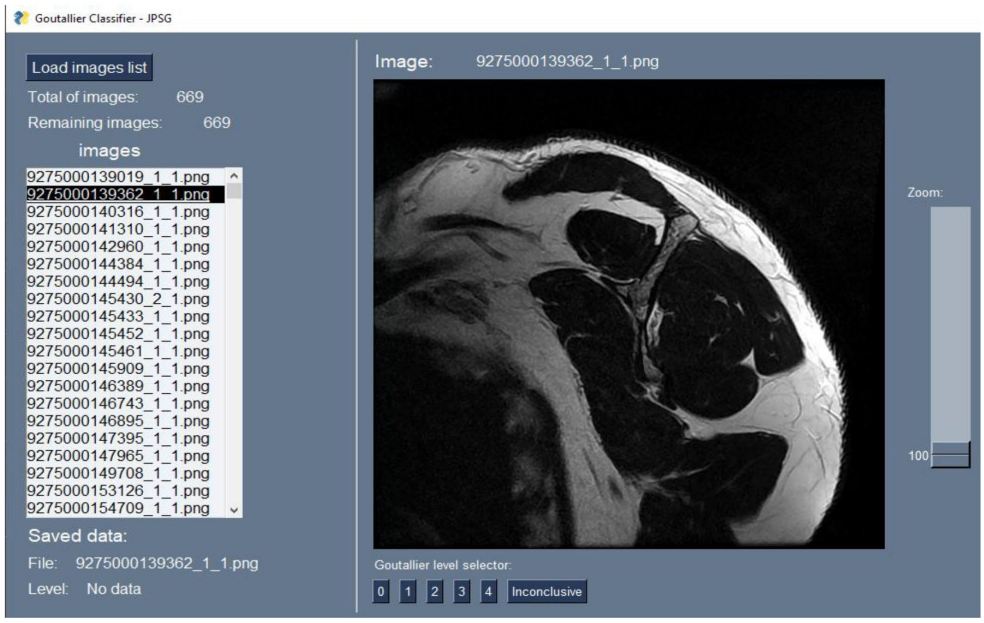

FIGURE 2
Custom software interface.

interest (ROI). All data in DICOM file format was processed with the MicroDICOM software to export images to PNG format. This allowed us to use fewer computational resources, as extracting images on the fly was unnecessary. Also, some Python libraries, such as PySimpleGUI, used to create the custom labeling software, only accept PNG format as input. We set the exported image resolution to the same as the original to avoid further mismatches between the image and its segmentation mask.

Regarding the segmentation of the ROI, the original DICOM files were used to create manual segmentation (identify the ROI in each image). The segmented areas were the supraspinatus fossa and the supraspinatus muscle. Figure 3 shows a sample segmentation. Panel (a) displays the original image, panel (b) the manually created segmentation masks, and panel (c) the segmented area masks. Each MRI image was segmented using the ITK-Snap software (27). At the end of the data preparation process, we obtained the original MRI images in PNG format, the segmentation masks, and label information for every image. The data preparation workflow is shown in Figure 4.

TABLE 1 Confusion matrix.

Classifier		Predicted	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

The definition and fatty infiltration criteria

We based our criteria on Goutallier's fatty infiltration definitions. The original paper proposed five levels of fatty infiltration (zero to four) about the qualitative presence of fat in the muscle. A level of zero means there is no fat in the muscle. As fatty infiltration increases, Goutallier's scale assigns a greater value. A level four means that there is more fat than muscle present. Figure 5 shows a representative MRI for every Goutallier's fatty infiltration level.

As shown in Table 2, we studied DL techniques' discriminatory (binary classification) power using five cases. In each case, we defined a positive and negative class composed of different Goutallier levels. Samples that belonged to the positive class were labeled as 1. Samples that belonged to the negative class were labeled as 0. The base case (case A) was used to assess the classification accuracy of no or low

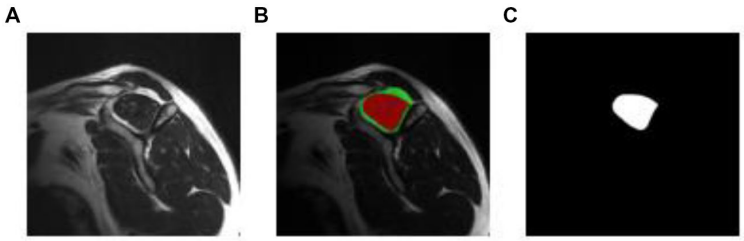

FIGURE 3
Manual segmentation process. Original, manual segmentation from ITK-Snap, and mask result, in figures **(A–C)**, respectively.

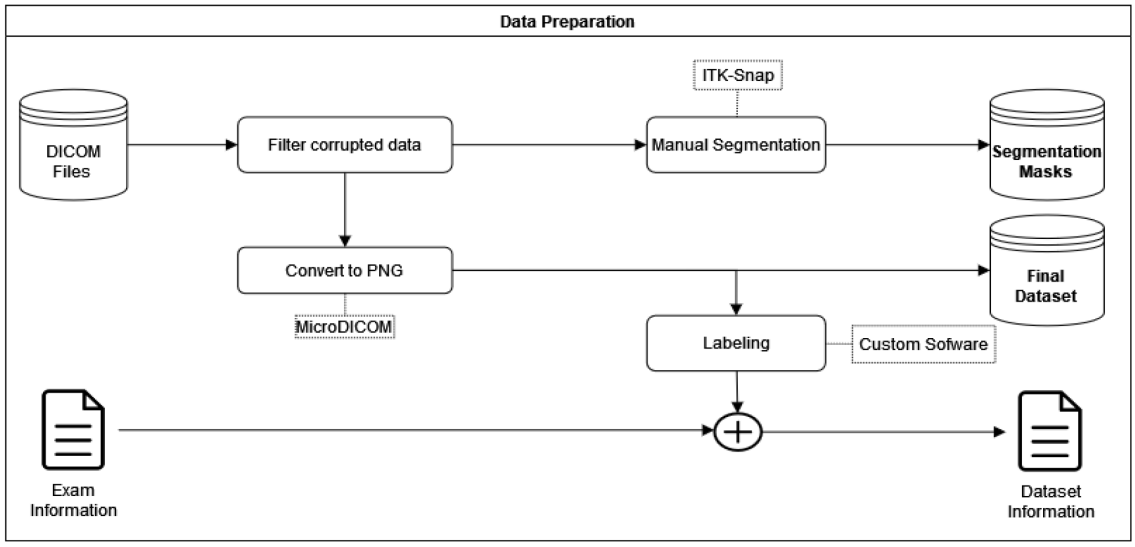

FIGURE 4
Data preparation.

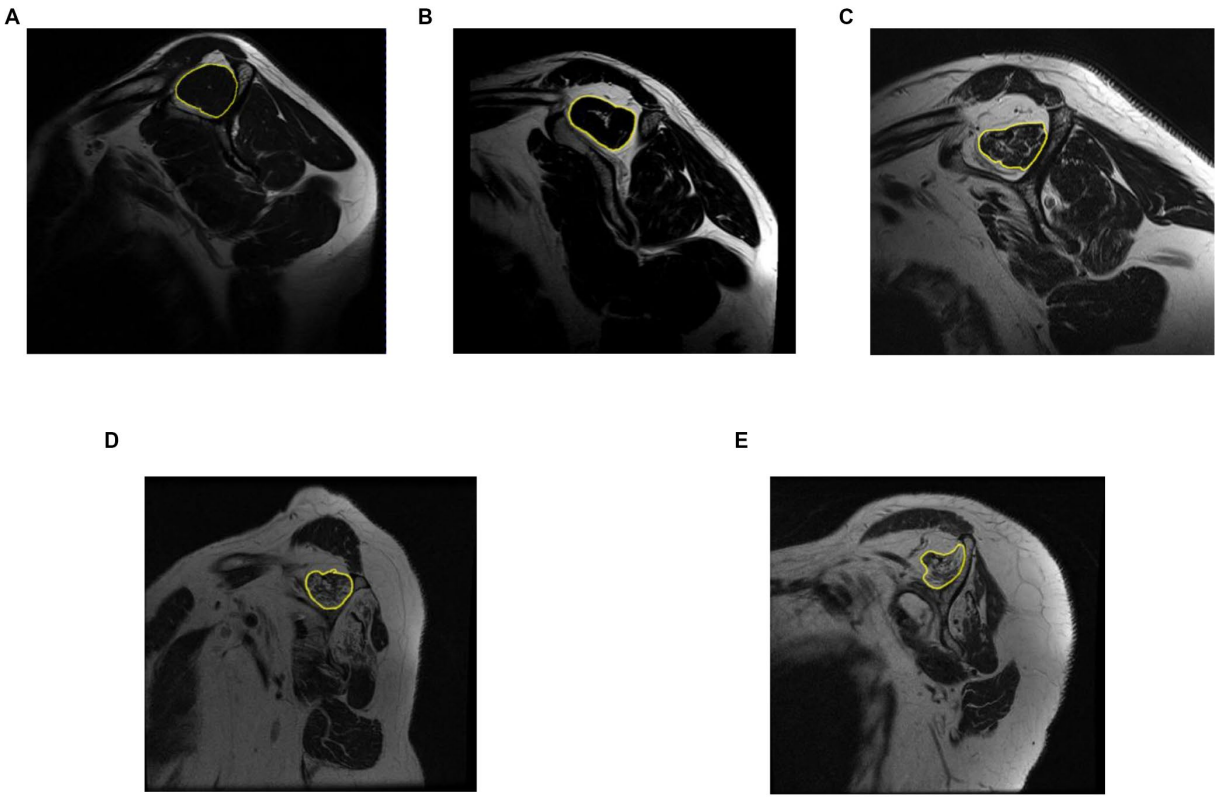

FIGURE 5
Representative MRI for each Goutallier's fatty infiltration scale. Level 0,1, 2, 3, and 4, are shown in sub-image **(A–E)**, respectively.

TABLE 2 Class designation in every case for fatty infiltration levels.

Case	Fatty infiltration levels in		Set size	
	Negative class	Positive class	Negative class	Positive class
A	0, 1	3, 4	517	38
B	0, 1	2, 3, 4	517	89
C	0, 1	2	517	51
D	0, 1, 2	3, 4	568	38
E	2	3, 4	51	38

fatty infiltration (Goutallier 0 and 1) against high fatty infiltration (Goutallier 3 and 4). Goutallier level 2 is not considered in this case. This allowed us to assess whether the DL techniques can differentiate between no-fatty and high-fatty infiltration cases. Cases B to E is used as a sensitivity analysis of the classification capacity of the DL techniques.

Based on the above definition of cases, a sample that belonged to class 1 (positive) was considered "risky." A sample that belonged to class 0 (negative) was considered "no risky." A few random samples from class 0 and class 1 are shown in Figure 6 for the case A. This classification is used since we aimed to help clinicians make decisions about proper treatment for patients based on the quality of the supraspinatus muscle. In every case, the positive and negative classes were different.

Model development and training

Three models based on well-known architectures were trained: VGG-19, Inception-v3, and ResNet-50, and compared their performance in terms of classification accuracy. For every model, the learning rate and average time were processed. Figure 7 shows the general training workflow. In terms of the architecture, the convolutional layers for every model remained the same as in the original, and only the classifier was modified. We replaced the last layer of every model with a 1,000-unit wide and SoftMax activation function with a single neuron with a sigmoid activation function because our problem was binary classification. In the case of VGG-19, we also reduced the size of the most outer fully connected layer from 4,096 neurons to 2048, which helped to avoid overfitting, Figure 8. We used transfer learning from ImageNet weights to train the models. The backbone of the original architecture was used as a feature extractor, and its layers were frozen. Then, only the fully connected layer parameters were optimized. In addition, every model architecture was created to admit three-channel images (RGB) as input. We simulate an RGB image from a gray-scale MRI by copying the same channel two times. Then, the three versions of the same single channel were stacked into a three-channel image.

Stratified k-fold cross validation

As we had a small dataset, stratified k-fold cross-validation was performed (28, 29). This method allowed us to use most of the data

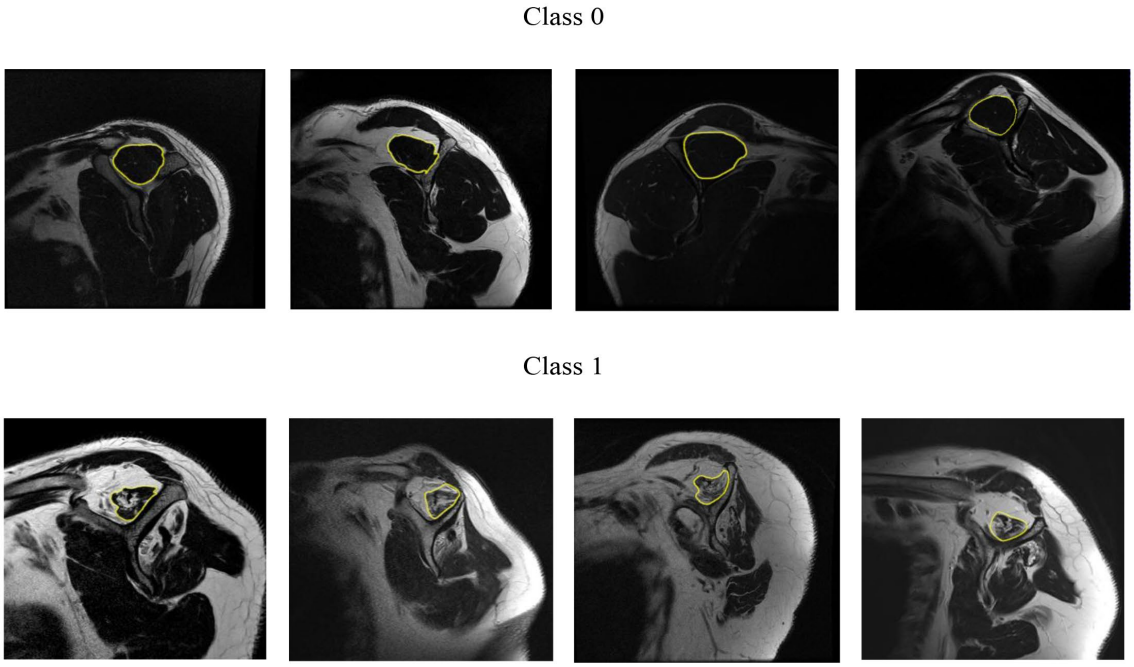

FIGURE 6
Random samples from class 0 and class 1 for the case A.

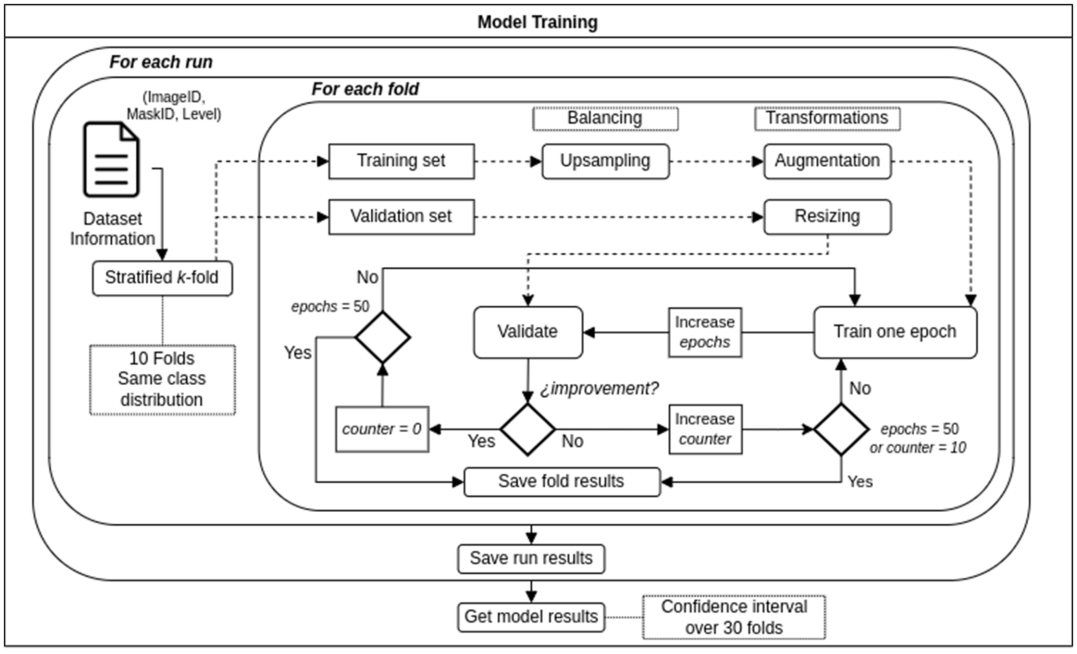

FIGURE 7
Training workflow.

for training and reduce the impact of the data selection in the results as would happen in a 20/80 random split, for example. We choose k equals to 10 and thus, 10 subgroups from the original data were created. That the cross-validation process is stratified means that every subgroup maintains the same class distribution of the original dataset. In each of the 10 training runs nine groups were used for training and one group for validation. We repeat three times the complete process of creating the 10 subgroups and running the training process. The performance of the model is calculated as the average of 30 training runs, and the confidence intervals for each were also found. The training and validation process based on stratified k-fold cross-validation follows the methodology described in (28, 29) when models are trained using small datasets.

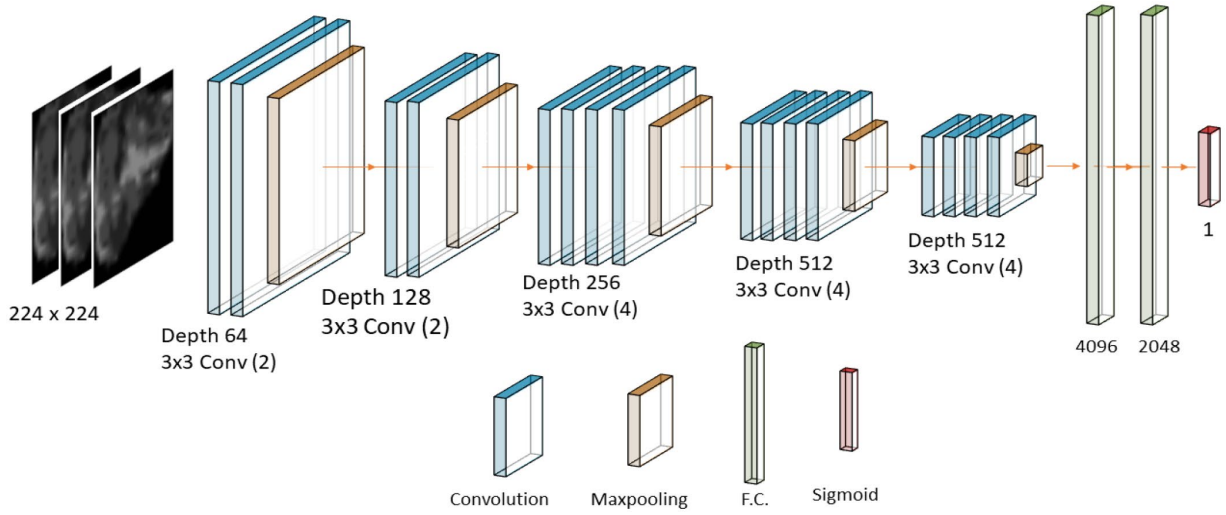

FIGURE 8
Diagram of the VGG/19 architecture.

Random data split

Additional to the assessment of the DL models using stratified K-fold cross validation, we evaluate the DL architectures using a new data set which has not been used during the training process. To do so, we trained the DL architectures using a random train/validation/test (70%/20%/10%, respectively) split. Downsampling of the majority class is performed over the training data only. The learning rate was set to 1e-06, 1e-04 and 1e-03 for VGG-19, Resnet50, and Inception V3, respectively. We train the model for 30 epochs and compute its accuracy, specificity, and sensitivity using the external new test data set (10% of the existing data) not used in training.

Augmentation and data balancing techniques

The data was highly imbalanced. This could lead the model to learn better from the most represented class than the minority class or lead to a highly overfitted model. We performed a balancing technique on the minority class to avoid or minimize these problems. In every 10 cross-validation processes, we over-sample the minority class on the training set until both classes have approximately the same number of samples. The validation set in the K-fold cross validation process remains imbalanced to validate the model similar to the real-world collection of images. Data augmentation was also performed on every image from the training set that was fed to the model. Augmentation is accomplished by rotating any grade value in ±35° and horizontally flipping with a probability of 0.5.

Training and optimization of hyper-parameters

All the DL models were trained using the Adam optimizer in standard configuration (weight decay = 0.9; beta = 0.999) for 50 epochs. The training process was stopped if there were no improvements in the last 10 epochs, and the best performance was saved. We only optimized the learning rate.

Before we fed the DL model with data, the region of interest was obtained from the segmentation mask for every image. This process is carried out automatically by the algorithm. It took the original image and the corresponding mask and cropped the region of interest. Then, only the ROI was fed to the DL models. The size of the input image was determined by the model's architecture requirements, which are 224 [px] squared images for the VGG-19 and the ResNet-50 architectures, and 299 [px] squared images for the InceptionV3. The cropped image was resized to meet those requirements.

Results

Statistical analysis results

A total of 606 patients (55% were males) with 606 MRI with RCTs were included in our analysis. The patient's average age was 55.1 ± 13.2 years. Data demonstrated the presence of all different Goutallier levels in imagological exams. An asymmetrical Goutallier distribution was found. More than 82% of the images belong to the 0 and 1 grades, showing an imbalance toward low fatty infiltration, as follows: Goutallier 0 (66.50%); Goutallier 1 (18.81%), Goutallier 2 (8.42%), Goutallier 3 (3.96%), and Goutallier 4 (2.31%). Also, the female group has more samples in higher grades than the male without statistical significance. The distribution of patient data is shown in Table 3.

Model performance

The learning rate used in every case and model and the average processing time were identified in Table 4. The shortest time was registered in the E case, using the Inception-v3 model with

TABLE 3 Quantity and proportions of sex by Goutallier's level.

Goutallier Level	*N* (%)	Female		Male		Value of *p*	
		N (%)	Age mean (SD)	*N* (%)	Age mean (SD)	*N*	Age
0	403 (66.50)	140 (35)	53.06 (10.55)	263 (65)	49.24 (13.13)	0.477	***
1	114 (18.81)	74 (65)	61.50 (10.37)	40 (35)	63.58 (8.17)	0.465	0.371
2	51 (8.42)	31 (61)	66.65 (9.53)	20 (39)	66.40 (10.13)	0.447	0.992
3	24 (3.96)	16 (67)	68.88 (7.74)	8 (33)	64.25 (7.59)	0.424	0.230
4	14 (2.31)	13 (93)	67.31 (7.33)	1 (7)	N.A.	0.354	0.8
Total	606 (100)	274 (45)	58.47 (11.67)	332 (55)	52.42 (13.81)	0.483	

Mann–Whitney or *t*-test were used to compute the significance (alpha 0.05). ***, statistically significant

TABLE 4 Learning rate and average processing time (C.I. 95%) for every case and model.

Case	Model	Learning rate	Processing time	Max. epochs
A	VGG-19	10^{-6}	3.51 ± 0.20	31.6 ± 3
	ResNet-50	10^{-4}	2.35 ± 0.22	20.7 ± 3.3
	Inception-v3	10^{-3}	1.55 ± 0.14	11.8 ± 2.5
B	VGG-19	10^{-6}	3.83 ± 0.27	33.1 ± 2.5
	ResNet-50	10^{-3}	1.12 ± 0.34	6.8 ± 2.2
	Inception-v3	10^{-3}	1.40 ± 0.19	9.3 ± 2.3
C	VGG-19	10^{-6}	3.87 ± 0.35	33.3 ± 2.9
	ResNet-50	10^{-5}	2.91 ± 0.22	26.9 ± 2.8
	Inception-v3	10^{-4}	2.98 ± 0.53	22.3 ± 2.9
D	VGG-19	10^{-6}	3.20 ± 0.25	27 ± 3.2
	ResNet-50	10^{-5}	3.09 ± 0.46	25.9 ± 2.9
	Inception-v3	10^{-3}	1.40 ± 0.01	8.8 ± 2.0
E	VGG-19	10^{-3}	0.42 ± 0.03	11.6 ± 3.0
	ResNet-50	10^{-4}	0.86 ± 0.20	22.5 ± 3.6
	Inception-v3	10^{-4}	0.34 ± 0.14	9.4 ± 4.0

0.34 ± 0.14 h. These results depend on the maximum number of epochs that the model runs until reaching its best validation loss, and thus, the training process is stopped and the training ends. In some cases, it is less than 50 epochs. In addition, the smaller the total size of the training set, the less time it takes to complete the training process. The E case has only 89 samples in total. On the other hand, the longest recorded time was registered in the VGG-19 model in the C case with 3.87 ± 0.35 h.

The DL architectures demonstrated outstanding performance using a shoulder MRI dataset. With a 10-fold cross-validation process, data was randomly divided into 10 non-overlapping folds. Nine folds were used as training sets and one as a validation set. The process was repeated three times; thus, three runs were obtained. This led to an average of 30 training loops.

Figure 9 shows the validation loss and AU-ROC curves for every model at every run. The three architectures show a decreasing validation loss at every epoch. At the beginning of the training process, the VGG-19 loss validation starts at 0.739 ± 0.006, 0.632 ± 0.007, and 0.631 ± 0.005 in the first, second, and third runs, respectively. Then, in the end, the validation loss was reduced to 0.225 ± 0.0053. In the case of Inception-v3, there is noticeably different behavior in one of the runs. This up-and-down loss value for the validation set could probably be explained due to the randomness in the process and the fact that the model could find a local minimum near the end. In any case, the last epoch showed an improvement in the validation loss value, and thus, it was recorded. Table 5 shows the starting value for the validation loss for every model. The model was run for a maximum of 50 epochs. We track the evolution of the loss function value. If the loss function did not decrease during 10 epochs, then the training process was terminated, and the results were computed.

The results confirm an optimized loss function. The loss function converges to zero as the learning progresses in the validation processes.

The model returns a value between 0 and 1, corresponding to the likelihood that the image belongs to the positive class. The value is then converted to binary based on a threshold. As the threshold value in our study, we utilized 0.5. The class will be considered positive if the model outputs a value greater than that. In contrast, if the model outputs a value lower than that threshold, the decision will be categorized as negative. One can compute the false positive and true positive rates under thresholds. The ROC curves in Figure 9 demonstrate the high performance of the models for various threshold values. The closer the curve is to (0.0, 1.0), the better the performance. To quantify curves, the area under the ROC curve was used. For our case A, VGG-19, ResNet-50, and Inception-v3 achieved 0.991 ± 0.003, 0.992 ± 0.003, and 0.991 ± 0.004, respectively for the area under the ROC curve (AU-ROC). Also, as shown in Figure 10, VGG-19 and ResNet-50 models showed the better performance when comparing precision-recall curves. When analyzing the per class prediction, the three models showed better performance in the negative class than in the positive class, which has fewer samples. Table 6 shows the confusion matrix for each model.

Subgroup analysis

A subgroup analysis was developed to determine the best combination of binary classes for Goutallier fatty infiltration level detection. Accuracy, sensitivity, specificity, AU-ROC, and loss performance for every single convolutional neural network model after three runs of 10 training cycles each are shown in Table 7. The reported metrics values shown are based on the results obtained from the repeated cross validation process. The process allowed us to have several validation groups and hence estimate

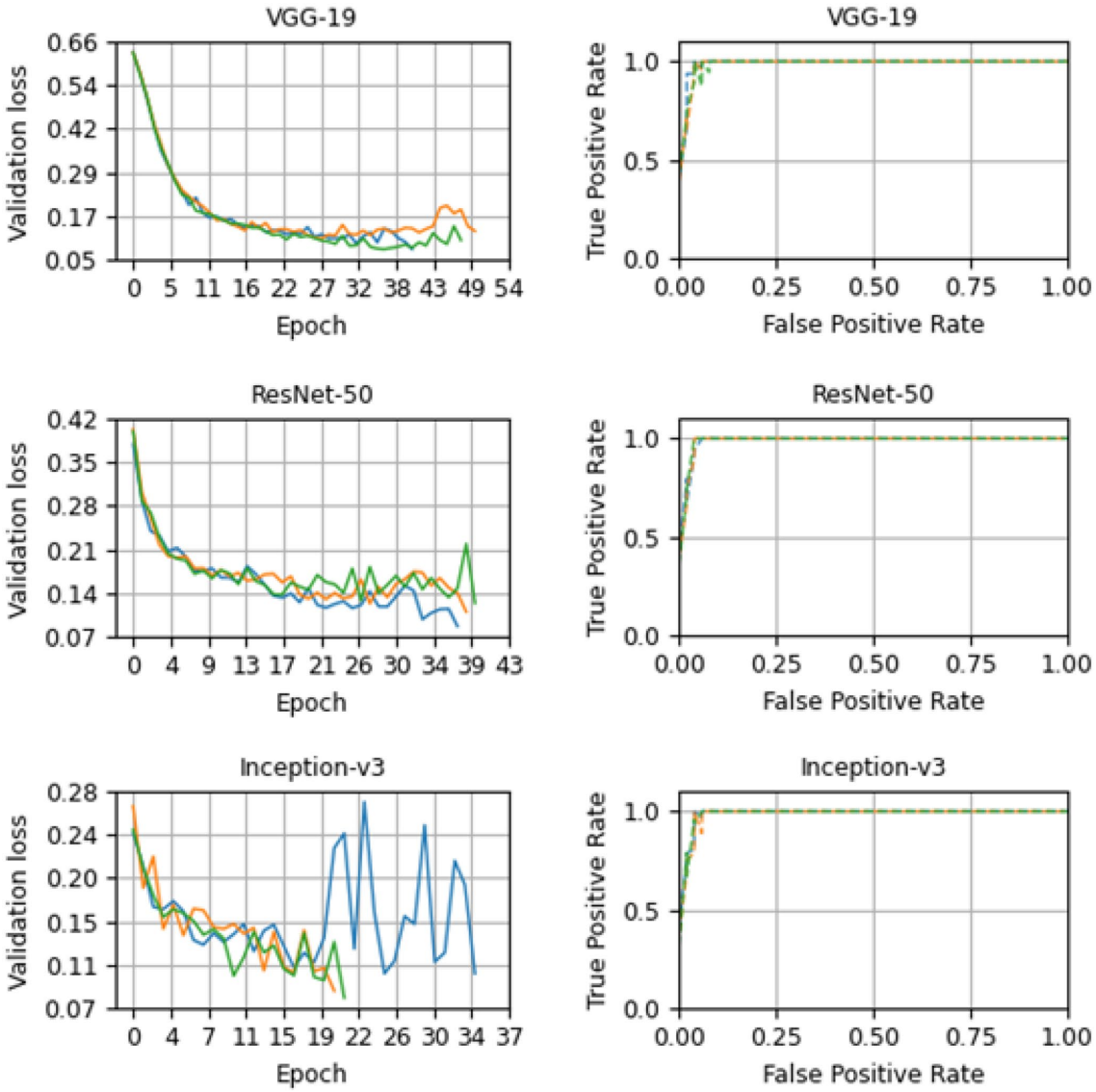

FIGURE 9
Loss and receiver operator curve plots for VGG-19, ResNet-50, and Inception-v3 models for base case (A). The results for the first, second, and third run are in color green, orange and blue, respectively.

TABLE 5 Confidence intervals (95%) for the starting validation loss in each run.

Model	Run 1	Run 2	Run 3
VGG-19	0.739 ± 0.006	0.632 ± 0.007	0.631 ± 0.005
ResNet-50	0.379 ± 0.024	0.403 ± 0.031	0.400 ± 0.047
Inception-v3	0.239 ± 0.050	0.265 ± 0.037	0.243 ± 0.068

the mean and the confidence level of each model in every experiment. Since DL models tend to learn the training data well, we do not report the training accuracy. Instead, we provide the evolution of the Loss Function, which depicts how the training error (learning process of the model) evolves. We also clarify this in the revised manuscript.

Excellent performance for the three architectures in every case was demonstrated. In three out of four cases, all model configurations had AU-ROC values higher than 0.91 on average and thus performed well when classifying fatty infiltration levels. In the base case, the models got an AU-ROC mean value of over 0.99, the highest among the cases. Here, models had to separate lower to no fatty infiltration images from high to extreme fatty infiltration levels, which were very dissimilar. In addition, sensitivity and specificity for this case are more homogeneous among models. This means that the models perform well when classifying negative and positive samples as the false positive rate and true positive rate are over 0.92, except for Inception-v3, which has a lower value for sensitivity. On the other hand, the same architecture showed a higher specificity, with a mean value of 0.981.

Random split performance

We also trained the model using a random train/validation/test split (training size: 413, validation size: 104, testing size: 58). Only the training data was down-sampled in order to account for unbalanced

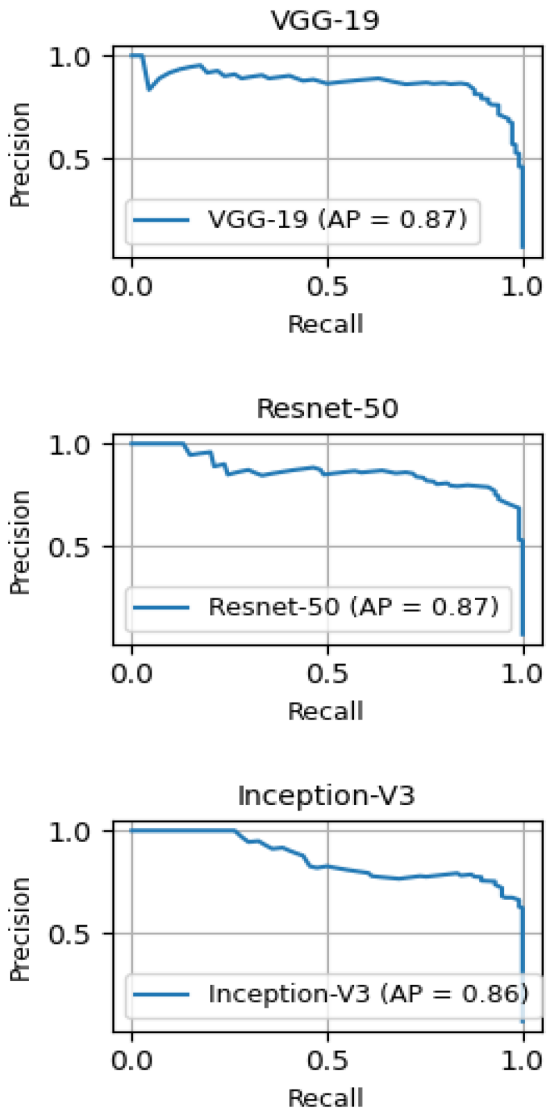

FIGURE 10
Mean Precision-recall curves for VGG-19, ResNet-50, and Inception-V3 models for the base case (A).

TABLE 6 Confusion matrix of VGG-19, Resnet-50, and Inception-V3 models for the case A validation set.

VGG-19		Predicted	
		0	1
Actual	0	1,512	39
	1	6	108

Resnet-50		Predicted	
		0	1
Actual	0	1,520	31
	1	9	105

Inception-V3		Predicted	
		0	1
Actual	0	1,522	29
	1	15	99

labels. As shown in Table 8, all models showed similar performance in the final testing data split (10% of the data) as the observed in the stratified k-fold cross-validation method, reaching, for instance for the VGG-19 model, 0.931, 1.0, and 0.925 for the accuracy, sensitivity, and specificity, respectively. This demonstrates the usability of DL techniques and that the models are not likely to be overfitted as demonstrated in the stratified k-fold process and in the 10% final random data split process. During the process of reviewing this paper, we were able to collect 20 more images. We added those images to the previous dataset and performed a random split experiment. We computed the performance of every model using this new dataset.

Discussion

This research is one of the first to demonstrate the capabilities of the DL models to classify SMFI in patients with RC conditions. The imagenological analysis considered an extensive novel shoulder T2-weighted MRI (30). This retrospective analysis applied various DL models, including the VGG-19, ResNet-50, and Inception-v3 architectures.

All diagnostics metrics demonstrated excellent results, achieving a high binary classification performance in every class of the Goutallier level. Distinctly high accuracy, sensitivity, and specificity among different architectures belonging to neural networks were found, specifically when the diagnosis was based on case A, that is, the negative class (Goutallier 0 or 1) and the positive class (Goutallier 3 or 4).

Traditionally, the scapular Y-view of the MRI, particularly the lateral-most T1 sagittal, is the most reliable indicator of the supraspinatus muscle status and is used for identifying FI (31). However, current standard shoulder protocols include sagittal oblique T2-weighted sequences to evaluate these findings (32). Despite that, recent data support ML methods' crucial function in identifying various structures in medical images (33). For this reason, we proposed evaluating the most extensive collections of T2 MRI sequences.

The approach we described allows a practical solution when the grading system of FI is presented, reducing diagnostic uncertainty. Other experiences using artificial intelligence have been published. We highlight the exciting work Ro et al. (24) carried out. They implemented a novel model using only 250 patients (all of whom were diagnosed with atrophy and fatty infiltration of the supraspinatus muscle) to analyze the occupation ratio using a DL framework. They calculated the amount of FI in the supraspinatus muscle using an automated region-based Otsu thresholding technique. Their method allows segmenting the supraspinatus muscle and fossa, which lets them figure out the occupation ratio without automatically classifying the Goutallier level.

In our case, results demonstrated that artificial intelligence tools, particularly the VGG-19 architecture, can be used to support shoulder MRI diagnosis. Few studies in the musculoskeletal radiology literature have addressed the evaluation of RC muscles using these methods (34). Even though supervised deep learning with CNNs has been highly successful in medical imaging, particularly in MRI (35). However, based on the CNN tool, different studies have determined the need to count with more analysis to detect the supraspinatus muscle's fatty infiltration (22).

TABLE 7 Mean train loss, validation loss, accuracy, sensitivity, specificity, and AU-ROC for every case and model (C.I. 95%).

Case	Model	Train loss	Validation loss	Accuracy	Sensitivity	Specificity	AU-ROC
A	VGG-19	0.225 ± 0.053	0.096 ± 0.010	0.973 ± 0.006	0.947 ± 0.039	0.975 ± 0.006	0.991 ± 0.003
	ResNet-50	0.394 ± 0.099	0.123 ± 0.011	0.976 ± 0.006	0.925 ± 0.053	0.980 ± 0.006	0.992 ± 0.003
	Inception-v3	0.474 ± 0.154	0.102 ± 0.009	0.974 ± 0.007	0.869 ± 0.085	0.981 ± 0.006	0.991 ± 0.004
B	VGG-19	0.345 ± 0.045	0.246 ± 0.014	0.925 ± 0.010	0.847 ± 0.041	0.939 ± 0.011	0.961 ± 0.013
	ResNet-50	0.563 ± 0.184	0.187 ± 0.022	0.936 ± 0.012	0.779 ± 0.057	0.963 ± 0.009	0.948 ± 0.017
	Inception-v3	0.332 ± 0.094	0.214 ± 0.012	0.933 ± 0.010	0.802 ± 0.039	0.956 ± 0.008	0.951 ± 0.013
C	VGG-19	0.453 ± 0.057	0.310 ± 0.016	0.900 ± 0.015	0.750 ± 0.078	0.914 ± 0.014	0.935 ± 0.022
	ResNet-50	0.605 ± 0.037	0.507 ± 0.008	0.896 ± 0.015	0.756 ± 0.079	0.909 ± 0.015	0.913 ± 0.025
	Inception-v3	0.587 ± 0.048	0.372 ± 0.013	0.914 ± 0.011	0.659 ± 0.056	0.939 ± 0.012	0.912 ± 0.019
D	VGG-19	0.299 ± 0.056	0.153 ± 0.018	0.942 ± 0.012	0.925 ± 0.056	0.942 ± 0.013	0.977 ± 0.007
	ResNet-50	0.631 ± 0.040	0.405 ± 0.010	0.928 ± 0.013	0.872 ± 0.066	0.932 ± 0.012	0.964 ± 0.012
	Inception-v3	0.494 ± 0.168	0.150 ± 0.011	0.941 ± 0.011	0.808 ± 0.078	0.950 ± 0.010	0.975 ± 0.007
E	VGG-19	0.519 ± 0.242	0.505 ± 0.138	0.779 ± 0.054	0.706 ± 0.088	0.831 ± 0.061	0.861 ± 0.050
	ResNet-50	0.664 ± 0.016	0.631 ± 0.012	0.700 ± 0.038	0.611 ± 0.102	0.756 ± 0.056	0.785 ± 0.053
	Inception-v3	0.696 ± 0.028	0.665 ± 0.008	0.678 ± 0.057	0.550 ± 0.103	0.766 ± 0.088	0.722 ± 0.072

TABLE 8 Accuracy, sensitivity, and specificity for case A and all DL models using a random training/validation/test data split.

	Accuracy	Sensitivity	Specificity
VGG-19 (Case A)	0.931	1.0	0.925
ResNET50 (Case A)	0.948	0.8	0.962
Inception V3 (Case A)	0.965	0.8	0.981

Also, we identified some limitations. Firstly, our results used a binary classification method, even though the classification proposed by Goutallier presents five types of fatty infiltration. However, the binary performance showed great classification results, with an AUC of 0.991 [95% CI, ± 0.003] for the low to nonfatty infiltration against severe to extreme fatty infiltration (VGG-19 model). Therefore, a Fuch-type classification (12) could be more accessible to learn than a Goutallier-type classification. For this reason, it is necessary to have future studies that use multilabel classification methods. In addition, since the number of samples (images) in the data set was small, a training and validation set were created for the cross-validation process, however. The training and validation process used in this study follows related papers which faced similar data limitations (22, 24). To further assess the model performance, we used a training/validation/test random data split using 70%/20%/10% (train size: 413 validation size: 104, testing size: 58) for training, testing and validation, respectively. This allowed us to further confirm the good model performance in predicting class 0 and 1. In the future, more data is needed to further test the proposed models.

On the other hand, when we included category two (Goutallier type 2), the analysis reduced the capability to classify correctly. However, better performance was achieved when the type two class was added to the negative class. As in other publications, the present study was an image analysis; clinical factors and the patient's history were not considered (24). Another essential point is that using these AI tools requires teamwork between clinical practitioners and engineering. Interdisciplinary work is necessary to improve people's health.

In conclusion, CNN models, particularly VGG-19, showed outstanding performance in classifying SMFI using shoulder T2-weighted MRI in patients with RC conditions. AI models could be used to support the radiological diagnosis.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Comité de Ética Científico Adulto del Servicio Metropolitano Oriente de la ciudad de Santiago de Chile (SSMO). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JS: conceptualization, data curation, formal analysis, investigation, methodology, model development, training, and writing original draft. GD: data curation, formal analysis, investigation, validation, and writing original draft. NG: validation and review. CJ: resources and validation. FF: conceptualization, formal analysis, investigation, supervision,

review, and editing. All authors contributed to the article and approved the submitted version.

Funding

The authors received no financial support for the research and authorship. The publication was financially supported by the Universidad Mayor.

Acknowledgments

The authors were grateful for the kind collaboration and assistance of the Sports Medicine Data Science Center MEDS-PUCV.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Urwin M, Symmons D, Allison T, Brammah T, Busby H, Roxby M, et al. Estimating the burden of musculoskeletal disorders in the community: the comparative prevalence of symptoms at different anatomical sites, and the relation to social deprivation. *Ann Rheum Dis*. (1998) 57:649–55. doi: 10.1136/ard.57.11.649

2. Sambandam SN, Khanna V, Gul A, Mounasamy V. Rotator cuff tears: an evidence based approach. *World J Orthop*. (2015) 6:902–18. doi: 10.5312/wjo.v6.i11.902

3. Parikh N, Martinez DJ, Winer I, Costa L, Dua D, Trueman P. Direct and indirect economic burden associated with rotator cuff tears and repairs in the US. *Curr Med Res Opin*. (2021) 37:1199–211. doi: 10.1080/03007995.2021.1918074

4. Dang A, Davies M. Rotator cuff disease: treatment options and considerations. *Sports Med Arthrosc Rev*. (2018) 26:129–33. doi: 10.1097/JSA.0000000000000207

5. Yamamoto A, Takagishi K, Osawa T, Yanagawa T, Nakajima D, Shitara H, et al. Prevalence and risk factors of a rotator cuff tear in the general population. *J Shoulder Elb Surg*. (2010) 19:116–20. doi: 10.1016/j.jse.2009.04.006

6. Dong X, Wang L. The imaging diagnosis of patients with shoulder pain caused by sports injury. *Appl Bionics Biomech*. (2022) 2022:1–12. doi: 10.1155/2022/5272446

7. Chang RF, Lee CC, Lo CM. Quantitative diagnosis of rotator cuff tears based on sonographic pattern recognition. *PLoS One*. (2019) 14:e0212741. doi: 10.1371/journal.pone.0212741

8. Liu F, Dong J, Shen WJ, Kang Q, Zhou D, Xiong F. Detecting rotator cuff tears a network Meta-analysis of 144 diagnostic studies. *Orthop J Sports Med*. (2020) 8:232596711990035–26. doi: 10.1177/2325967119900356

9. Gladstone JN, Bishop JY, Lo IKY, Flatow EL. Fatty infiltration and atrophy of the rotator cuff do not improve after rotator cuff repair and correlate with poor functional outcome. *Am J Sports Med*. (2007) 35:719–28. doi: 10.1177/0363546506297539

10. Warner JJP, Higgins L, Parsons IM IV, Dowdy P. Diagnosis and treatment of anterosuperior rotator cuff tears. *J Shoulder Elb Surg*. (2001) 10:37–46. doi: 10.1067/mse.2001.112022

11. Goutallier D, Postel J, Bernageau J, Lavau L, Voisin MC. Fatty muscle degeneration in cuff ruptures. Pre-and postoperative evaluation by CT scan – pub med. *Clin Orthop Relat Res*. (1994) 304:78–83. doi: 10.1097/00003086-199407000-00014

12. Fuchs B, Weishaupt D, Zanetti M, Hodler J, Gerber C. Fatty degeneration of the muscles of the rotator cuff: assessment by computed tomography versus magnetic resonance imaging. *J Shoulder Elb Surg*. (1999) 8:599–605. doi: 10.1016/S1058-2746(99)90097-6

13. Somerson JS, Hsu JE, Gorbaty JD, Gee AO. Classifications in brief: Goutallier classification of fatty infiltration of the rotator cuff musculature. *Clin Orthop Relat Res*. (2016) 474:1328–32. doi: 10.1007/s11999-015-4630-1

14. Naimark M, Trinh T, Robbins C, Rodoni B, Carpenter J, Bedi A, et al. Effect of muscle quality on operative and nonoperative treatment of rotator cuff tears. *Orthop J Sports Med*. (2019) 7:232596711986301. doi: 10.1177/2325967119863010

15. Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial intelligence in musculoskeletal imaging: current status and future directions. *Am J Roentgenol*. (2019) 213:506–13. doi: 10.2214/AJR.19.21117

16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

17. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using image net. *Comput Biol Med*. (2021) 128:104115. doi: 10.1016/j.compbiomed.2020.104115

18. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *arXiv*. (2015). doi: 10.48550/arXiv.1512.00567

19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv*. (2015). doi: 10.48550/arXiv.1512.03385

20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*. (2014). doi: 10.48550/arXiv.1409.1556

21. Medina G, Buckless CG, Thomasson E, Oh LS, Torriani M. Deep learning method for segmentation of rotator cuff muscles on MR images. *Skelet Radiol*. (2021) 50:683–92. doi: 10.1007/s00256-020-03599-2

22. Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. *Comput Methods Prog Biomed*. (2019) 182:105063. doi: 10.1016/j.cmpb.2019.105063

23. Yao J, Chepelev L, Nisha Y, Sathiadoss P, Rybicki FJ, Sheikh AM. Evaluation of a deep learning method for the automated detection of supraspinatus tears on MRI. *Skelet Radiol*. (2022) 51:1765–75. doi: 10.1007/s00256-022-04008-6

24. Ro K, Kim JY, Park H, Cho BH, Kim IY, Shim SB, et al. Deep-learning framework and computer assisted fatty infiltration analysis for the supraspinatus muscle in MRI. *Sci Rep*. (2021) 11:1–12. doi: 10.1038/s41598-021-93026-w

25. Otsu N. Threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cyber*. (1979) 9:62–6. doi: 10.1109/TSMC.1979.4310076

26. Taghizadeh E, Truffer O, Becce F, Eminian S, Gidoin S, Terrier A, et al. Deep learning for the rapid automatic quantification and characterization of rotator cuff muscle degeneration from shoulder CT datasets. *Eur Radiol*. (2021) 31:181–90. doi: 10.1007/s00330-020-07070-7

27. Yushkevich PA, Piven J, Cody Hazlett H, Gimpel Smith R, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*. (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015

28. Chen XH, Xu D, Ji W, Li S, Yang M, Hu B, et al. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. *Geoderma*. (2021) 400:115159. doi: 10.1016/j.geoderma.2021.115159

29. Martens HA, Dardenne P. Validation and verification of regression in small data sets. *Chemom Intell Lab Syst*. (1998) 44:99–121. doi: 10.1016/S0169-7439(98)00167-1

30. Matsuki K, Watanabe A, Ochiai S, Kenmoku T, Ochiai N, Obata T, et al. Quantitative evaluation of fatty degeneration of the supraspinatus and infraspinatus muscles using T2 mapping. *J Shoulder Elb Surg*. (2014) 23:636–41. doi: 10.1016/j.jse.2014.01.019

31. Lee YB, Yang CJ, Li CZ, Zhuan Z, Kwon SC, Noh KC. Can a single sagittal magnetic resonance imaging slice represent whole fatty infiltration in chronic rotator cuff tears at the supraspinatus? *Clin Orthop Surg*. (2018) 10:55–63. doi: 10.4055/cios.2018.10.1.55

32. Ashir A, Lombardi A, Jerban S, Ma Y, Du J, Chang EY. Magnetic resonance imaging of the shoulder. *Pol J Radiol*. (2020) 85:e420:420–39. doi: 10.5114/pjr.2020.98394

33. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Digital Med*. (2021) 4:1–23. doi: 10.1038/s41746-021-00438-z

34. Khoury V, Cardinal É, Brassard P. Atrophy and fatty infiltration of the supraspinatus muscle: sonography versus MRI. *Am J Roentgenol*. (2012) 190:1105–11. doi: 10.2214/AJR.07.2835

35. Johnson PM, Recht MP, Knoll F. Improving the speed of MRI with artificial intelligence HHS public access. *Seminars Musculoskelet Radiol*. (2020) 24:012–20. doi: 10.1055/s-0039-3400265

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact