frontiers | Research Topics

Machine learning in clinical decision-making

Edited by
Tyler John Loftus, Amanda Christine Filiberto, Ira L. Leeds and Daniel Donoho

Published in
Frontiers in Digital Health

ISSN 1664-8714
ISBN 978-2-8325-3325-3
DOI 10.3389/978-2-8325-3325-3

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

Machine learning in clinical decision-making

Topic editors

Tyler John Loftus — University of Florida, United States
Amanda Christine Filiberto — University of Florida, United States
Ira L. Leeds — Yale University, United States
Daniel Donoho — Children's National Hospital, United States

Citation

Loftus, T. J., Filiberto, A. C., Leeds, I. L., Donoho, D., eds. (2023). *Machine learning in clinical decision-making*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3325-3

Table of contents

frontiers in Digital Health

EDITORIAL
published: 18 November 2021
doi: 10.3389/fdgth.2021.784495

Editorial: Machine Learning in Clinical Decision-Making

Amanda C. Filiberto[1], *Ira L. Leeds*[2] *and Tyler J. Loftus*[1]*

[1] *Department of Surgery, University of Florida, Gainesville, FL, United States,* [2] *Department of Surgery, Yale School of Medicine, New Haven, CT, United States*

Keywords: artificial intelligence, machine learning, reinforcement learning, decision analysis, decision-making

Editorial on the Research Topic

Machine Learning in Clinical Decision-Making

The "Machine Learning in Clinical Decision-Making" Research Topic assimilates evidence and perspectives from researchers and thought leaders that are pursuing the safe, effective development, and clinical application of machine learning systems to augment clinical decision-making across a wide array of specialties including cardiology, neurology, audiology, intensive care, and oncology. This editorial summarizes key points from the Research Topic.

Electronic health record (EHR) systems have become widespread amongst health care systems globally. The resulting EHR databases have generated large, heterogeneous datasets that offer new opportunities to design and implement smarter health care systems by minimizing manual data entry, introducing objectivity where hypothetical-deductive reasoning fails, and providing accurate predictions and classifications that tailor care to individual patients' needs. Despite the promising role for artificial intelligence (AI) techniques and technologies to improve patient care, several substantial barriers to clinical adoption remain.

Many ethical dilemmas surround the use of AI in healthcare. Machine learning algorithms trained to optimize a certain endpoint may make medically sound recommendations without reflecting a patient's ultimate goals of care, which often differ from textbook medical outcomes. Additionally, algorithms trained on biased data sets may produce biased outputs, which may be detrimental if the target population demographics and other characteristics are misaligned. There is also a lack of transparency regarding how many AI algorithms arrive at predictions, which has important implications for care delivery. Currently, despite impressive efficacy in retrospective and observational studies, there is limited level I evidence supporting the use of health care AI for decision support, suggesting a need for more high-level evidence, especially randomized trials.

Electrocardiography (ECG) is an efficient, easily accessible, and commonly performed method for screening and diagnosing cardiovascular disease, a major contributor to potentially preventable mortality and morbidity. Among cardiovascular diseases, heart failure is particularly difficult to recognize due to heterogeneity of underlying pathology and clinical manifestations. Grun et al. demonstrate the ability of AI to accurately predict heart failure from standard 12-lead ECGs, highlighting the potential for AI to promote early diagnosis and treatment using routine clinical data.

Electroencephalography (EEG) is used in the diagnosis, monitoring, and prognostication of many neurological ailments including seizure, coma, sleep disorders, brain injury, and behavioral abnormalities. Similar to cardiovascular disease, these neurologic diseases are heterogenous and often present with diagnostic uncertainty. Saba-Sadiya et al. propose a flexible, unsupervised model that applies to novel EEG data for a variety of clinical decision tasks, including coma prognostication and neurodegenerative illness detection. This work represents an important

OPEN ACCESS

Edited and reviewed by:
Dean Ho,
National University of Singapore, Singapore

***Correspondence:**
Tyler J. Loftus
tyler.loftus@surgery.ufl.edu

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

Received: *28 September 2021*
Accepted: *20 October 2021*
Published: *18 November 2021*

Citation:
Filiberto AC, Leeds IL and Loftus TJ (2021) Editorial: Machine Learning in Clinical Decision-Making. Front. Digit. Health 3:784495. doi: 10.3389/fdgth.2021.784495

foundation for future investigations. Epilepsy affects 50 million people worldwide; approximately one third of all cases are refractory to medications. If a discrete cerebral focus is identified, then neurosurgical resection can be curative. Alim-Marvasti et al. provide evidence that machine learning models trained on a combination of chronological clinical seizure manifestations and an imaging feature can enhance epileptogenic lobe localization, which is a necessary step in achieving optimal surgical outcomes for medication-refractory epilepsy.

In audiology, large amounts of patients' data are measured but are distributed primarily over local clinical databases with unique structures, data elements, and variable names, which hinders external validation and multi-center investigations. Saak et al. illustrated the feasibility of automatically predicting common audiological functional parameters from audiological measures using separate lasso regression, elastic net, and random forest algorithms, which had similar, strong predictive performance. The trained models underlie a prototype for a broadly applicable audiology clinical decision-support system that would function well across local clinical databases despite their unique structures, data elements, and variable names.

Intensive care units (ICUs) serve critically ill patients who require near-continuous surveillance or advanced organ support. Medication dosing can be challenging for critically ill patients because they are often affected by gastrointestinal, hepatic, and renal dysfunction, which affect medication absorption, metabolism, and excretion. Several data-driven medication dosing models have been proposed but have limited ability to assess inter-individual differences and compute individualized doses. Eghbali et al. developed a sedation management agent using deep reinforcement learning which was associated with improved ICU blood pressure management compared with clinicians' performance. The framework proposed by the authors holds promise for automating dosing for other medications commonly used in ICUs.

Smartphones, wearables, and other devices providing medically relevant information generated directly by individuals outside the healthcare system are an emerging trend and can augment existing EHR data for model training purposes. Kyriazakos et al. describe how physiological, psychological, social, and environmental biomarkers can be used to train machine learning algorithms to determine the quality of life of cervical cancer patients and identify novel treatments. Shickel et al. explored the benefits of incorporating novel measurements from wrist-worn activity sensors into EHR data and using resulting datasets and temporal deep learning models to predict patients' illness severity. These results demonstrate the power of non-traditional patient data for making predictions and classifications that have the potential to enhance clinical decision-making.

The editors hope you enjoy the "Machine Learning in Clinical Decision-Making" Research Topic and that your clinical and research efforts in this realm are enriched by the evidence and wisdom shared by the contributing authors.

AUTHOR CONTRIBUTIONS

ACF, ILL, and TJL made substantial contributions to the conception and interpretation of data for the work, provided approval for publication of the content, and agreed to be accountable for all aspects of the work. ACF drafted the manuscript. ILL and TJL provided critical revisions. All authors contributed to the article and approved the submitted version.

FUNDING

TJL was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number K23 GM140268.

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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 Digital Health

TYPE General Commentary
PUBLISHED 31 July 2023
DOI 10.3389/fdgth.2023.1214111

OPEN ACCESS

EDITED BY
Max A. Little,
University of Birmingham, United Kingdom

REVIEWED BY
Paraskevi Papadopoulou,
American College of Greece, Greece
Konstantinos Markatos,
Salamina Medical Center, Greece

*CORRESPONDENCE
Tyler Loftus
tyler.loftus@surgery.ufl.edu

RECEIVED 28 April 2023
ACCEPTED 17 July 2023
PUBLISHED 31 July 2023

CITATION
Filiberto AC, Donoho DA, Leeds IL and Loftus TJ (2023) Commentary: Machine learning in clinical decision-making.
Front. Digit. Health 5:1214111.
doi: 10.3389/fdgth.2023.1214111

Commentary: Machine learning in clinical decision-making

Amanda C. Filiberto[1], Daniel A. Donoho[2], Ira L. Leeds[3] and Tyler J. Loftus[1*]

[1]Department of Surgery, University of Florida, Gainesville, FL, United States, [2]School of Medicine and Health Sciences, George Washington University, Washington, DC, United States, [3]School of Medicine, Yale University, New Haven, CT, United States

KEYWORDS

surgery, machine learning, artificial intelligence, decision support, data science

A Commentary on
Editorial: Machine learning in clinical decision-making

By Filiberto AC, Leeds IL, and Loftus TJ. (2023) Front. Digit. Health. 3:784495. doi: 10.3389/fdgth.2021.784495

Given the success of the Research Topic *Machine Learning in Clinical Decision-Making* published in *Frontiers in Digital Health*, we—the editors of the Research Topic—were pleased to expand the Topic by adding manuscripts that highlight dynamic prediction of mortality among critically ill patients, machine learning models for individualized antimicrobial use duration, electronic health record (EHR) tokenization approaches to patient acuity predictions, and a review of specific artificial intelligence (AI) applications, limitations, and requisites in the United States. These are important Research Topic additions because they embody the principles that clinicians make complex decisions under time constraints and uncertainty using hypothetical-deductive reasoning and individual judgement, which vary from clinician to clinician. Time constraints are imposed by acute diseases and high clinical workloads in which uncertainty results from insufficient knowledge, data, and evidence regarding possible diagnoses and treatments. Clinical decision-support systems often require time-consuming manual data acquisition and entry, which limit their ready adoption by physicians working in high acuity environments with critically ill patients. This General Commentary summarizes key points from the work by Patel, Giordiano, Bolton, Shickel, and their colleagues.

Patients in an intensive care unit (ICU) require close monitoring with a plethora of data points collected in an EHR that are updated frequently. Models predicting mortality have traditionally been used for research purposes rather than individual patient risk assessment at the bedside, and do not consider dynamic clinical status of individual patients. These risk scores are often calculated to produce a score at a single timepoint, overlooking subtle yet important updates in patients' physiology. Despite the rapidly expanding use of EHR data for model training purposes in research environments, current monitoring strategies in clinical use remain limited in their ability to accurately represent changes in patient status.

In volume two of this Research Topic, Patel and colleagues introduce a novel study designed to assess the performance of a dynamic method of updating mortality risk every three hours using a criticality index mortality (CI-M) neural network methodology (1). The data were collected from 2018 to 2020 at the Children's National Hospital, comprising 72 pediatric ICU beds. EHR data were extracted and ICU courses were stratified into three-hour intervals, using a neural network to predict outcomes. The CI-M uses a neural network which incorporates physiology, therapy, and intensity of care to compute a morality risk for pediatric ICU patients in a clinically relevant model using updated data every three hours. The area under

the receiver operating characteristic curves had a minimum value of 0.778 (95% confidence interval 0.689–0.867) at hour three and a maximum value of 0.885 (0.841,0.862) at hour 81. The ten most important variables for risk prediction were duration of ICU stay, ventilator-free days, hours on mechanical ventilation, coma scores, age, and neutrophil counts. The CI-M has the potential to enhance prognostic assessments of critically ill pediatric patients, toward improving clinical decision-making and care. Ideally, this risk model will be externally validated and applicable to other institutions.

Bacterial antimicrobial resistance is a global threat and is associated with increased risk of mortality not only for index patients who develop resistant infections, but also for other patients who suffer collateral harm from spread of resistant organisms, often through healthcare worker vectors. Clinical decision support systems have the potential to increase antimicrobial stewardship, thus mitigating antimicrobial resistance. Bolton et al. use a machine learning and synthetic control-based approach to estimate patients' length of stay (LOS) and mortality outcomes for any given day if they were to stop vs. continue antibiotic treatment (2). Comparisons between decision support system use and control experiences demonstrated minimal difference for both stopping and continuing scenarios, indicating that decision support estimations were reliable (average LOS differences of 0.24 and 0.42 days, respectively). Their approach is novel, can assist with individualized antibiotic cessation, and establishes the safety of patient-specific shortening of antibiotic treatment durations.

Shickel et al. describe their use of a transformer-based patient acuity prediction framework in the critical care setting with a data embedding scheme that captures both concept and corresponding measurement values of many disjoint clinical descriptors (3). The authors introduce a mechanism for combining both absolute and relative temporality as an improvement over traditional positional encoding. They highlight the future of this promising approach while noting that more research is needed to emphasize analyzing self-attention distributions between input variables and clinical outcomes to further the clinical understanding and enhance the trust of clinicians using transformers in healthcare settings.

In a comprehensive review of peer-reviewed literature describing access to AI for clinical decision making, Giordano et al. highlight the use of machine learning models for risk stratification, early warning of acute decompensation, potential bias in machine learning algorithms, and the paradigm shift in medical training towards emerging biomedical informatics applications (4). With the widespread adoption of EHRs there are vast repositories of data sets that are ideal for AI training and testing, and many healthcare disciplines have developed and validated promising solutions for improved risk stratification and optimization of patient outcomes. Healthcare workers will be expected to comfortably work within this new AI frontier and, in turn, relate it to their patients. This review provides an optimal overview and introduction to the novel methods that should be considered.

We hope that you have enjoyed and learned from these important additions to the *Machine Learning in Clinical Decision-Making* and *Machine Learning in Clinical Decision-Making—Volume II* Research Topics published in *Frontiers in Digital Health*.

Author contributions

AF, DD, IL, and TL made substantial contributions to the conception and interpretation of data for the work, provided approval for publication of the content, and agreed to be accountable for all aspects of the work. AF drafted the manuscript. DD, IL, and TL provided critical revisions. All authors contributed to the article and approved the submitted version.

Funding

TL was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers K23 GM140268 and R01 GM149657. TL was also supported by the Thomas Maren Junior Investigator Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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. Patel Anita K, Trujillo-Rivera E, Morizono H, Pollack MM. The criticality Index-mortality: a dynamic machine learning prediction algorithm for mortality prediction in children cared for in an ICU. *Front Pediatr*. (2022) 10:1023539. doi: 10.3389/fped.2022.1023539

2. Bolton WJ, Rawson TM, Hernandez B, Wilson R, Antcliffe D, Georgiou P, et al. Machine learning and synthetic outcome estimation for individualised antimicrobial cessation. *Front Digit Health*. (2022) 4:997219. doi: 10.3389/fdgth.2022.997219

3. Shickel B, Silva B, Ozrazgat-Baslanti T, Ren Y, Khezeli K, Guan Z, et al. Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks. *Front Digit Health*. (2022) 4:1029191. doi: 10.3389/fdgth.2022.1029191

4. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. *Front Digit Health*. (2021) 3:645232. doi: 10.3389/fdgth.2021.645232

frontiers in Digital Health

ORIGINAL RESEARCH
published: 15 December 2020
doi: 10.3389/fdgth.2020.596433

Predicting Common Audiological Functional Parameters (CAFPAs) as Interpretable Intermediate Representation in a Clinical Decision-Support System for Audiology

Samira K. Saak[1,2*], *Andrea Hildebrandt*[1,2], *Birger Kollmeier*[2,3,4,5] *and Mareike Buhl*[2,3*]

[1] Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, [2] Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, [3] Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, [4] HörTech gGmbH, Oldenburg, Germany, [5] Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology (IDMT), Oldenburg, Germany

The application of machine learning for the development of clinical decision-support systems in audiology provides the potential to improve the objectivity and precision of clinical experts' diagnostic decisions. However, for successful clinical application, such a tool needs to be accurate, as well as accepted and trusted by physicians. In the field of audiology, large amounts of patients' data are being measured, but these are distributed over local clinical databases and are heterogeneous with respect to the applied assessment tools. For the purpose of integrating across different databases, the Common Audiological Functional Parameters (CAFPAs) were recently established as abstract representations of the contained audiological information describing relevant functional aspects of the human auditory system. As an intermediate layer in a clinical decision-support system for audiology, the CAFPAs aim at maintaining interpretability to the potential users. Thus far, the CAFPAs were derived by experts from audiological measures. For designing a clinical decision-support system, in a next step the CAFPAs need to be automatically derived from available data of individual patients. Therefore, the present study aims at predicting the expert generated CAFPA labels using three different machine learning models, namely the lasso regression, elastic nets, and random forests. Furthermore, the importance of different audiological measures for the prediction of specific CAFPAs is examined and interpreted. The trained models are then used to predict CAFPAs for unlabeled data not seen by experts. Prediction of unlabeled cases is evaluated by means of model-based clustering methods. Results indicate an adequate prediction of the ten distinct CAFPAs. All models perform comparably and turn out to be suitable choices for the prediction of CAFPAs. They also generalize well to unlabeled data. Additionally, the extracted relevant features are plausible for the respective CAFPAs, facilitating interpretability of the predictions. Based on the trained models, a prototype of a clinical decision-support system in audiology can be implemented and extended towards clinical databases in the future.

Keywords: CAFPAs, clinical decision-support systems, machine learning, audiology, interpretable machine learning, precision diagnostics

OPEN ACCESS

Edited by:
Amanda Christine Filiberto, University of Florida, United States

Reviewed by:
Alex Jung, Aalto University, Finland
Meenakshi Chatterjee, Johnson & Johnson, United States

***Correspondence:**
Samira K. Saak
samira.kristina.saak@uni-oldenburg.de
Mareike Buhl
mareike.buhl@uni-oldenburg.de

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

Received: *19 August 2020*
Accepted: *26 November 2020*
Published: *15 December 2020*

Citation:
Saak SK, Hildebrandt A, Kollmeier B and Buhl M (2020) Predicting Common Audiological Functional Parameters (CAFPAs) as Interpretable Intermediate Representation in a Clinical Decision-Support System for Audiology. Front. Digit. Health 2:596433. doi: 10.3389/fdgth.2020.596433

INTRODUCTION

Clinical decision-making is a complex and multi-dimensional process which comprises gathering, interpreting, and evaluating data in the context of a clinical case, in order to derive an evidence-based action (1). Due to the complexity of the process, clinical decision-making is obviously prone to errors. Their rates in general practice have been estimated as high as 15% (2). Arguably, wrong clinical decisions can have considerable negative impact on the quality of life of the affected individuals (3). This is also true for decision-making in audiology. Considering that [1] about 5% of the world population and one third of individuals aged above 65 years suffer from disabling hearing loss (4), [2] that the age group above 65 years is the fastest growing population (5), and [3] that decisions are prone to error also in audiology, it is important to continuously improve the precision of clinical decision-making in this domain.

Flaws in clinical decision-making are partly caused by individual differences between physicians with respect to their level of expertise, the subjective nature of the decision-making process, as well as environmental factors. For instance, highly experienced physicians tend to be more accurate in their choice of treatment as compared to novices (6). Furthermore, also experts, similarly to novice physicians, like humans in general, are susceptible to cognitive processing biases. Most often occurring distortions were described as the availability bias, confirmation bias, and premature closure, amongst others (7). Lastly, different physicians may have access to different measurements (data) because different clinics may use different test batteries in their assessment kits which can vary with respect to their measurement precision and validity (8). Additionally, it is possible that in the longitudinal evaluation of a patient, required data from previous potential examinations is missing, or inconsistencies in the administered tests entail difficulties for a physician newly involved in the case (8). In summary, the aforementioned factors arguably lead to variability in the clinical decision-making process across physicians and clinics, and facilitate distortions in diagnostic outcomes. To improve the objectivity, precision and reproducibility of physicians' decision-making, clinical decision-support systems (CDSS) have received an increased attention in many health care domains.

CDSS are information systems that aim to improve clinical decision-making by providing relevant information on relationships between measurements and diagnosis to physicians, patients, or other individuals involved in the clinical context (9). They aim to reduce the information load of physicians by summarizing it through the extraction of patterns and predictions from large amounts of data (10). For instance, physicians can be informed with probabilities of certain medical findings and treatment recommendations, based on imputed case-relevant data which can help to achieve well-informed judgements (9). In addition, CDSS can rule out subjectivity in clinical judgements. Not only can they reduce the impact of processing biases on diagnostic outcomes, but also support novice physicians in their decision-making process to eliminate inter-physician variability in diagnostic outcomes.

The advantage of CDSS has been demonstrated in many previous studies. Just to exemplify with a few, Paul et al. (11) introduced a computerized CDSS for antibiotic treatment. Based on a sample of 2,326 patients in three different countries, the study demonstrated that TREAT improved the hits for an appropriate antibiotic treatment to 70% as compared with physicians who only achieved 57% hits. Another example for a successful CDSS was provided by Dong et al. (12). The authors developed a rule-based CDSS for the classification of headache disorders which correctly identified several types of conditions with an accuracy above 87.2%.

Despite the demonstrated potential of using CDSS, in practice a widespread usage is oftentimes lacking. Developed CDSS may not go beyond the trial stage and physicians may choose not to adopt them (13). Consequently, research has tried to identify potential reasons that lead physicians to refrain from using a CDSS. The Technology Acceptance Model developed by Davis (14) aims to explain this problem of users acceptance with respect to Information Technology in general. It concludes that user's acceptance is influenced by design features, perceived usefulness, and perceived ease of use. The perceived ease of use represents how effortless a system can be adopted and it will causally affect the perceived usefulness. This, in turn, entails how such a system would benefit the user and enhance his or her performance. However, it is believed that physicians may be more prone to assess a system based on trust, rather than its usefulness or ease of use (15). Wendt et al. (16) state that the extent to which users are convinced of the validity of the information provided by the CDSS is crucial for acceptance. On the one hand, this can be achieved by including physicians in the development of such CDSS, by means of interviewing physicians along with extensive piloting. This could lead towards a CDSS that addresses the physicians needs and, additionally, incorporate it in such a way that it fits into the physician's workflow. On the other hand, enabling physicians to understand how the CDSS works may further increase their trust towards them. As a result, physicians evaluate and interpret the system's output and determine its validity, enhancing the level of comfort in utilizing the CDSS (17). Consequently, black box CDSS are rarely accepted, so that understandable algorithms need to be established for achieving physician's trust.

In the medical discipline of audiology, in addition to the aforementioned issues, the heterogeneity of the applied assessment tools among different clinics leads to further challenges in clinical decision-making (8). As a result, comparability in audiological diagnostics and treatment recommendations across clinics is compromised. This in turn may lead to some of the errors that occur in provided diagnostic decisions. Moreover, the differences in applied audiological measures may turn out to pose challenges for the development of a CDSS, aiming to enhance diagnostic precision. This is because data from different measurement sources need to be accounted for and integrated in a CDSS. Thus far, the use of machine learning and CDSSs in the field of audiology is restricted to automatizing audiological measures (18, 19), predicting specific diseases, e.g. vertiginous disorders (20), or for a broad classification of individuals into auditory profiles (21).

For instance, Song et al. (18) proposed an automated audiometry based on machine learning that resulted in similar estimates at audiogram frequencies, while requiring fewer samples than the traditional manual procedure. Further, Sanchez Lopez et al. (21) identified four different auditory profiles using unsupervised learning, which differ on the dimension of audibility and non-audibility related distortions and may be used for the development of audiological test batteries. However, to the best of our knowledge, no CDSS was yet proposed aiming to support physicians in their general diagnostic endeavor for a variety of audiological findings.

To address this issue and to work out the relevant constituents of a more generally applicable CDSS in the field of audiology that are transparent to the physicians with respect to their underlying properties, Buhl et al. (8) developed the Common Audiological Functional Parameters (CAFPAs). The CAFPAs aim to represent the functional aspects of the human auditory system in an abstract and measurement-independent way. They can act as an interpretable intermediate representation in a CDSS, i.e. CAFPAs are estimated from audiological measures, and the CAFPAs can be used to infer probabilities of audiological findings or treatment recommendations. In other words, the CAFPAs aim to integrate audiological data from a variety of sources, next to allowing physicians to interpret and validate them. This is achieved through ten different parameters, describing relevant conditions which help to determine hearing disorders (8).

Due to their characteristic of being an abstract representation that does not depend on specific audiological measures, the CAFPAs provide a common framework for physicians, regardless of environmental factors, i.e. differences in audiological measures and clinical expertise. In addition, the CAFPAs were defined in an expert-driven way, through discussions among experts (8) and by considering the statistical analysis performed by Gieseler et al. (22). By including audiological experts into the development process of the CAFPAs, the crucial aspect of users involvement, here physicians, has been addressed. In summary, the need for a CDSS with decision-making steps that become transparent to physicians is addressed by the CAFPA framework aiming to act as interpretable intermediate layer in a CDSS. This property ensures that a future CDSS based upon the CAFPAs will not be a black box.

Buhl et al. (8) already demonstrated the general feasibility of the CAFPAs to be used as abstract representation of audiological knowledge. By an expert survey conducted in the opposite direction as compared with the typical diagnostic process, audiological experts rated outcomes of audiological measures and CAFPAs for given diagnostic cases (i.e., audiological findings as well as treatment recommendations). This resulted in audiologically plausible distributions. As a next step towards a CDSS for audiology, Buhl et al. (23) built a labeled data set in the typical direction of audiological diagnostics, i.e. experts rated audiological findings, treatment recommendations, and CAFPAs based on individual patients' data from audiological measures. The suitability of the given data set as a training distribution for future algorithmic audiological classification tasks was assessed and confirmed. Hence, Buhl et al. (23) provided a data set with expert-derived CAFPAs for given audiological measure data in a sample of individual patients. Based on this data set, machine learning models for the automatic estimation of CAFPAs from audiological measures can now be built and evaluated as a next step towards a CDSS in audiology.

The current study therefore aims at:

1. Predicting expert determined CAFPAs for given audiological measures using machine learning models;
2. Identifying the most relevant features for the prediction of ten different CAFPAs from the audiological measures, in order to ensure the interpretability of the models and increase physicians' future acceptance of automatically derived CAFPAs;
3. Evaluating the potential of the trained models in predicting CAFPAs for unlabeled data i.e., unlabeled patient cases from available databases.

METHOD

Data Set

As outlined above, Common Audiological Functional Parameters (CAFPAs) are intended as intermediate representations between audiological measures and diagnostic decisions in a CDSS. To empirically instantiate CAFPAs, Buhl et al. (23) conducted an expert survey on a data set containing audiological measures ($N_{\text{total}} = 595$) provided by the Hörzentrum Oldenburg GmbH (Germany). Thus, given the audiological data, experts were asked to assess CAFPAs, as well as to provide diagnostic decisions for $N_{\text{labeled}} = 240$ patients. The remaining data of $N_{\text{unlabeled}} = 355$ patients will be used as unlabeled cases for further evaluations of the trained algorithms. With the labeled data set we intend to quantify the link from audiological measures to CAFPAs.

Common Audiological Functional Parameters

The CAFPAs describe functional aspects of the human auditory system and are thereby independent of the choice of audiological measures. The covered functional aspects are summarized in **Table 1** and **Figure 1A**.

In a CDSS for audiology, the CAFPAs are planned to act as an interpretable intermediate layer. They should be determined from audiological measures. Subsequently, a classification of audiological findings, diagnoses, or treatment recommendations for the provision with hearing devices could be performed based on their basis. The CAFPAs are defined on a continuous scale in the interval [0 1], indicating the degree of impairment. Their scale can be graphically displayed in a traffic-light-like color scheme (cf. **Figure 1B**), where for the respective functional aspect green [0] represents "normal" and red [1] represents "maximally pathological" status.

Expert Survey

The database of the Hörzentrum Oldenburg GmbH (Germany) contains audiological measures, cognitive tests, and self-reports on multiple questionnaires from more than 2,400 patients. Complete data on main variables relevant for the expert survey was available for 595 patients. A detailed description of this database was published by Gieseler et al. (22). In the expert survey by Buhl et al. (23), a part of this database was labeled for the

TABLE 1 | Overview and description of CAFPAs.

Functional aspects	CAFPA	Description
Hearing Threshold	C_{A1} C_{A2} C_{A3} C_{A4}	The CAFPAs CA1-CA4 refer to the hearing threshold at increasing frequencies. Hearing threshold refers to the minimum sound level that is required to hear a sound. It is indicated as the threshold at which a sound is detected at least 50% of the time. The hearing thresholds are given in decibels of the hearing level (dB HL) for given frequencies in comparison to the normal population. Values between 0 and 20 dB HL are considered to be within the normal range, whereas increasing dB HL values correspond to increasing hearing loss for the given frequencies (24).
Suprathreshold deficits	C_{U1} C_{U2}	These components refer to deficits at levels above the threshold (24) for lower (C_{U1}) and higher frequencies (C_{U2}). Even if hearing threshold levels are within the normal range, deficits may still be present in the suprathreshold range, e.g. with deficits in speech recognition (25).
Binaural hearing	C_B	Binaural hearing reflects processes taking place in the central nervous system, which enables hearing with two ears simultaneously (24, 26). On the one hand, this entails the ability to perceive different signals that reach the two ears as one, termed binaural fusion (24). On the other hand, binaural hearing allows spatial hearing and sound localization (26, 27).
Neural processing	C_N	This CAFPA broadly defines the involvement of neural components in the hearing process, such as the cochlear and auditory neurons (24).
Cognitive components	C_C	Cognitive components play a role in hearing deficits. Studies have widely indicated a correlation between age-related hearing loss and cognitive decline, even though the causal mechanisms remain unclear (28). Cognitive decline may reduce available cognitive resources for auditory processing. Conversely, reduced auditory input caused by hearing loss may lead to a degradation of inputs to the brain, causing cognitive decline. In any case, a strong association between cognitive measures and hearing loss has been found (29).
Socio-economic status	C_E	This CAFPA contains information regarding the socio-economic status of an individual, which is a combined measure of economic and social status, found to be positively associated with better health (30).

purpose of linking CAFPAs to audiological diagnostics. Thereby, audiological experts were asked to label individual cases from the database. They were asked to indicate expected CAFPA values as well as audiological findings and treatment recommendations on a one-page survey sheet on which the patients' data were displayed in a graphical manner.

The following audiological measures and subjective patients' reports were displayed to the experts. The audiogram (for air and bone conduction), which characterizes the hearing threshold of a patient, i.e. which minimum sound pressure level can be perceived at different frequencies. The adaptive categorical loudness scaling [ACALOS; (31)] which aims to assess the loudness perception of the patient. Furthermore, speech intelligibility was captured with the Goettingen sentence test [GOESA; (32)]. The Vocabulary test [German: Wortschatztest (WST); (33)] was used as a measure of verbal intelligence. Information regarding the socio-economic status was assessed with the Scheuch-Winkler index [SWI; (34)]. The DemTect (35) was selected as a measure of cognitive performance which also serves as a screening measure for dementia. Finally, self-reports on age, gender, first language, the presence of tinnitus in the left/right ear, and hearing problems in quiet and in noise were additionally displayed to the experts.

Experts were asked to indicate expected CAFPA values on a continuous color bar based on their clinical experience in audiology. Furthermore, they had to tick diagnostic cases from a provided list of options. Audiogram and loudness scaling results were available for both ears. If there was an asymmetry between the ears in a given case, experts were instructed to consider only the worse ear for estimating respective CAFPAs and diagnostic classes. According to the above procedure, expert labels were obtained for 240 different patient cases. Out of these, for consistency check, a subset was given to multiple experts. Thus, in total 287 labeled expert survey sheets were available. The mean age of the sample including labeled cases was 67.5 ($SD = 11.3$). For the present analyses, the expert labels provided for the CAFPAs are assumed to reflect the ground clinical truth. They will be denoted as 'labeled' CAFPAs in the following.

Model-Building

CAFPAs, which serve as labels, are defined on a continuous scale, leading to a regression problem to be solved for automatic generation of CAFPA values given the above mentioned audiological data (features) for the patients (data points). The model space of the given regression problem contains the lasso regression, elastic nets, and random forests approaches. These predictors will be applied and evaluated in comparison with regard to the loss function. The model space covers the range between higher interpretability and lower flexibility (lasso regression, elastic net) and lower interpretability and higher flexibility [random forests; see (36)]. The comparative evaluation aims at capturing the well-known trade-off between interpretability and potentially higher predictive performance accuracy, whereby the first is a similarly crucial feature for a CDSS in order to be accepted in applied context.

We use a 10-fold Cross-Validation (CV) in the model-building process. The data set for the prediction of each CAFPA was randomly split into training (80% of the sample, containing the validation set) and test sets (20%). The validation set is used for hyperparameter tuning. In contrast, the test set is not being used in the model-building process, but for evaluating the model with respect to prediction accuracy for future cases.

Features and Labels

Each of the ten CAFPAs was treated as individual label. Features are the audiological measures as used in the expert survey (**Table 2**). If an audiological measure includes several

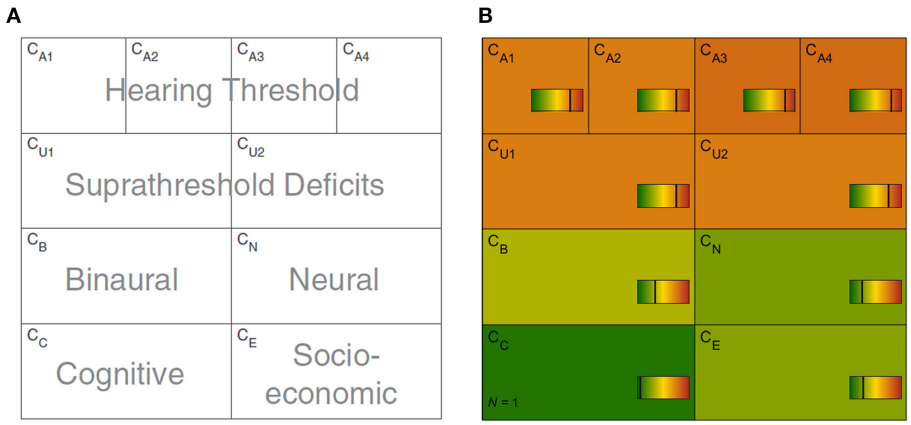


FIGURE 1 | Common Audiological Functional Parameters (CAFPAs). **(A)** Functional aspects of the human auditory system represented by the CAFPAs. **(B)** Exemplary CAFPA representation. The color bar corresponds to the interval [0 1]. The respective value of each CAFPA is indicated by the color of the area, as well as by the vertical line within the color bar.

TABLE 2 | Overview of audiological measures and features.

Measure	Number of Features	Features
Audiogram (air conduction)	11	Frequencies: {0.125, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0, 6.0, 8.0} kHz; worse ear (according to PTA) selected
Audiogram (bone conduction)	7	Frequencies: {0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0} kHz worse ear (according to PTA) selected
Asymmetry score	1	Difference of pure-tone average (PTA) hearing loss for left and right ear in dB
Adaptive categorical loudness scaling (ACALOS)	12	With 1.5 & 4 kHz narrowband noise; worse ear selected – Lcut (juncture point between linear parts of the loudness function) – Mlow (slope of first linear part) – Mhigh (slope of second linear part) – L2.5 (hearing threshold level) – L25 (medium-loudness level) – L50 (uncomfortable level) (37)
Goettingen sentence test (GOESA)	3	SRT (speech reception threshold) Slope SI (speech intelligibility) (32)
Vocabulary test (WST)	1	Sum of correct answers (33)
DemTect	1	Sum score of five tests (08: suspect of dementia; 912: slight cognitive impairment; 1318: normal cognitive behavior) (35)
Hearing problems (HP)	2	quiet; noise 0 (no hearing loss) to 5 (very severe)
Scheuch-Winkler Index (SWI)	1	Sum score for categories profession, education, and income (34)
Age	1	Age in years
Language	1	Native speaker (German); non-native speaker
Gender	1	Male; female
Tinnitus	2	Presence; right and left ear

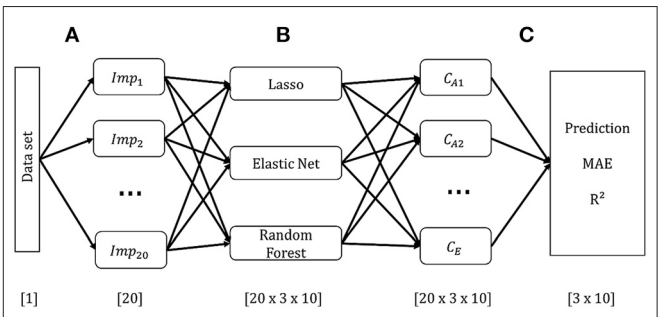


FIGURE 2 | Schematic overview of the model-building pipeline. Numbers in brackets indicate dimensions. **(A)** Pre-processing and generation of 20 data sets based on multiple imputation of missing values. **(B)** Building models on each imputed data set for each CAFPA. **(C)** Prediction of CAFPAs with the three selected models on each imputed data set. Evaluation of prediction accuracy using the Mean Absolute Error (MAE), and R^2 averaged across multiple data sets.

measurement variables (e.g., the audiogram is measured for different frequencies), each of these variables is used as feature. In total, 44 features were used for modeling. Corresponding to the instruction in the expert survey to rate CAFPAs for the worse ear in case of an asymmetric hearing loss, only audiogram and adaptive categorical loudness scaling data for the respective worse ear of each patient are included as features. To retain information regarding the asymmetry between ears, an asymmetry score serves as an additional feature. This score reflects the absolute difference in dB between the pure-tone average hearing loss (PTA; audiogram (air conduction) averaged over the frequencies 0.5, 1, 2, and 4 kHz) of the left and right ear [e.g., (38)]. **Figure 2** depicts the general analysis pipeline for predicting the CAFPAs.

Pre-processing

To avoid statistical dependency due to multiple evaluations of certain patients by multiple experts, for all analyses we randomly selected the CAFPA results of one experts' response only. For all features, but for hearing problems in quiet and noise (74.3%), at least 94.2% of the data were available. Where necessary, we imputed missing data on features by using Multivariate

Imputation with Chained Equations [MICE; (39)]. MICE is an approach in which missing values on one feature are estimated based on the remaining features included into the imputation model. Missing values are replaced by predicted values with an added random error term. To minimize potential bias due to one single addition of the random error, the imputation process is repeated multiple times. Imputed values are updated in each iteration, resulting in a given imputed data set. By generating multiple such imputed data sets, MICE accounts for the uncertainty that stems from predicting missing values (39). It is a superior missing data technique as compared with single imputation methods, such as mean or predicted values imputation (40). We used 20 iterations for each imputed data set and generated a total amount of 20 imputed data sets. This amount was shown to be sufficient for successful estimation of the missing data (39, 41). The plausibility of the imputed values was visually inspected across iterations and imputed data sets, as well as through a density plot of the imputed values for each feature. Modeling was carried out on each of the 20 imputed data sets, instead of averaging the data prior to the model-building process (41). Thus, we averaged the predicted CAFPAs after being estimated over multiple data sets.

Missing labels were not imputed. For the prediction of each CAFPA label only those cases were included for which the corresponding CAFPA label was available. In total, 97.5% of the labeled CAFPAs were available. Thus, for each predicted CAFPA, only minor sample size differences occurred.

Lasso Regression and Elastic Net

Lasso regression and elastic net are both linear regression models that are closely related to each other. As with linear regression, coefficients are estimated, such that the Residual Sum of Squares (RSS) is minimized. Both lasso regression and elastic net perform feature selection by introducing a penalty for the size of the coefficients (36). By feature selection, a more parsimonious model is being achieved, so that model flexibility and interpretability is optimized. Lasso regression and elastic nets use different penalties. Whereaslasso regression introduces the l_1 penalty (Equation 1), elastic nets combine the l_1 with the l_2 penalty (Equation 2).

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (1)$$

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (2)$$

With l_1, the model will penalize the sum of the absolute values of the regression coefficients depending on the tuning parameter λ and thus, sparse models result because coefficients can be shrunken exactly to zero. The size of the selected λ determines the strength of the penalty, with larger values of λ corresponding to a stronger regularization (36). The tuning parameter is being selected by cross-validation in the model-building process (see below).

In contrast, the l_2 penalty does not eliminate coefficients, but shrinks irrelevant features towards zero, next to grouping correlated features together by assigning them similar coefficient sizes (36). Combining both penalties, as in elastic nets, will have three consequences: Irrelevant features will be eliminated, less important features will be shrunken towards zero and correlated features will be grouped together. The relative contribution of each penalty can be fine-tuned with α, a tuning parameter ranging on a scale from [0 1]. As part of the model building process features were standardized for both lasso regression and elastic net, to ensure an equal impact on all coefficients.

For lasso regression, we evaluated λ values that cover the range between the least squares estimate (simple linear regression including all features, $\lambda = 0$) to the null model (including no feature and using the mean of the labels as predicted value, $\lambda \rightarrow$ inf). The λ value minimizing the loss function of the validation set was selected by means of 10-fold CV separately for each imputed data set.

For elastic net, we performed a grid search of the length 10 for α and λ, using the caret train() function in R. That is, we considered a combination of ten potential values for both α and λ in the grid. Values for α and λ minimizing the loss function on the validation set were selected with 10-fold CV for each imputed data set (cf. **Figure 4**).

Random Forests

Random forests combine multiple decision trees for improving the accuracy and robustness of predictions as compared to those achieved by a single decision tree. Decision trees perform recursive binary splitting of the feature space, that is, a feature that leads to the largest reduction of the RSS is being selected for a split, such that two distinct regions are obtained at every step of the tree building process. In every step, the splitting procedure is repeated based on other features, such that multiple regions in the multivariate space of the observed data are obtained. The prediction is different for each determined region and it corresponds to the mean of the observed response variable in the respective regions. For random forests, multiple trees are built. To avoid building the same decision tree multiple times, only a specified number of features was considered at each split. This enforces different structures of the achieved decision trees and it has the effect of de-correlating the trees before being averaged for the final prediction. As such, the variance of the prediction for future cases (test data) is being minimized (36).

For the current analyses, we tuned the number of features considered at each split (*mtry*) using the tuneRF() function from the randomForest package in R (42). TuneRF() searches for optimal values for *mtry* given the data. The final number of features selected at each split was then determined using the proposed *mtry* values for the 500 trees built for each fold of the 10-fold CV.

Model Evaluation Based on Labeled Cases

Prediction of the CAFPAs

We evaluated the models' performance using the Mean Absolute Error (MAE) as loss function and the coefficient of determination (R^2) between labeled and predicted CAFPA values. As mentioned above, for each of the ten CAFPAs, 20 imputed data sets exist. Accordingly, we built all models (lasso, elastic net and random

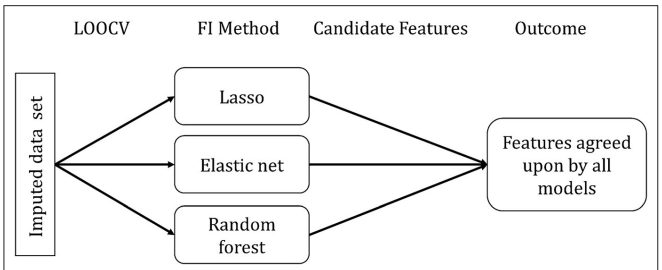


FIGURE 3 | Feature importance (FI) analysis pipeline. FI Method indicates the specific method for each model to extract the relevant features.

forest) multiple times on each imputed data set. This resulted in 20 × 3 models for each CAFPA [20 × 3 × 10]. For final model evaluation, we then averaged the MAE and R^2 values across multiple estimations for each CAFPA [3 × 10]. In addition, the correlation between the labeled and predicted values were estimated and plotted. Density plots for labeled and predicted values are provided as well. The null model was chosen as a general baseline to improve upon.

Feature Importance

For assessing feature importance, we randomly selected one of the 20 imputed data sets, as we did not expect significant differences between the data sets. Furthermore, we did not observe differences when inspecting the standard deviation of the predicted CAFPAs across multiple imputed data set. The selected data set was used to build all three models using Leave-One-Out-Cross-Validation (LOOCV) for each CAFPA [1 × 3 × 10]. LOOCV performs CV by leaving out one observation to be considered as validation set. No additional test set was set aside (differently from the prediction of the CAFPAs), considering that no predictions on future data are made. **Figure 3** depicts the feature importance analysis pipeline.

Feature importance assessment is identical for lasso regression and elastic net and it directly follows from the definition of the methods. Due to the different approaches of feature selection that characterize the specific models, selected features differ across models. We used the selection frequency of each feature across all LOOCV models to determine feature importance. Features selected for more than 50% of the LOOCV models are candidate features to be considered relevant.

For each random forest model, we calculated a feature importance measure. For each tree ($n = 500$) in the random forest, 2/3 of the data was used for resampling with replacement. The remaining 1/3 of the data is termed out-of-bag (OOB). Predictions for each data point i were made by averaging all trees in which i was part of the OOB sample. The loss function can be calculated from the resulting predictions (36). Subsequently, the importance of a given feature p was determined by calculating the loss function for each tree in the forest, including all features, next to calculating them with a permuted feature p' (36). The average difference between the two loss functions was then normalized and scaled to range from 0 to 100, with 100 being the most important (43). Here, all features with importance values above 50 were considered candidate features. Features selected as candidates by all three models were taken as most relevant features for the prediction of a respective CAFPA.

Model Application to Unlabeled Cases

Our aim was to obtain a model that allows predicting CAFPAs in the context of a CDSS. Thus, it is crucial that the obtained model(s) are accurate at estimating CAFPAs on unlabeled cases. Therefore, the models were applied to the additional 355 cases ($mean\ age = 67.6$, $SD = 12.3$) of our data set ($N_{total} = 595$) for which no expert labels on CAFPAs are available. To evaluate the predictions on unlabeled cases, we applied model-based clustering (section Prediction of CAFPAs and Clustering for the Unlabeled Data Set). Ideally, we should find the same number of clusters in the CAFPAs predicted by the models from the unlabeled data set, as on the labeled data set.

Pre-processing

For the purpose of imputing missings in the unlabeled data set using MICE, we merged this data set with the labeled, previously imputed data set. Because in the future CAFPAs should be predicted for individual cases as part of a CDSS for audiology, potential missing data in single patients will have to be imputed on the basis of larger databases. Thus, merging the unlabeled data set with the labeled one to deal with missingness is in line with procedures suitable for a prospective CDSS. Apart from merging the data sets, the imputation procedure for the features was identical to the one described before. After imputation, we separated the two data sets. In contrast to the model-building analysis, for clustering purposes missing data on CAFPAs were also imputed. However, the imputation was performed exclusively on the basis of the available labeled CAFPAs without considering the features in the imputation model.

To obtain a comparable data set to the labeled one with respect to its size as well as demographic characteristics of the cases (i.e., age, gender, and first language), we applied propensity score matching [PSM; (44)]. The propensity score is defined as the conditional probability that a data point belongs to a treatment group (e.g., in our case to the labeled vs. unlabeled sample) given a set of covariates. It can be estimated by logistic regression (45). Data points with a similar propensity score in the labeled vs. unlabeled data are matched according to the Nearest Neighbor (NN) matching technique (46). NN refers to matching each propensity score from the treatment group (unlabeled data) with the nearest propensity from the control group (labeled data). As a result of the PSM, the unlabeled data set used for unsupervised prediction of the CAFPAs and for subsequent evaluation with model-based clustering consists of 240 cases ($mean\ age = 67.4$, $SD = 11.8$) that are maximally similar to the labeled cases with respect to demographic features.

Prediction of CAFPAs and Clustering for the Unlabeled Data Set

We predicted CAFPAs for the unlabeled cases using the three previously trained models (lasso, elastic net, random forests), each containing 20 models, resulting from the 20 imputed data sets in the model-building part of the present analysis.

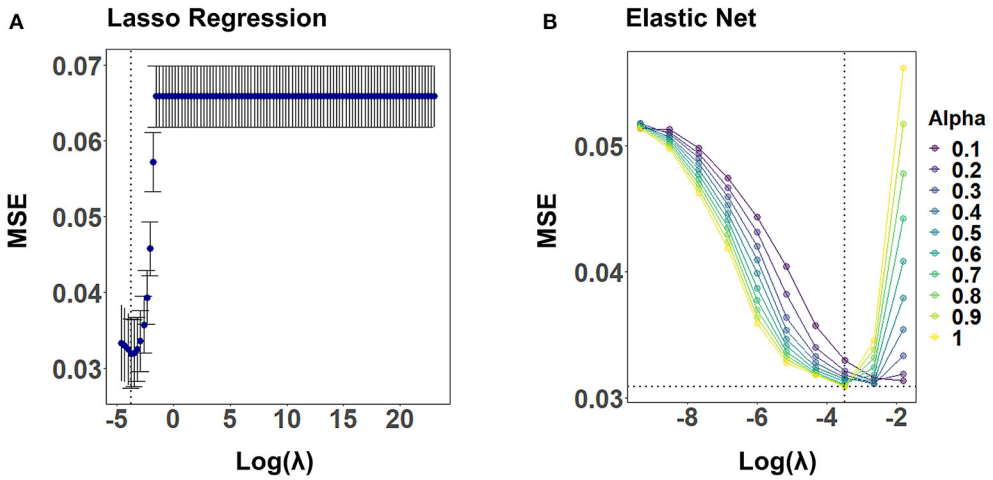


FIGURE 4 | Hyperparameter selection for lasso regression and elastic net using 10-fold CV, exemplarily for C_{A2} and a randomly selected imputed data set. Both plots display the mean-squared error (MSE) as a function of log(λ). The dotted line indicates the values leading to the smallest MSE. **(A)** Tuning of λ for lasso regression. The standard error of λ across CV-folds is shown. **(B)** Tuning of λ for different levels of α with elastic net.

To evaluate the predictions for unlabeled cases, we applied model-based clustering to [1] the labeled CAFPAs and [2] predicted CAFPAs from the data not containing labels. Model-based clustering assumes the data to stem from a mixture of gaussian distributions, where each cluster k is represented by a cluster specific mean vector μ_k and a covariance matrix Σ_k (38). The covariance matrix determines the shape, volume, and orientation of the clusters (e.g., varying or equal shape, volume, and orientation). Thus, to determine the most suitable number of clusters for given data, model-based clustering applies different parameterizations of the covariance matrix for different numbers of components [see (47) for the different parameterizations of the covariance matrix]. Accordingly, multiple clustering models can be compared with regard to their properties (i.e., covariance structure and number of components) and the best fitting model selected for the cluster analysis. Model selection can be performed by means of the Bayesian Information Criterion (BIC), which evaluates the likelihood of the model given the data and parameterization, with larger BIC values indicating better fit of a model (48).

To select the optimal model and number of clusters for the data set including labeled CAFPAs, we inspected the BIC to choose the parameterization of the covariance matrix. Thereafter, we determined the optimal number of clusters via visual inspection of the resulting average CAFPA patterns for each cluster. That is, the largest number of clusters differentiating labeled CAFPA patterns was selected (cf. **Supplementary Figures 6, 7**). As the clusters exist in a multidimensional space, i.e. the ten CAFPA dimensions, we applied principle component analysis (PCA) to visualize the clusters. PCA is a dimensionality reduction method that linearly combines features to result in a new set of orthogonal principle components (PCs). The PCs are ordered with regard to variance, i.e. the first PC explains the largest amount of variance in the data (49). This allows a visualization of clusters in a 2D space (PC1 and PC2), while retaining a large amount of variance existing in the data (50). We then intended to reproduce the same number of clusters of CAFPAs estimated, in the unlabeled data set using the same covariance parameterization, for the purpose of providing comparability between labeled and predicted clusters.

RESULTS

Model Evaluation Based on Labeled Cases

Model-Building

Figure 4 illustrates the CV results from tuning λ for lasso regression, as well as α and λ for the elastic net, exemplarily, for C_{A2} of a randomly selected imputed data set. Values for α and λ were selected that lead to the largest error reduction in the validation set, as indicated by the dotted line. The results for the remaining CAFPAs for the given imputed data set are provided in the **Supplementary Figures 1, 2**. **Figure 5** depicts the MAE of the trained models for the training and test set across CAFPAs, in comparison to the MAE of the null model. The performance of the lasso regression and the elastic net is comparable. The test error for random forest is slightly higher as compared to the training error but not yet indicative of overfitting.

Prediction of CAFPAs

Figure 6 displays the models' performance at predicting the CAFPAs. In case of all three models, the predicted CAFPAs in the test set were averaged over the imputed data sets. **Figure 6A** shows the mean absolute error (MAE) between labeled and predicted CAFPAs for the three models as compared with the null model. Although different models perform best for different CAFPAs as indicated by the color bars, the performance across models is comparable, and all models improve upon the null model. The average reduction of MAE over CAFPAs is also

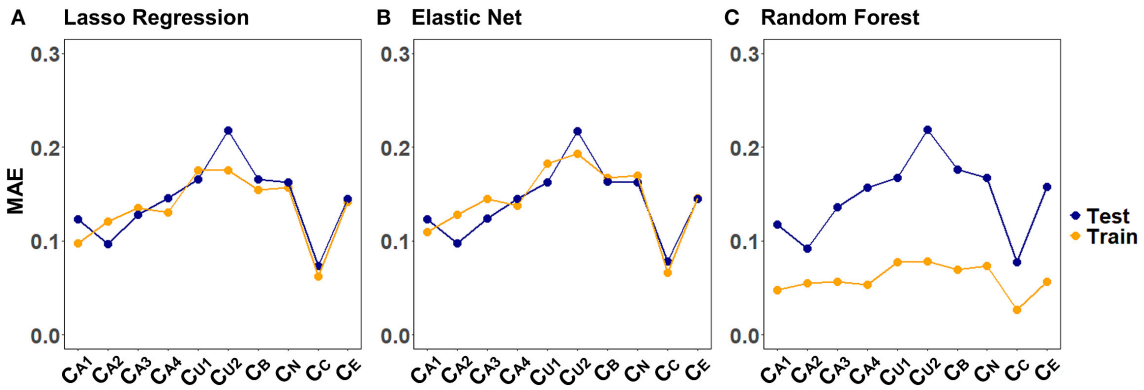


FIGURE 5 | Training and test set loss function (MAE) across CAFPAs for the three models **(A)** lasso regression, **(B)** elastic net, and **(C)** random forest. MAE values correspond to a randomly selected imputed data set.

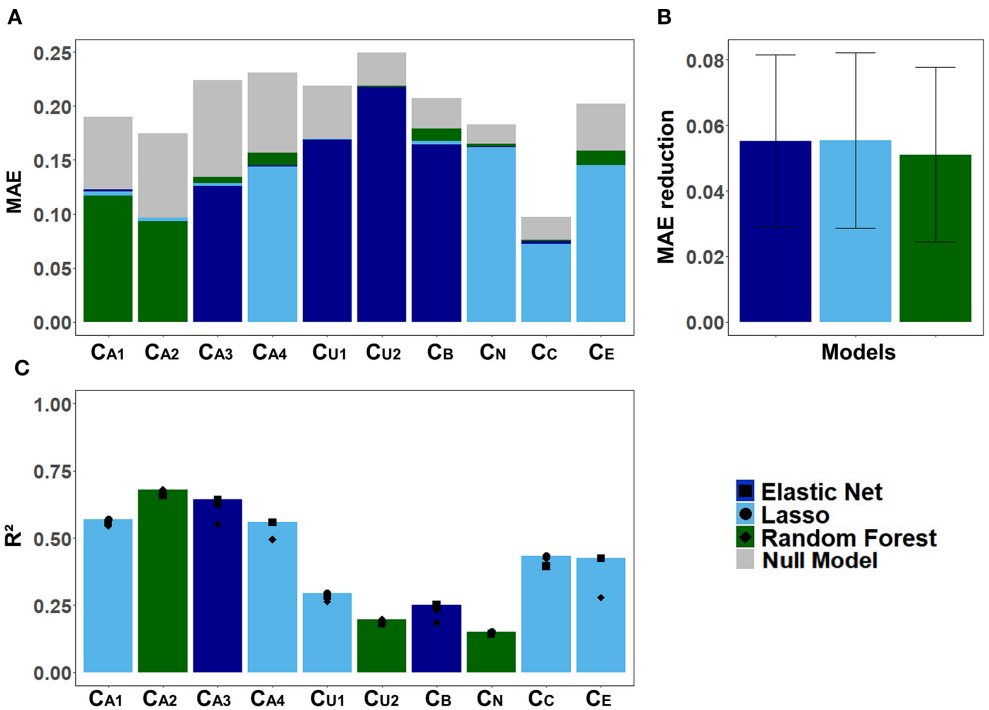


FIGURE 6 | Model-specific predictive performance accuracy for the CAFPAs on the test set, averaged over multiple imputed data sets. Different models are color-coded. **(A)** Mean absolute error (MAE) for each CAFPA. indicates the predictive performance of the null model, and the foremost bar color denotes the model with best predictive performance. **(B)** Mean and standard deviation of the MAE reduction as compared to the null model, averaged over CAFPAs. **(C)** Coefficient of determination (R^2) for each CAFPA. The depicted bar color indicates the model with the best predictive performance. The symbols denote the performance of the respective comparison models.

similar for the different models (**Figure 6B**), with the random forests performing slightly worse.

Figure 6C shows the coefficient of determination (R^2) for labeled CAFPAs in the test set. In line with the MAE results, the plot indicates that the performance of lasso regression, elastic net, and random forests was very similar. However, the random forest performed slightly worse for some CAFPAs (C_{A3}, C_B, C_E). In comparison over CAFPAs, larger differences in predictive performance occurred. The audiogram-related CAFPAs C_{A1}-C_{A4} were predicted best, while performance accuracy was lowest for the suprathreshold CAFPA C_{U2} and the neural CAFPA C_N.

With **Figure 7** we provide a more detailed view on the models' predictive performance for different CAFPAs. The scatter plots (**Figure 7A**) indicate the labeled vs. predicted CAFPAs for individual patients. In addition to the depicted correlations, the range of the labeled and predicted CAFPA values with regard

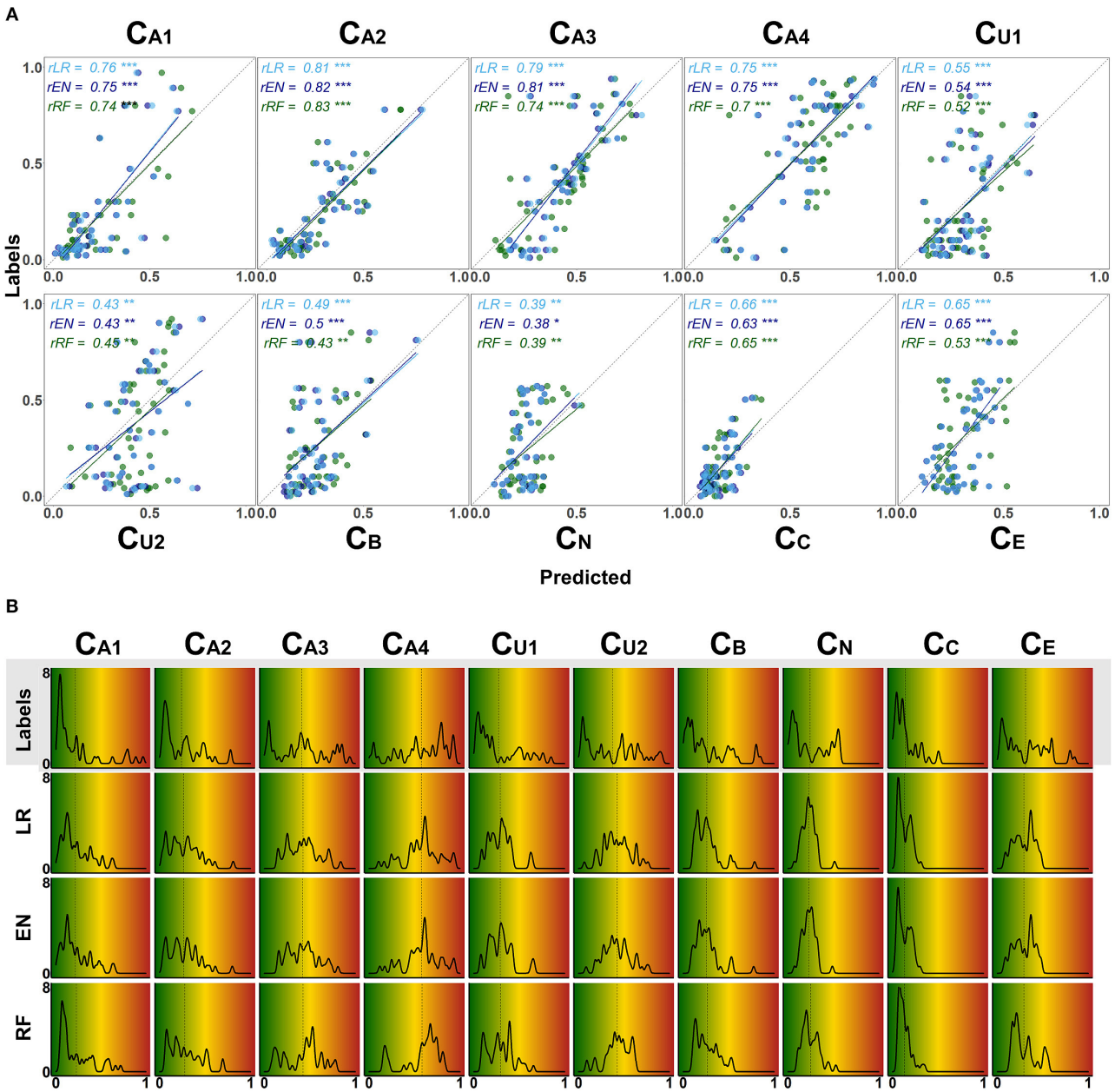


FIGURE 7 | Detailed overview on model performance for different CAFPAs. LR (lasso regression); EN (elastic net); RF (random forests) **(A)** Predicted vs. labeled CAFPA values (for the test set, averaged over imputed data sets) for all CAFPAs. The three models are color-coded. The Pearson correlation *r* and corresponding linear associations are depicted in the corresponding color. The level of significance *p* is indicated by *, **, ***, for $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively. The dotted line indicates perfect prediction. **(B)** Absolute frequency density plots (bandwidth = 0.015) for the, ten CAFPAs. Different rows depict the distributions of the labeled CAFPAs, next to the distributions achieved by the three models. The *x*-axis indicates the CAFPA values; the *y*-axis the absolute frequency of the labeled CAFPAs and those predicted by the respective model. In the background, the color codes corresponding to the CAFPA interval [0 1] are depicted (cf. **Figures 1B**, **9**).

to the interval [0 1] is being visualized in the plot. Except for the neural CAFPA C_N and the cognitive CAFPA C_C, all labeled CAFPAs cover the complete range of potential values. The predicted CAFPAs for all three models generally cover a smaller range of potential CAFPA values, that is, very high values are rarely predicted by the models. Only for the audiogram-related CAFPAs C_{A2}-C_{A4} both labeled and predicted values span the complete interval [0 1].

The range of the predicted CAFPAs is further visualized in **Figure 5B**. Frequency density plots for all CAFPAs are depicted for labeled and predicted values. The labeled CAFPAs are generally distributed over the whole interval [0 1], with a

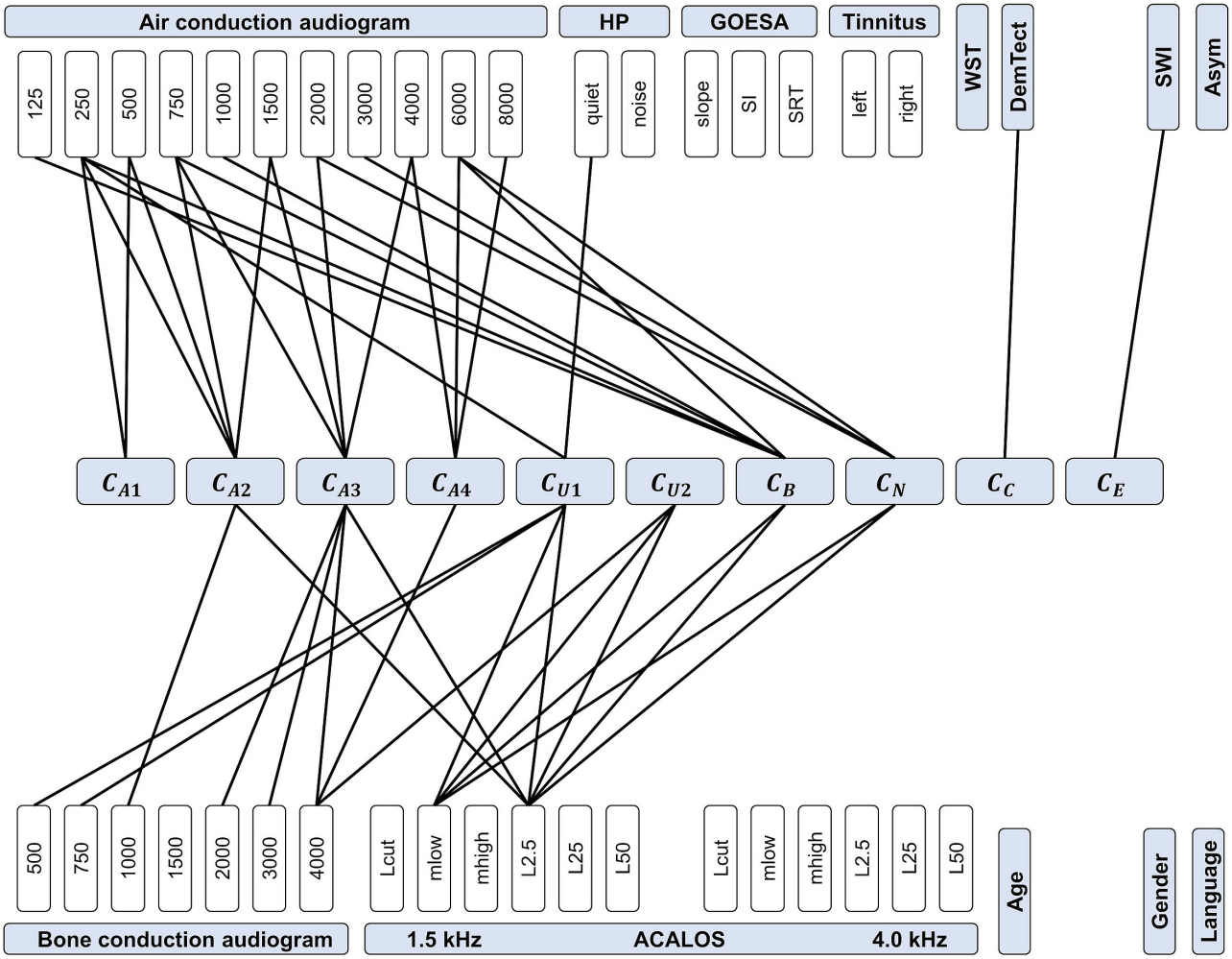


FIGURE 8 | Feature importance for predicting CAFPAs. CAFPAs are displayed in the center of the figure; features in the upper and bottom parts. Measurement parameters are represented by white boxes (cf. **Table 1** for abbreviations and units). Black lines displayed as connections between the measures and the CAFPAs indicate features selected by all three models to be relevant for the prediction of the respective CAFPA.

tendency towards lower (green) values especially for the CAFPAs C_{A1}, C_{A2}, C_{U1}, and C_C which characterizes the expert ratings, but also the underlying audiological data. To conclude on a sound prediction of CAFPAs, in addition to a high correlation between labeled and predicted CAFPA values and an overlapping value range of the two, the shape of the predicted CAFPA distribution should be similar to the one of the labeled CAFPA scores (see **Figure 7B**). For most CAFPAs and models, the label distributions are well reproduced by the distributions of the predicted CAFPA scores. Differences between models are smaller than differences between CAFPAs. The strongest similarity between labeled and predicted scores is obtained for the audiogram-related CAFPAs C_{A1}-C_{A4}. However, the distributions for C_N and C_C are limited to a restricted CAFPA range as compared with the label distributions. For example, the two maxima of the label C_N distribution are not covered by the distributions of the predicted scores.

Feature Importance

For all models, we assessed feature importance using Leave-One-Out-Cross-Validation (LOOCV). **Figure 8** provides a summary of the most relevant features for predicting the different CAFPAs. All features (audiological measures) included in the data set (cf. **Table 2**) are represented in the plot, and those measures that were selected as relevant features by all three models are connected with the respective CAFPA. The candidate features for each model separately are provided in the **Supplementary Figures 3–5**.

The most important features for the audiogram-related CAFPAs C_{A1}-C_{A4} are air and bone conduction audiogram for plausible frequencies, i.e., frequencies that increase over the four CAFPAs defined for different frequency ranges. For the cognitive CAFPA C_C and the socio-economic CAFPA C_E, the models agreed on only one respective feature, namely DemTect and the Scheuch-Winkler-Index, respectively. In contrast, the

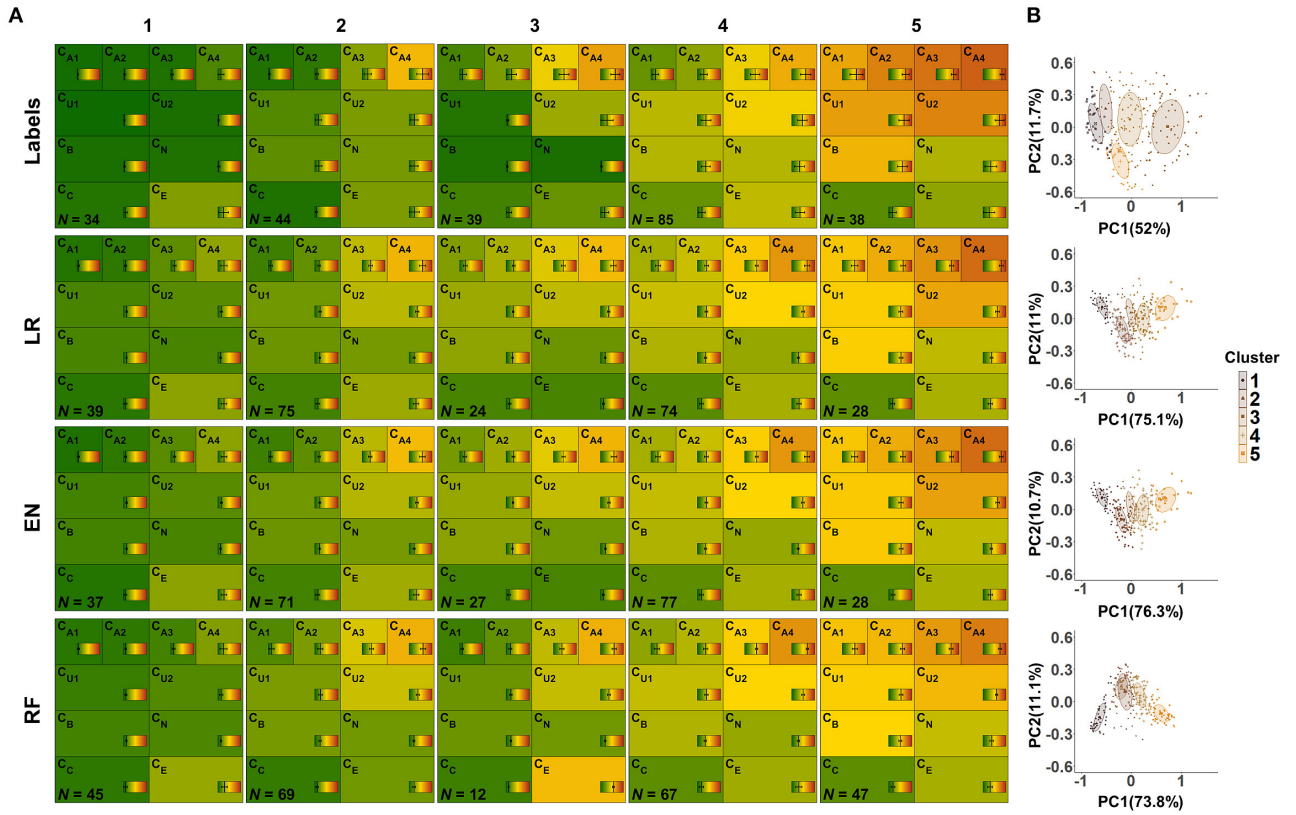


FIGURE 9 | CAFPA clusters estimated using model-based clustering. Rows depict the clusters for labeled CAFPAs (labeled data set), as well as the predicted CAFPAs for matched, unlabeled patients, originating from the models lasso regression (LR), elastic net (EN), and random forest (RF), respectively (unlabeled data set). **(A)** CAFPA patterns for five estimated clusters (columns). The average CAFPAs assigned to each cluster are depicted by the color of the respective area as well as by the vertical line in the color bar. Standard deviations are depicted by horizontal lines in the color bars. *N* indicates the number of patients assigned to each cluster. **(B)** Combined representation of the five estimated clusters. For visualization purposes, the clusters are displayed in the plane given by the first and second principal component (PC) estimated for the respective CAFPA data. The percentage of explained variance by PC1 and PC2 is indicated on the respective axis. Different colors of data points and two-dimensional Gaussian distributions correspond to the different clusters in the columns of **(A)**.

selected features for the suprathreshold CAFPAs C_{U1} and C_{U2}, as well as the binaural CAFPA C_B and neural CAFPA C_N are more widely distributed over different audiological measures. Some audiological measures such as ACALOS at 4.0 kHz or tinnitus, and demographic information as well as the asymmetry score were not selected by all of the models for any CAFPA as relevant features, but at least by one model (see **Supplementary Figures 3–5**).

Model Evaluation Based on Unlabeled Cases

Next, we applied the three models to unlabeled cases for the purpose of investigating the feasibility of predicting plausible CAFPAs also for unlabeled cases. This is an important step toward a CDSS for audiology. Model-based clustering was then used to estimate distinguishable clusters in the ten-dimensional CAFPA data. According to a combination of visual inspection and the BIC, the labeled CAFPAs were best characterized by five clusters using the model $\lambda_k A_k$ with the identifier VVI. Accordingly, the distribution of the covariance matrix Σ_k is diagonal, with varying volume and shape, and an orientation aligned with the coordinate axes (51). Six clusters with the same covariance parameterization reached a marginally higher BIC value (BIC = 1698.8) as compared to five clusters (BIC = 1695.3, **Supplementary Figure 6**). The additional cluster, however, mainly leads to a separation of the healthy patients into two clusters with higher and lower values for the socio-economic CAFPA C_E (**Supplementary Figure 7**). As separating healthy patients solely on socio-economic status is undesirable, we argue for using five clusters for further analysis. We then applied the same clustering method to the CAFPAs for the 240 matched, unlabeled cases which we predicted using the previously trained lasso regression, elastic net, and random forest. The obtained clusters are depicted in **Figure 9A** using the typical CAFPA representation that was introduced and used in Buhl et al. (8, 23). **Figure 9B** additionally displays a combined representation of the five clusters for assessing how well the clusters can be distinguished.

From the left to the right, the labeled CAFPA patterns (labeled data set; first row of **Figure 9A**) indicate an increasing

degree of hearing loss which is expressed by increasing average CAFPA values. The largest differences between the clusters occur for the audiogram-related CAFPAs C_{A1}-C_{A4}. In comparison to the CAFPA distributions published in Buhl et al. (23), the obtained clusters are in line with normal hearing (cluster 1), different degrees of high-frequency hearing loss (cluster 24), and a more severe, broadband hearing loss (cluster 5). The corresponding plot in **Figure 9B** shows five distinguishable clusters.

The clusters for predicted CAFPAs on unlabeled cases using the three models (unlabeled data set; second to fourth row in **Figure 9A**) show CAFPA patterns that are very similar to the labeled CAFPA patterns. However, different numbers of cases were associated to the different clusters, with generally more patients allocated to the clusters with lower CAFPAs. The largest deviation in terms of patients' allocation frequency occurred for random forest, where cluster 5 includes more patients, but on average with less severe hearing loss. This is consistent with the generally lower CAFPA values that the models predicted, in contrast to labeled CAFPAs (cf. **Figure 7B**). Clusters 2 and 3 for random forest are very similar, with the main difference in the socio-economic CAFPA C_E. Cluster 3 only contains 12 patients, which is also visible in **Figure 9B**. In general, similar clusters were obtained for the three models, i.e. the models agreed on the cluster allocation for most of the cases. The agreement between lasso regression and elastic net amounts to 96% and for both lasso regression and random forests and elastic net and random forests to 68%. Further, this similarity becomes evident in **Figure 9B**), where clusters are displayed on a similar plane in the dimensions of the two first principal components, i.e., PC1 and PC2 are explaining similar amount of variance. In contrast to the clusters for lasso regression and elastic net, the clusters for random forest are depicted with opposite sign with respect to PC2, which is however the same due to symmetry of principal component analysis. Here, the clusters 2 and 3 overlap considerably.

DISCUSSION

The present study proved the feasibility of automatically predicting Common Audiological Functional Parameters (CAFPAs) from audiological measures. For developing a clinical decision-support system (CDSS) using CAFPAs as interpretable, intermediate representation of audiological knowledge, the automatic prediction of CAFPAs comprises the last step towards a full working first prototype of such a system. We predicted CAFPAs on the expert-determined data from Buhl et al. (23) using lasso regression, elastic net, and random forests. Interpretability of the model predictions was assessed by feature importance measures, and the potential of predicting CAFPAs for unlabeled cases was evaluated using model-based clustering.

Prediction of CAFPAs

The three models worked reasonably well in predicting the CAFPAs, even though optimal predictive performance cannot yet be achieved. One reason is the limited amount of available data, especially in the range of hearing deficits, and second the choice of the models to some extent. That is, due to the small number of available labeled clinical cases, it was plausible to start with rather simple models to avoid overfitting. As soon as more data becomes available, model flexibility and complexity could be increased, and the here trained methods can be further evaluated to determine which of them turns out to be optimal for CAFPA prediction within a CDSS. Given the available data, the prediction accuracy of the three models was similar, while larger differences occurred between the different CAFPAs, i.e. not all CAFPAs were equally well predicted.

One explanation for performance differences among CAFPAs could be that some CAFPAs are more directly related to the audiological measures than others. This aspect is further discussed in the next section, where we turn to feature importance. A second explanation may be that experts more strongly agree when labeling some of the CAFPAs. Especially given a continuous scale, experts' ratings can be expected to differ from each other to some extent. For example, a meta-analysis of inter-rater reliability on performance status assessment in cancer patients indicated good agreement between raters for about half of the studies; the other half achieved only low to moderate agreement (52). Another study investigated the inter- and intra-rater reliability of audiologists in the estimation of hearing thresholds in newborns, using auditory brainstem response (53). The intra-class correlation of 0.873 was concluded to be satisfactory. However, this value indicates that differences between raters exist. Thus, labels provided by experts, as in the current study, may introduce some bias themselves, although Buhl et al. (23) qualitatively found a good agreement among experts for two reference cases which were given to multiple experts. Such experts' biases, in turn, could lead to less optimal predictions for some of the CAFPAs by using statistical models. To account for these biases and to measure the extent of error introduced by experts, future studies are needed to generate labels by multiple experts for the same cases.

Model Interpretability via Feature Importance Assessment

By analyzing feature importance, we gained crucial insights into the model-building process as well as into the relationships between audiological measures and CAFPAs. Without exception, all models selected audiologically plausible features for predicting different CAFPAs. This means that the automated generation of CAFPAs could be demonstrated to build upon similar audiological measures like physicians are expected to use in their decision making. Thus, the differences in predictive performance of the models for different CAFPAs (cf. section Prediction of CAFPAs) can be assumed to be due to the measures contributing to the respective CAFPA, as indicated by feature importance. For example, the threshold-related CAFPAs C_{A1} and C_{A2} are among the best-predicted ones. These are closely related to the audiogram (8). For predicting them, the models selected suitable audiogram frequencies, as well as the hearing threshold level L2.5 at 1.5 kHz from the adaptive categorical loudness scaling (ACALOS). In contrast, the CAFPAs that were not as well

predicted (e.g., neural CAFPA C_N; binaural CAFPA C_B) may be more vaguely related to the measures. That is, impairment in the neural and binaural domain cannot be directly inferred from a single audiological measure, but rather from a combination of audiological measures. Thus, for these CAFPAs, additional measures that better characterize the respective functional aspect need to be included in future test batteries.

In several regards, assessing feature importance contributes to interpretability of the decision-making process. In model-building, it gives access to information with respect to features which were selected by the model. Thereby, it also allows analyzing how experts derived the CAFPAs in the current study, as well as characterizing the data set itself. In addition, being provided with audiological measures (as input of the model) and the derived CAFPAs (output), physicians may be able to understand and trust the automatized generation of the CAFPAs in a CDSS. Therefore, feature importance also helps to achieve physicians trust towards the diagnostic system and could ensure the physician about the validity of decisions provided by the model. Both are crucial for enhancing acceptability and for reinforcing future implementations of an audiological CDSS into the clinical routine (15, 16). The models considered in this study all belong to "intrinsically interpretable" models according to Jung and Nardelli (45), that is, the selected features directly provide interpretability to the experts. However, if in the future more complex models are used, explanations of model predictions that are most informative to specific users could be constructed using the probabilistic model described in Jung and Nardelli (45).

Additionally, by demonstrating that the CAFPAs can be predicted by plausible audiological measures, assessed by commonly used test batteries, here, we provide further empirical support for the concept of the CAFPAs as an abstract representation of the human auditory system. That is, machine learning models were generally capable to learn the underlying relation between audiological measures and the CAFPAs. This is especially relevant for future applications of a CDSS employing the CAFPAs, since predictions in the medical field need to be grounded on available knowledge in the given domain to avoid flawed predictions (54). For instance, in Cooper et al. (55) a neural network predicted low or high risk of in-hospital mortality for pneumonia patients. Subsequent studies analyzing feature importance, however, have revealed that the model assumed asthma to be a protective factor, even though in reality the opposite is true. The prediction error was caused by asthma patients being more carefully treated, due to their higher mortality risk (56). Clearly, this example highlights the importance of the interpretability of predictions within a CDSS in general, and together with the presented results it demonstrates the benefit of the interpretability of the CAFPA predictions that we could achieve in this work. Based on our hitherto available results on CAFPAs, physicians can be provided with the audiological measures that are most influential for the respective CAFPA prediction. As a next step towards a CDSS for audiology, it will be of interest to further enhance interpretability, i.e. by providing physicians with the exact proportions of measurement importance.

Model Evaluation on the Unlabeled Data Set

A future CDSS would have to be applied to unlabeled cases. Thus, it must be possible to evaluate if plausible CAFPAs can be predicted for unlabeled cases. For this purpose, we applied the trained models on a demographically matched data set of cases for which no labeled CAFPAs were available. Subsequently, we applied model-based clustering on the predicted CAFPAs and obtained five distinguishable clusters that resemble the clusters contained in the labeled CAFPAs.

In clinical practice, different audiological findings occur, such as cochlear hearing loss related to inner ear dysfunction, conductive hearing loss related to middle ear dysfunction, or central hearing loss related to impaired transmission of neuronal signals to the brain. As the data set used in this study consists of a rather small number of clinical cases, it seems plausible that not all audiological findings are well represented in the data set. In particular, the most frequent cases in the current data set are high-frequency hearing loss patients, broadband hearing loss patients, and normal hearing individuals. Thus, the five clusters represent the most frequent audiological findings in the underlying data set well, including different degrees of hearing loss (23). Consequently, it can be assumed that collecting a sufficient amount of more severe clinical cases for additional audiological findings would allow differentiating more clusters.

The performance differences between models for different CAFPAs are reflected in the resulting clusters, as these models were used for the prediction of the CAFPAs for unlabeled cases. If prediction accuracy can be improved in the future for certain CAFPAs, e.g., by including larger data sets and more measures, the separation of audiological findings by the clustered CAFPA patterns will further improve. However, already with the current prediction accuracy, plausible and distinguishable patterns were demonstrated.

Finally, assessing the obtained clusters using the graphical representation of CAFPA patterns, which was introduced by Buhl et al. (8), allows for direct comparability of audiological findings, and it contributes to interpretability of the CDSS by providing a visualization of the functional aspects which describe the group of patients belonging to the respective cluster.

Clinical Decision-Support System Using CAFPAs

On the way of setting up a CDSS using CAFPAs as interpretable, intermediate layer, the current study closes the gap towards a CDSS working with the input data from a single patient: The prediction models trained here can be used in the future to automatically generate CAFPAs, based on which a classification of audiological findings can be performed. The classification performance could be compared to the classification performance based on the labeled CAFPAs from the expert data set (57).

Most potential for improving toward a testable CDSS lies in applications of the here described models and their extension to larger clinical databases in the model-building process. This is because currently we obtained different performance for

different CAFPAs. The analysis of feature importance revealed that the CAFPAs were backed up by different amounts of appropriate audiological measures. Hence, data sets are needed that contain a higher number of patients for all clinically relevant audiological findings, which are characterized by a test battery with information about all functional aspects covered by CAFPAs. In addition, feature importance analysis could also be used in the future to identify redundant audiological measures contained in test batteries used in clinical settings.

For the purpose of integrating data from different clinical test batteries comprising different audiological measures, the CAFPAs act as abstract representation and data standardization format which is independent from the exact choice of measures. Especially data from electronic health records (EHR), i.e. digitally available data from different clinics, could be easily integrated as training data, if CAFPA labels are available for at least some of them. Expert-based estimations of CAFPAs are arguably the most time-consuming. Our future aim is to estimate CAFPAs by a combination of algorithmic generation and expert-coding. For example, experts could confirm and revise automatically estimated CAFPAs instead of labeling each patient case based on audiological data alone.

CONCLUSION

In the current study, we applied three modeling approaches, lasso regression, elastic net, and random forests, for the prediction of Common Audiological Functional Parameters (CAFPAs). As all three models provide similar predictive performance, currently all appear suitable choices for an algorithmic prediction of the CAFPAs. We demonstrated that it was possible to estimate CAFPAs as intermediate layer in a clinical decision-support system for audiology, that is, as abstract and interpretable representation for potential users of a CDSS for audiological decision-making.

In line with the aim of setting up an interpretable CDSS for audiology, different aspects provide interpretability to the future users of the tool. First, the CAFPAs themselves act as interpretable representation of audiological knowledge which is independent of the exact choice of measurements, that is, the user can assess the functional aspects that are responsible for the classification of a certain audiological finding. Second, the analysis of feature importance helps the user to reproduce which measures are influential to the estimation of CAFPAs.

Finally, the reported cluster analysis allowed assessing CAFPA prediction performance on unlabeled cases. This is an important property to be covered in a future CDSS. The achieved cluster similarity between labeled and predicted CAFPAs revealed that the trained models generalize well to unlabeled cases, which could also be visually assessed by the CAFPA patterns. Building upon previous work by Buhl et al., the present work is a substantial step towards a CDSS for audiology. However, the models still need to be applied and evaluated on new, larger and more variable clinical data sets in the future. Interpretability needs to be always maintained, even if the models described here might become more flexible when tuned and applied to future data.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: According to the data usage agreement of the authors, the datasets analyzed in this study can only be shared upon motivated request. The analyses scripts can be found here: http://doi.org/10.5281/zenodo.4282723. Requests to access these datasets should be directed to Mareike Buhl, mareike.buhl@uni-oldenburg.de, Samira K. Saak, samira.kristina.saak@uni-oldenburg.de.

AUTHOR CONTRIBUTIONS

MB provided the data. SS conducted the data analysis which was continuously discussed with all authors. SS and MB drafted the manuscript and all authors contributed to editing the manuscript. All authors conceptualized and designed the study.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 – Project ID 390895286.

ACKNOWLEDGMENTS

We thank the Hörzentrum Oldenburg GmbH for the provision of the patient data and all audiological experts for their participation in the expert survey.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2020.596433/full#supplementary-material

REFERENCES

1. Tiffen J, Corbridge SJ, Slimmer L. Enhancing clinical decision making: development of a contiguous definition and conceptual framework. (2014) 30:399–405. doi: 10.1016/j.profnurs.2014.01.006
2. Schwartz A, Elstein AS. Clinical reasoning in medicine. *Clinical reasoning in the health professions*. Philadelphia, PA. (2008). p. 223–34.
3. Khullar D, Jha AK, Jena AB. Reducing diagnostic errors—why now? (2015) 373:2491. doi: 10.1056/NEJMp1508044
4. Organization WH. *Deafness and hearing loss*. Available online at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed February 10, 2020).
5. Nations U. *World population ageing* (2015). Available online at: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf (accessed August 12, 2020).

6. Lamond D, Farnell S. The treatment of pressure sores: a comparison of novice and expert nurses' knowledge, information use and decision accuracy. (1998) 27:280–6. doi: 10.1046/j.1365-2648.1998.00532.x
7. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. (2003) 78:775–80. doi: 10.1097/00001888-200308000-00003
8. Buhl M, Warzybok A, Schädler MR, Lenarz T, Majdani O, Kollmeier B. Common Audiological Functional Parameters (CAFPAs): statistical and compact representation of rehabilitative audiological classification based on expert knowledge. (2019) 58:231–45. doi: 10.1080/14992027.2018.1554912
9. Shortliffe EH, Cimino JJ. *Biomedical informatics*. Springer (2006). doi: 10.1007/0-387-36278-9
10. Beam AL, Kohane IS. Big data and machine learning in health care. *Jama*. (2018) 319:1317–8. doi: 10.1001/jama.2017.18391
11. Paul M, Andreassen S, Tacconelli E, Nielsen AD, Almanasreh N, Frank U, et al. Improving empirical antibiotic treatment using TREAT, a computerized decision support system: cluster randomized trial. (2006) 58:1238–45. doi: 10.1093/jac/dkl372
12. Dong Z, Yin Z, He M, Chen X, Lv X, Yu S. Validation of a guideline-based decision support system for the diagnosis of primary headache disorders based on ICHD-3 beta. (2014) 15:40. doi: 10.1186/1129-2377-15-40
13. Shibl R, Lawley M, Debuse J. Factors influencing decision support system acceptance. (2013) 54:953–61. doi: 10.1016/j.dss.2012.09.018
14. Davis FD. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *Int anach tud*. (1993) 38:475–87. doi: 10.1006/imms.1993.1022
15. Walter Z, Lopez MS. Physician acceptance of information technologies: ole of perceived threat to professional autonomy. (2008) 46:206–15. doi: 10.1016/j.dss.2008.06.004
16. Wendt T, Knaup-Gregori P, Winter A. Decision support in medicine: a survey of problems of user acceptance. (2000) 77:852–6.
17. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inform*. (2018) 6:e24. doi: 10.2196/medinform.8912
18. Song XD, Wallace BM, Gardner JR, Ledbetter NM, Weinberger KQ, Barbour DL. Fast, continuous audiogram estimation using machine learning. *Ear ear*. (2015) 36:e326. doi: 10.1097/AUD.0000000000000186
19. Barbour DL, Howard RT, Song XD, Metzger N, Sukesan KA, DiLorenzo JC, et al. Online machine learning audiometry. (2019) 40:918–26. doi: 10.1097/AUD.0000000000000669
20. Goggin LS, Eikelboom RH, Atlas MD. Clinical decision support systems and computer-aided diagnosis in otology. *Otolaryngology*. (2007) 136:s21s6. doi: 10.1016/j.otohns.2007.01.028
21. Sanchez Lopez R, Bianchi F, Fereczkowski M, Santurette S, Dau T. Data-driven approach for auditory profiling and characterization of individual hearing loss. (2018) 22:2331216518807400. doi: 10.1177/2331216518807400
22. Gieseler A, Tahden MA, Thiel CM, Wagener KC, Meis M, Colonius H. Auditory and non-auditory contributions for unaided speech recognition in noise as a function of hearing aid use. (2017) 8:219. doi: 10.3389/fpsyg.2017.00219
23. Buhl M, Warzybok A, Schädler MR, Majdani O, Kollmeier B. Common Audiological Functional Parameters (CAFPAs) for single patient cases: deriving statistical models from an expert-labelled data set. (2020) doi: 10.1080/14992027.2020.1728401
24. Gelfand SA. *Essentials of audiology* (2016). doi: 10.1055/b-006-161125
25. Bharadwaj HM, Verhulst S, Shaheen L, Liberman MC, Shinn-Cunningham BG. Cochlear neuropathy and the coding of supra-threshold sound. (2014) 8:26. doi: 10.3389/fnsys.2014.00026
26. Joris P, Yin TC. A matter of time: internal delays in binaural processing. (2007) 30:70–8. doi: 10.1016/j.tins.2006.12.004
27. Yin TC. *Neural mechanisms of encoding binaural localization cues in the auditory brainstem*. Springer (2002). p. 99159. doi: 10.1007/978-1-4757-3654-0_4
28. Loughrey DG, Kelly ME, Kelley GA, Brennan S, Lawlor BA. Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: a systematic review and meta-analysis. (2018) 144:115–26. doi: 10.1001/jamaoto.2017.2513
29. Fortunato S, Forli F, Guglielmi V, De Corso E, Paludetti G, Berrettini S, et al. A review of new insights on the association between hearing loss and cognitive decline in ageing. (2016) 36:155. doi: 10.14639/0392-100X-993
30. Baker EH. Socioeconomic status, definition. *The Wiley Blackwell Encyclopedia of health, illness, behavior, and society*. Oxford: Wiley Blackwell (2014). p. 2210–4. doi: 10.1002/9781118410868.wbehibs395
31. Brand T, Hohmann V. An adaptive procedure for categorical loudness scaling. (2002) 112:1597–604. doi: 10.1121/1.1502902
32. Kollmeier B, Wesselkamp M. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. (1997) 102:2412–21. doi: 10.1121/1.419624
33. Schmidt K, Metzler P. *WST-Wortschatztest*. Göttingen: Beltz Test 1992.
34. Winkler J, Stolzenberg H. *Adjustierung des Sozialen-Schicht-Index für die Anwendung im Kinder-und Jugendgesundheitssurvey (KiGGS)* Wismar: Wismarer Diskussionspapiere2009.
35. Kalbe E, Kessler J, Calabrese P, Smith R, Passmore A, Brand Ma, et al. DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. (2004) 19:136–43. doi: 10.1002/gps.1042
36. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media (2009).
37. Oetting D, Brand T, Ewert SD. Optimized loudness-function estimation for categorical loudness scaling data. *HeaRes*. (2014) 316:16–27. doi: 10.1016/j.heares.2014.07.003
38. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. (2002) 97:611–31. doi: 10.1198/016214502760047131
39. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. 2010:1–68. doi: 10.18637/jss.v045.i03
40. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. (2007) 16:277–98. doi: 10.1177/0962280206074466
41. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? (2011) 20:40–9. doi: 10.1002/mpr.329
42. Liaw A, Wiener M. Classification and regression by randomForest. *R ews*. (2002) 2:18–22.
43. Grömping U. Variable importance assessment in regression: linear regression versus random forest. *AmStat*. (2009) 63:308–19. doi: 10.1198/tast.2009.08199
44. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. (1983) 70:41–55. doi: 10.1093/biomet/70.1.41
45. Jung A, Nardelli PHJ. An information-theoretic approach to personalized explainable machine learning. *IEEE Signal Process Lett*. (2020) 27:825–9. doi: 10.1109/LSP.2020.2993176
46. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. (2008) 22:31–72. doi: 10.1111/j.1467-6419.2007.00527.x
47. Fraley C, Raftery AE, Murphy TB, Scrucca L. *mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation*. Technical report
48. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics*. (2001) 17:977–87. doi: 10.1093/bioinformatics/17.10.977
49. Abdi H, Williams LJ. Principal component analysis. (2010) 2:433–59. doi: 10.1002/wics.101
50. Jake Lever MK, Naomi Altman. Principal component analysis. *Nat Methods*. (2017) 14:641–2. doi: 10.1038/nmeth.4346
51. Fraley C, Raftery AE. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *JClass*. (2003) 20:263–86. doi: 10.1007/s00357-003-0015-3
52. Chow R, Chiu N, Bruera E, Krishnan M, Chiu L, Lam H, et al. Inter-rater reliability in performance status assessment among health care professionals: a systematic review. *Ann Palliat Med*. (2016) 5:83–92. doi: 10.21037/apm.2016.03.02

53. Zaitoun M, Cumming S, Purcell A, O'brien K. Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results. (2016) 14:59–63. doi: 10.3109/21695717.2016.1110957
54. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N, editors. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015). doi: 10.1145/2783258.2788613
55. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. (1997) 9:107–38. doi: 10.1016/S0933-3657(96)00367-3
56. Ahmad MA, Eckert C, Teredesai A, editors. Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (2018). doi: 10.1145/3233547.3233667
57. Buhl M, Warzybok A, Schädler MR, Kollmeier B. Sensitivity and specificity of automatic audiological classification using expert-labelled audiological data and Common Audiological Functional Parameters. *Int J Audiol.* (2020) 1–11. doi: 10.1080/14992027.2020.1817581

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 Digital Health

ORIGINAL RESEARCH
published: 22 January 2021
doi: 10.3389/fdgth.2020.608920

Unsupervised EEG Artifact Detection and Correction

Sari Saba-Sadiya[1,2], Eric Chantland[2], Tuka Alhanai[3], Taosheng Liu[2] and Mohammad M. Ghassemi[1]*

[1] Human Augmentation and Artificial Intelligence Lab, Department of Computer Science, Michigan State University, East Lansing, MI, United States, [2] Neuroimaging of Perception and Attention Lab, Department of Psychology, Michigan State University, East Lansing, MI, United States, [3] Computer Human Intelligence Lab, Department of Electrical & Computer Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Electroencephalography (EEG) is used in the diagnosis, monitoring, and prognostication of many neurological ailments including seizure, coma, sleep disorders, brain injury, and behavioral abnormalities. One of the primary challenges of EEG data is its sensitivity to a breadth of non-stationary noises caused by physiological-, movement-, and equipment-related artifacts. Existing solutions to artifact *detection* are deficient because they require experts to manually explore and annotate data for artifact segments. Existing solutions to artifact *correction* or removal are deficient because they assume that the incidence and specific characteristics of artifacts are similar across both subjects and tasks (i.e., "one-size-fits-all"). In this paper, we describe a novel EEG noise-reduction method that uses representation learning to perform patient- and task-specific artifact detection and correction. More specifically, our method extracts 58 clinically relevant features and applies an ensemble of unsupervised outlier detection algorithms to identify EEG artifacts that are unique to a given task and subject. The artifact segments are then passed to a deep encoder-decoder network for unsupervised *artifact correction*. We compared the performance of classification models trained with and without our method and observed a 10% relative improvement in performance when using our approach. Our method provides a flexible end-to-end unsupervised framework that can be applied to novel EEG data without the need for expert supervision and can be used for a variety of clinical decision tasks, including coma prognostication and degenerative illness detection. By making our method, code, and data publicly available, our work provides a tool that is of both immediate practical utility and may also serve as an important foundation for future efforts in this domain.

Keywords: electroencephalography, artifact rejection, brain computer interface, unsupervised learning, artifact removal

OPEN ACCESS

Edited by:
Tyler John Loftus, University of Florida, United States

Reviewed by:
Aishwarya Bandla, National University of Singapore, Singapore
Amanda Christine Filiberto, University of Florida, United States

****Correspondence:***
Sari Saba-Sadiya
sadiyasa@msu.edu

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

Received: *22 September 2020*
Accepted: *14 December 2020*
Published: *22 January 2021*

Citation:
Saba-Sadiya S, Chantland E, Alhanai T, Liu T and Ghassemi MM (2021) Unsupervised EEG Artifact Detection and Correction. Front. Digit. Health 2:608920. doi: 10.3389/fdgth.2020.608920

1. INTRODUCTION

Electroencephalography (EEG) devices are pervasive tools used for clinical research, education, entertainment, and a variety of other domains (1). However, most EEG *applications* remain limited by the low signal to noise ratio inherent to data collected by EEG devices. EEG noise sources include movement artifacts, physiological artifacts (e.g., from perspiration), and instrument artifacts (resulting from the EEG device itself). While researchers have developed a number of

methods to identify specific instances of these artifacts (2) in EEG data, most methods require manual labeling of exemplary artifact segments[1] or special hardware, such as Electrooculography electrodes that are placed around the eyes, or large data-sets of templates, such as independent component scalp maps (3).

Manual annotation of artifacts in EEG data is problematic because it is time-consuming and may even be untenable if the specific profiles of artifacts in the EEG data vary as a function of the task, the subject, or the experimental trial within a given task for a given subject, as they so often do. These realities quickly scale the complexity of the artifact annotation problem and make the use of a one-size-fits-all artifact detection method infeasible for many practical use cases.

Even if artifacts could be identified with perfect fidelity, their simple removal (e.g., by deletion of the corrupted segment) may introduce secondary analytic complications that confound the performance of downstream methods that leverage these data. For instance, methods that rely on the stationarity of EEG segments will be confounded by simple removal of the artifact segments. Even the simplest approaches, such as averaging many EEG trials before extracting features (4), may be less effective if artifact occurrence is correlated with the trail type or experimental condition, thereby increasing the likelihood of a type II error and the consequent reduction in experimental power.

An essential challenge of artifact detection in EEG processing is that the definition of "artifact" depends on the specific task at hand. That is, a given EEG segment is an artifact if and only if it impacts the performance of downstream methods by manifesting as uncorrelated noise in a feature space that is relevant to those methods. For instance, muscle movement signatures confound comma-prognostic classification but are useful features for sleep stage identification (5).

The task-specific nature of artifacts makes their detection especially suitable for data-driven unsupervised approaches as the only requirement for the identification of artifacts using such methods is that the artifacts are *relatively* infrequent. That is, when mapping our data into feature spaces that are relevant to the specific EEG task, artifacts should stand out as rare anomalies. Indeed, many state-of-the-art approaches use unsupervised methods for the detection of specific artifact types under specific circumstances. For instance, the *Blink* algorithm described by Agarwal et al. is a fully unsupervised EEG artifact detection algorithm (6) that is effective for the detection of eye-blinks. While existing methods provide excellent performance for specific artifact types, there is a need for additional progress toward generalized artifact detection approaches, that make no assumptions about the task, subject, or circumstances.

It is also possible to go beyond artifact detection to *correct* the EEG trial by removing the artifact signal. EEG artifact removal is one instance of a more general class of noise reduction problems. The removal of noise from signal data has been a topic of scientific inquiry since Shannon laid the foundation for information theory in the 1940s (7); over the years, multiple signal processing approaches to this problem have found their way into EEG research. One such technique for artifact removal that is ubiquitous for EEG processing is Independent Component Analysis (ICA). This method and its modern derivative remain popular among the research community for unsupervised artifact correction. However, ICA still requires EEG experts to review the decomposed signals and manually classify them as either signal or noise. Furthermore, while ICA is undeniably an invaluable tool for many EEG applications, it also has limitations that are particularly poignant when the number of channels is low; ICA can only extract as many independent components as there are channels and will therefore be unable to isolate all independent noise components if the total number of independent noise components and signal sources exceeds the number of EEG electrodes (8).

Artifact removal is an especially common practice for a particular artifact type: the electrode "*pop*." These artifacts result from abrupt changes in impedance, often due to loose electrode placement or bad conductivity (9, 10). Unlike muscle and movement artifacts, electrode pop is extremely localized, often affecting only one electrode channel. Channel interpolation is the process of replacing the signal of a corrupted channel with one that is interpolated from surrounding clean channels. Patrichella et al. demonstrated that knowing specific electrode locations (namely the exact electrode locations for each subject), and the distances between them can improve interpolation results (11, 12). However, this type of additional information is rarely available and often requires special dedicated hardware. Recently, Sadiya et al. proposed a deep learning convolutional auto-encoder based approach to learn task and subject-specific interpolation (13). By iteratively occluding channels in the input and using original data as the ground truth, the model learned how to interpolate channels in a self-supervised manner with no human annotation. Moreover, not only was the model able learn idiosyncratic information, such as subject-specific electrode location, beating state-of-the-art models, it was also possible to use transfer learning to improve performance on previously unseen tasks and subjects.

In this paper, we extend the aforementioned state-of-the-art approaches in artifact detection and rejection by building an end-to-end pipeline that solves both the detection and rejection problems together without making any assumptions concerning the task or artifact type.

Our artifact detection approach uses a collection of quantitative EEG features that are relevant for a wide variety of tasks including coma prognostics (14), diagnosing mental-illness (15), decoding mental representations (16), decoding attention deployment (17), and brain–computer interface design (18). Unsupervised outlier detection algorithms utilize these extracted features to identify artifacts in the EEG data. These unsupervised algorithms only require an estimate of the *frequency* of artifacts in the data, and can detect any artifact type, irrespective of the task. To guarantee that our results accurately represent the capabilities of these unsupervised outlier detectors we carefully selected algorithms that are qualitatively different from each other (for instance relying on local vs global characteristics of the data distributions) and explored hundreds of different possible

[1]Which may be used as "templates" by statistical or rule-based methods for the identification (and potential rejection) of noisy data epochs.

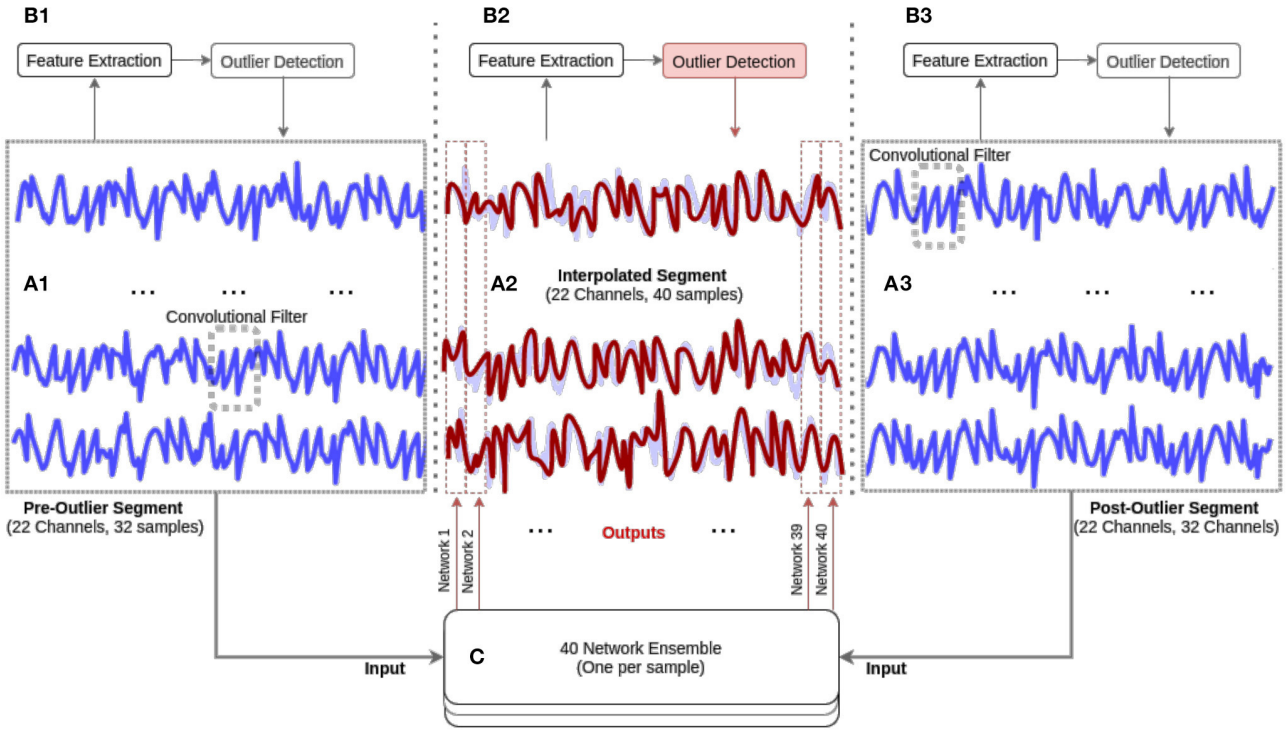


FIGURE 1 | Our methodological approach. The EEG data is first segmented into epochs (see A_1, A_2, A_3). Next, 58 features are extracted and an ensemble of unsupervised outlier detection methods are used (see B_1, B_2, B_3) to identify EEG epochs that are artifact-ridden and require interpolation (see A_2 and B_2). The artifact-ridden epochs are then interpolated by an ensemble of deep encoder-decoder networks (see red line in C).

configurations. Sub-section 2.2.1 provides a comprehensive review of the feature extraction process. Sub-section 2.2.2 details our experimentation with different outlier detection algorithms.

Our artifact correction approach uses a deep encoder-decoder network to correct artifacts that are *not restricted to only one channel*. Specifically, we frame our learning objective as a modified "*frame-interpolation*" task. Frame interpolation is the filling in of missing frames in a video (19). To the best of our knowledge, this is the first work that takes this approach to EEG artifact correction. The proposed approach is also unique in that it does not require the maintenance of any large data-set of templates or annotated data similarly to other state-of-the-art artifact removal methods (6). The model architecture as well as the exact objective formulation are discussed in detail in subsection 2.3.

The data-sets used in this work are discussed in detail in subsection 2.1. The results of the different experiments we conducted can be found in section 3. Finally, we discuss our findings, their broad implications, and the limitations of our approach in section 4.

2. METHODS

In this paper, we propose an end-to-end pre-processing pipeline for the automated identification, rejection, and removal/correction of EEG artifacts using a combination of feature-based and deep-learning models which is intended for use as a general-purpose EEG pre-processing tool. To begin, we provide a brief overview of the data and methodological pipeline, calling out the specific subsections where the full details of each component of the pipeline are discussed.

In **Figure 1** we provide a visualization of our proposed pre-processing pipeline; our method begins by performing unsupervised detection of epoched EEG segments in a 58-dimensional feature space (subsection 2.2). The trials that were not rejected in this initial stage are used to train a deep encoder-decoder network designed to correct artifacts segments (subsection 2.3).

While we demonstrate this method on a particular data set (described below), it is applicable (with no modifications) for any EEG pre-processing work. The methods are presented in the order of their processing within our proposed pipeline.

2.1. Data-Sets

2.1.1. Data Acquisition

Our aim is to demonstrate that unsupervised anomaly detection is successfully used to identify artifacts in EEG data and that these artifacts can be corrected via representation learning methods (see section 2.3). To demonstrate the feasibility of our approach, it is necessary to not only have ground truth artifact annotations but also the ground truth labels for all trials, including those that were annotated as artifacts. While the artifact annotations

allow us to test the unsupervised outlier detection methods, the trial labels allow us to verify that corrected EEG data can indeed be used in conjunction with that regular data for downstream analytic tasks (e.g., training a classification model). Unfortunately, available data sets usually do not contain rejected trials, and even when these annotations are available the original trial label is not included[2]. Therefore, our work is validated on two data-sets, hereinafter referred to as the *orientation* and *color* data-sets, that were previously collected by Saidya et al. (20). We briefly describe these datasets here; additional information about the data-sets is provided in the **Supplementary Material**.

Both experiments were passive viewing tasks. The orientation task stimulus consisted of 6 oriented gratings, the color task stimulus consisted of random dot fields in six different colors. The stimulus was generated using MGL, a library running in Matlab (Mathworks). The data was collected using a 32-electrode actiCHamp cap at 1,000 Hz. For each task, we collected data from seven subjects (four male) for a total of ∼10,000 EEG Trials. All subjects reported normal or corrected to normal vision. The data were examined for noisy trials by expert annotators. Fully annotated and anonymized data-sets will be made available online. Participants gave informed consent and compensated at the rate of 15$ per hour. The experimental procedures were approved by the Michigan State University Institutional Review Board and adhered to the tenets of the Declaration of Helsinki.

2.2. Unsupervised Artifact Detection

To benchmark the different outlier detection methods we collected a list of common features used in EEG research in different domains and applied various unsupervised outlier detection algorithms. Our main objective was to thoroughly investigate the feasibility of unsupervised artifact rejection for EEG.

2.2.1. Feature Extraction

Building on the previous work of Ghassemi et al. (21), we reviewed the EEG literature and constructed a permissive list of several features that are commonly used for EEG classification tasks. In total, we identified and extracted 58 features. The code that extracts these features was written to allow for parallelization of the calculations and is accessible as a downloadable python 3.5 package[3]. See **Table 1** for breakdown and references for all 58 features.

These features can be grouped into three categories that measure the complexity, continuity, and connectivity of EEG activity. Before continuing to discuss our pipeline we will provide high-level intuition behind the inclusion of each category. We encourage the interested reader to refer to the previous work of Ghassemi et al. for a more detailed discussion of the specific features (21).

2.2.1.1. Complexity features (n = 25)

These features measure the complexity of the EEG signal from an information-theoretic perspective and are known to correlate with impaired cognitive functions and the presence of degenerative illnesses. Our first set of features is therefore a collection of information-theoretic complexity measures. Of special interest are the first three features shown in **Table 1** as they are particularly prominent in EEG research: *Shannon's entropy* has been associated with neurological outcomes in post-anoxic coma patients (14); the entropy of the decomposed EEG wavelet signals (known as the *Subband Information Quantity*) have similarly been used in cardiac arrest studies (36, 37). *Tsalis entropy* is a generalization of Shannon's entropy that does not make assumptions about the independence of data channels (as Shannon's entropy does) and has been shown to be particularly useful for the characterization of complexity in EEG data (23).

2.2.1.2. Continuity features (n = 27)

These features capture the regularity and volatility of EEG activity. Bursts, spikes, and unusual changes in the mean and standard deviation in the frequency and power domains are examples of continuity features that are relevant for a variety of clinical tasks. See Hirsh et al. for an in-depth review of continuity and it's relevance to clinical care (38).

2.2.1.3. Connectivity features (n = 6)

. These features reflect the statistical dependence of EEG signal activity *across* two or more channels. Functional connectivity networks are established features of normal brain functioning. We draw on the rich literature on measuring connectivity from EEG signals (39) extracting features that have previously been used for designing brain computer interfaces (18) as well as in mental illness, perception, and attention research [see (15), (16), and (17), respectively].

2.2.2. Outlier Detection Methods

We explored a set of ten algorithms for unsupervised artifact detection; the explored algorithms were inspired by the work of Zhao et al. (40). The algorithms can be divided into two general groups: statistical methods and representation learning methods; they are described in more detail in the "*Statistical Methods*" and "*Representation Learning Based Methods*" sections below. The hyper-parameters of each method were determined by randomly exploring the hyper-parameter space and choosing the settings that yielded the best performance of the methods on the data according to our artifact annotations.

2.2.2.1. Statistical methods

Statistical methods identify anomalies based on statistical measures extracted from the data, thereby producing an "anomaly score" for each trial. The Histogram-Based Outlier detection (HBOS) method uses histograms with dynamic bin widths to detect clusters and anomalies in different feature dimensions. Despite the simplicity of the approach it has been shown to work well on a variety of data types (41). The Local Outlier Factor (LOF) method similarly calculates an "outlier score"; however, instead of global measures, it relies on the local density of the data as it's main indicator (42). Another popular local algorithm, the Angle-Based Outlier Detector (ABOD), calculates the cosine similarity of data points with their neighbors and uses the variance of these scores to generate anomaly

[2]For instance BCI competitions data: http://bbci.de/competition/.
[3]Code available at: https://github.com/sari-saba-sadiya/EEGExtract.

TABLE 1 | EEG Features.

Signal Descriptor	References	Brief description
Complexity features		Degree of randomness or irregularity
Shannon entropy	(22)	Additive measure of signal stochasticity
Tsalis entropy (n = 10)	(23)	Non-additive measure of signal stochasticity
Information quantity ($\delta, \alpha, \theta, \beta, \gamma$)	(24)	Entropy of a wavelet decomposed signal
Cepstrum coefficients (n = 2)	(25)	Rate of change in signal spectral band power
Lyapunov exponent	(26)	Separation between signals with similar trajectories
Fractal embedding dimension	(27)	How signal properties change with scale
Hjorth mobility	(28)	Mean signal frequency
Hjorth complexity	(28)	Rate of change in mean signal frequency
False nearest neighbor	(29)	Signal continuity and smoothness
ARMA coefficients (n = 2)	(30)	Autoregressive coefficient of signal at (t-1) and (t-2)
Continuity features		Clinically grounded signal characteristics
Median frequency		The median spectral power
δ band power		Spectral power in the 0–3 Hz range
θ band power		Spectral power in the 4–7 Hz range
α band power		Spectral power in the 8–15 Hz range
β band power		Spectral power in the 16–31 Hz range
γ band power		Spectral power above 32 Hz
Standard deviation	(31)	Average difference between signal value and it's mean
α/δ ratio	(14)	Ratio of the power spectral density in α and δ bands
Regularity (burst-suppression)	(14)	Measure of signal stationarity/spectral consistency
Voltage $<$ (5, 10, 20 μ)		Low signal amplitude
Diffuse slowing	(32)	Indicator of peak power spectral density $<$8 Hz
Spikes	(32)	Signal amplitude exceeds μ by 3σ for 70 ms or less
Delta burst after spike	(32)	Increased δ after spike, relative to δ before spike
Sharp spike	(32)	Spikes lasting $<$70 ms
Number of bursts		Number of amplitude bursts
Burst length μ and σ		Statistical properties of bursts
Burst band powers ($\delta, \alpha, \theta, \beta, \gamma$)		Spectral power of bursts
Number of suppressions		Segments with contiguous amplitude suppression
Suppression length μ and σ		Statistical properties of suppressions
Connectivity features		Interactions between EEG electrode pairs
Coherence – δ	(14)	Correlation in 0–4 Hz power between signals
Mutual information	(18)	Measure of dependence
Granger causality – All	(33)	measure of causality
Phase lag index	(34)	Association between the instantaneous phase of signals
Cross-correlation magnitude	(35)	Maximum correlation between two signals
Cross-correlation – lag	(35)	Time-delay that maximizes correlation between signals

The 58 EEG features fell into three EEG signal property domains: Complexity features (25 in total), Category features (27 in total), Connectivity features (six in total).

scores (43). Finally, we also trained a One Class SVM Detector (OCSVM), a classic algorithm for outlier detection (44). In this algorithm, an SVM is trained on the entire data-set and afterwards every instance is scored based on its distance from the class boundary; the intuition is that the infrequent outliers will contribute less to the decision boundary calculation and will be more likely to be on the margin of the learned boundary.

As previously mentioned, we selected these detectors to be different in the type of statistical measurements they use. Therefore, it makes sense to also train ensemble classifiers to further improve the outlier detection accuracy. Specifically, we trained five hundred *Locally Selective Combination in Parallel* (LSCP) Outlier Ensembles (45) with different combinations of the algorithms mentioned above.

2.2.2.2. Representation learning based methods

Unlike statistical methods, representation-learning-based outlier detectors do not simply calculate statistical properties of featurized data. The most basic classifier uses auto-encoder (AUTO) based deep learning architectures to learn a lower-dimensional representation of the data that enables the best possible reconstruction of the original signal; the embedding

would be optimized for the common regular data points thereby producing distinctly noisy reconstructions for the outlier trials (46). This classifier can be viewed as a modern update of similar classic outlier detection methods that use methods, such as PCA reconstruction instead of training a deep auto-encoder (PCA) (47). A more sophisticated approach uses Variational Auto-Encoders (VAE). This class of algorithms tries to ensure that the learned embedding captures the structure of the original data by penalizing the classifier if the embedding does not follow a standard normal distribution (48). Finally, we also examine a Generative Adversarial Active Learning (GAAL) outlier detector (49), which uses generative adversarial networks to generate outliers. This method can be used to improve any of the statistical methods described in 2.2.2.1. We also use an extension of the original method to learn multiple generators (MGAAL).

2.3. Artifact Correction

As previously mentioned, encoder-decoder based deep learning methods have proven useful for channel interpolation (13). In this section we discuss an extension of this approach that utilizes the same framework for artifact correction. Namely, given an EEG data segment with an isolated artifact we remove the corrupted segment and use the data samples preceding and proceeding it to fill in the resulting void. This problem is equivalent to the "*frame-interpolation*" task of filling in missing frames in a video (19).

2.3.1. The Model

2.3.1.1. Input representation

The channel interpolation model proposed in Saba-Sadiya et al. (13) represented the EEG as a time series of 2D topologically organized arrays. This reflects the spatial nature of the EEG channel interpolation issue; the interpolated values at different time points are treated as independent. To the best of the author's knowledge, this is a standard assumption for EEG interpolation algorithms. For instance, Petrichella et al. and Courellis et al. calculate the interpolated values of the missing data at each time point separately (11, 12). However, research on convolutional neural networks for EEG decoding and visualization have shown performance benefits from presenting the input as a column of electrodes unfolding in time, as this facilitates the learning of temporal modulations (50). Since artifact correction is first and foremost a process of completing gaps across time we decided to depart from Saba-Sadiya et al. (13) and use a 2D array representation with the number of time steps as the width of the array.

2.3.1.2. Architecture

The best frame interpolation models involve calculating object trajectory and accounting for possible occlusion (e.g., if one object moves behind another). With these "flow computations" and a stack of the frames before and after the missing image a convolutional encoder-decoder can generate realistic intermediate images (19). Unlike video, EEG data have only one spatial dimension (see subsection 2.3.1.1) and are not analogous to local phenomena, such as occlusion or object movement; these can occur as EEG modulations and are often thought of as mostly global in nature (50). Therefore, we only concern ourselves with a stacked convolutional auto-encoder. This architecture is shared by previously discussed state-of-the-art algorithms for both frame interpolation and channel interpolation (13, 50).

The interpolation of each frame is done separately, thus to predict n frames it is necessary to train n networks. Technically this is equivalent to training one ensemble model, however, by separating the networks we allow for easier parallelization of the training process. Specifically, given a series of EEG frames $x_1, x_2, \ldots, x_n$ where x_t is a vector of all the channel values at time t, and assuming that the series is missing all frames between time points t_b and t_e, our network learns to predict x_{t_q} from the two stacks, $x_{t_b-h}, x_{t_b-h+1}, \ldots, x_{t_b}$ and $x_{t_e}, x_{t_e+1}, \ldots, x_{t_e+h}$ where $t_q \in (t_b, t_e)$ and h is some small positive integer representing how many frames before and after the missing segment can be perceived. Every network is trained to predict the value at one specific value of q. Every network takes the same $2h$ frames (half preceding the missing segment and half following it) to calculate the value at a given frame.

2.4. Model Validation Approach

2.4.1. Artifact Detection Method

The performance of the artifact detection methods was assessed by inspecting the agreement between the artifact detection approach and the expert annotations from the two data sets (color and orientation). More specifically, the agreement was measured using the f-score and Cohen's Kappa (first and second values in each cell, respectively). We compared the performance of our model against the expected performance of a classifier with knowledge of the exact number of artifacts; this random classifier is expected to have an f-score of 0.172 and a Kappa of 0.029. We ran the detection algorithms in two configurations, for each subject separately and for the entire aggregated data. We hypothesize that the performance will drop when using the aggregated configuration, as each individual setup for an EEG recording is likely to introduce unique artifacts (due to loose connections or subject-specific circumstances, such as perspiration).

2.4.2. Artifact Correction Method

To optimize the parameters of the artifact correction model, we produced training data from trials that were marked as artifacts free by our unsupervised artifact detection method (section 2.2.2) and randomly removed a segment from the middle of the trial. The h samples proceeding the removed segment and h samples preceding it were used as input for the model while the removed segment was the ground truth (h was a hyper-parameter optimized on the training set). For the purposes of validating the artifact correction model, all EEG data were re-sampled to $200Hz$. The reconstructed segments were $200ms$ each.

2.4.3. End-to-end Assessment Approach

We ran a number of tests to examine if the trials reconstructed by our artifact correction method could be used to enhance the performance of downstream EEG tasks. More specifically, we trained two SVM models to predict the label of the trial from the color data-set: one SVM was trained using the *raw data*, and

TABLE 2 | Comparison of the different unsupervised outlier detection methods when applied to each subject separately.

Statistical methods	**HBOS**	**LOF**	**ABOD**	**OCSVN**	**LSCP**
Orientation	0.564 0.473	0.218 0.065	0.11 0.06	0.41 0.29	0.577 0.489
Color	0.5 0.4	0.241 0.091	0.1 −0.08	0.36 0.23	0.51 0.411
Representation learning	**AUTO**	**PCA**	**VAE**	**GAAL**	**MGAAL**
Orientation	0.53 0.44	0.527 0.426	0.477 0.368	0.429 0.311	0.428 0.309
Color	0.51 0.42	0.477 0.367	0.478 0.368	0.241 0.086	0.389 0.263

We calculated the mean f-score and Cohen's Kappa (first and second row in every cell) across all subject. HBOS, Histogram based outlier detection; LOF, Local outlier factor Method; ABOD, Angle-based outlier detector; OCSVM, One class support vector machine; LSCP, Locally selective combination of parallel outlier Ensembles; AUTO, Auto-encode based method; VAE, Variational auto-encoder based method; GAAL, Generative Adversarial Active Learning; MGAAL, Multi-object Generative Adversarial Active Learning.

TABLE 3 | The performance of the models trained on data aggregated from all the subjects.

Statistical methods	**HBOS**	**LOF**	**ABOD**	**OCSVN**	**LSCP**
Orientation	0.502 0.4	0.246 0.095	0.07 −0.11	0.362 0.234	0.537 0.441
Color	0.476 0.35	0.305 0.15	0.09 −0.108	0.377 0.238	0.463 0.332
Representation learning	**AUTO**	**PCA**	**VAE**	**GAAL**	**MGAAL**
Orientation	0.488 0.338	0.448 0.338	0.447 0.336	0.383 0.246	0.393 0.258
Color	0.414 0.283	0.437 0.312	0.436 0.31	0.185 0.022	0.393 0.258

The f-score and Cohen's Kappa are presented in the first and second row in every cell.

the other was trained using the raw data *after artifact correction*. Both models were validated using 5-fold cross-validation, and the performance of the models on the test set (μ and σ) was reported.

We also evaluated the impact of our artifact correction method on downstream EEG tasks when applied to *clean trials, exclusively*; this evaluation allowed us to test for inadvertent degeneration in signal quality of clean segments when processed by our method. More specifically, we applied our artifact correction method to 20% of *clean* trials and used the resulting data to train an additional SVM model.

3. RESULTS

This section presents the results of the two main components in our pipeline, the artifact detection method and the artifact correction method on the data described in 2.1.

3.1. Artifact Detection Results

In **Table 2**, we compare the *average* performance of the outlier detection methods described in section 2.2.2 when applied to each subject *separately*. Therefore, each value is the mean of the algorithm's performance across subjects. As previously mentioned, the expected performance of a baseline random classifier with knowledge of the exact number of artifacts is an f-score of 0.172 and a Kappa of 0.029. Hence, all models other than the *ABOD* classifier performed significantly better than the baseline (one tailed t-test with a $p = 0.05$ significance level).

Unsurprisingly, the best outlier detector was an *LSCP* ensemble classifier that performed 16.86x better than the baseline method, and 1.03x better than the next best approach; the best performing configuration of the classifier consisted of two *HBOS* classifiers and one *OCSVM*. While it is difficult to interpret ensemble classifiers it is worth noting that the two histogram-based classifiers diverged quite substantially; one using a high number of histogram bins and a rigid outlier scoring policy ($tol = 0.1$) while the other using a smaller number of bins and more relaxed policy ($tol = 0.5$). A simple auto-encoder was the best representation learning algorithm, closely followed by the PCA algorithm. We speculate that the auto-encoder could have possibly had better performance if more data were available for each subject. See our **Supplementary Material** for a breakdown of trial and artifact numbers for each subject.

In **Table 3**, we compare the performance of the outlier detection methods described in section 2.2.2 when applied to the subjects *aggregated* data; that is, subject were not considered separately as they were in the results from **Table 2**. When compared to the results shown in **Table 2**, the performance decreased for most models. This is not surprising as the fundamental assumption of unsupervised methods is that the data are homogeneous with the exceptions of the outliers. Here again, the LSCP method performed the best of the tested approaches. A comparison of the results in **Tables 2**, **3** provide motivation for the development of subject-specific anomaly detection approaches. Moreover, the comparison also highlights that the unsupervised algorithms and the features we extracted can successfully capture both common EEG artifacts and subject-specific idiosyncrasies.

3.2. Artifact Correction Results

3.2.1. Network Optimization

Our first step was to optimize the network hyper-parameter configurations. This included testing different sizes of both the layers and convolution filter, as well as exploring different hyper-parameters, such as optimization algorithms, dropout rates, and activation functions. To train the network we followed the method discussed in section 2.2.2: we randomly extracted 104 samples from the data, the first and last 32 samples were stacked and used as the input to the model, and the sample at position i from the remaining 40 samples was used as the ground truth. Essentially we are training a network to predict the values after removing 40 samples (200ms) using the 32

TABLE 4 | Mean accuracies of simple SVM classifiers.

	Original EEG	EEG with random correction	EEG with artifact correction
All trials	0.3	0.31	0.33
Rejected trials	0.23	0.23	0.29

A simple t-test confirmed that all accuracies were significantly above chance level (1/6 for six different colors) at a $p = 0.05$ level. Original EEG: The Original EEG data. EEG with Random Correction: The EEG data after random artifact free trials were "corrected." EEG with artifact correction: The data after we applied the EEG artifact correction on the trials that were marked as artifact ridden.

samples that before and after the removed segment. The best performing network (lowest loss) was different for different *t*s. The optimal topology for reconstructing sample 20 is available in the **Supplementary Material** as a reference of the type of convolutional U-net architecture used.

3.2.2. End-to-End Assessment

In **Table 4** we compare the classification accuracy of a 5-fold SVM model trained to perform a downstream classification of trial type using down-sampled EEG data with three different configurations of the data: (1) the raw EEG data, (2) the data after correction of artifact segments, and (3) the data following "correction" of a random 40 samples of 20% of the non-artifact segments. Note that while simple this type of analysis is used in actual EEG research (4).

The performance remained comparable after using the artifact correction on trials that did not contain any artifacts. This is a strong indication that the model is indeed able to learn how to reconstruct the original EEG signal. When using the corrected trials with EEG artifacts the classification accuracy improved by 10% overall and over 20% for trials that were marked as containing artifacts. These results successfully demonstrate that our unsupervised end-to-end artifact correction pipeline improves down-stream analysis.

4. DISCUSSION

4.1. Significance of Our Results

In this paper, we presented an end-to-end pipeline that is capable of unsupervised artifact detection and correction. Our results demonstrate that data-driven approaches for unsupervised outlier detection can be extremely useful when applied to the problem of EEG artifact detection. Interestingly, the classifiers with the best performance (HBOS, OCSVM, and the best performing LSCP) are global classifiers; this might indicate that EEG artifacts are better discriminated by global characteristics. This supports our previous observation that artifacts are task specific and infrequent occurrences of uncorrelated noise. It is worth noting that, as demonstrated in **Table 3**, the classifiers we trained were able to learn subject-specific idiosyncrasies.

While the accuracy and agreement between the annotators and the detectors were far from perfect, the Cohen Kappa of the best performing algorithm was comparable to the inter-rater agreement levels of expert annotators reported in the literature; for instance, when asked to annotate, "periodic discharges" (a specific type of artifact) and "electrographic seizure" annotators had a Cohen's Kappa of 0.38 and 0.58, respectively (51). Our results indicate that an unsupervised outlier detection is a feasible approach for generalized EEG artifact detection.

4.2. The Data-Sets

We validated our framework on two novel data-sets. To test the impact of artifact correction algorithms on downstream analysis it is necessary to have ground truth artifact annotation as well as knowledge of the labels of all trials, including those that are artifact ridden. Unfortunately, public data-sets often exclude trials that contain artifacts. Even in the rare occasions in which these trials are made available, the labels are often replaced with a special identifier for rejected trials[4]. We hope our data-sets inspire other researchers to adopt more thorough data publishing practices as data-availability is perhaps the primary limiting factor in artifact correction research.

4.3. The Strength of Unsupervised End-to-End Methods

The accuracy of simple classifiers improved modestly after artifact removal. It is possible that replacing our deep-learning-based artifact removal components with an ICA artifact removal algorithm (52) could yield better results. However, two important distinctions should be made: First, the proposed method does sidestep many weaknesses inherent to ICA (8) (such as the number of independent components being limiting by the number of channels, which is particularly problematic for lightweight commercial EEG setups). Secondly, while the independent component deconstruction itself is data driven and unsupervised, the ICA method still requires visual inspection and analysis of the decomposed signal by human experts. In contrast, our method can be put into effect without any human intervention, making it is suitable for online EEG applications or as a no-cost first step before a more thorough analysis. In general, supervised methods unquestionably out-perform unsupervised ones and we fully acknowledge that the pipeline proposed in this work is no different. It is therefore useful to consider unsupervised methods not as replacements of currently existing algorithms but as complimentary additions to the toolbox of the EEG researcher. With this in mind, we intentionally designed our end-to-end pipelines to be highly modular; An experienced researcher can easily substitute our last component with an ICA artifact removal algorithm, and in contrast, researchers that have access to artifact annotations (for instance by virtue of employing specialized hardware during data acquisition) will be able to use their method in conjunction with ours or sidestep the first processes completely and apply only the artifact correction component before carrying on with the analysis process.

4.4. Limitations

We did not formally evaluate the reconstruction performance of the model because (1) there is not an authoritative literature

[4]For an example of standard EEG publishing practices see the BCI Competition data-sets.

baseline, and (2), insofar as the reconstruction enhances the ability of the downstream classification model to perform their intended classification tasks, the reconstruction is valid and valuable. There are a few limitations that we hope to address in future work. First and foremost, this artifact detection method can only be used if the frequency of the artifacts is low enough for them to be considered outliers. While this is indeed the case for the vast majority of EEG use cases, tasks, such as seizure detection often involve long periods of unusually low signal to noise ratio. Additionally, the performance of our artifact correction network would likely benefit from introducing more complex component into the architecture. For instance, introducing temporal dependencies via an LSTM component would guarantee that the corrected frame at time t influences the frame at time $t + 1$. Finally, our method is in dire need of being validated on additional tasks and data-sets.

Despite the challenges described above, we believe that our work demonstrates the feasibility of an EEG pre-processing pipeline which if adopted could facilitate and expedite the often tenuous process of artifact annotation and removal, and could therefore be extremely beneficial for the general EEG research community.

5. CONCLUSION AND FUTURE WORK

The applications of EEG are numerous and diverse, and while this impacts the particularities of what components are classified as part of the signal vs. artifacts, data homogeneity is a common concern in this area of research. Building on this data science perspective, in this work we appropriated state-of-the-art data-driven methods to construct an end-to-end unsupervised pipeline for general artifact detection and correction. We introduced two new data-sets and demonstrated that the inter-rater reliability of our artifact detection component against expert annotators is comparable to reported inter-human levels. Furthermore, we demonstrated how applying the complete pipeline on a data-set can improve the performance of common downstream analysis. The pipeline makes use of a wide range of handcrafted clinically relevant features, and we believe the released python package will be of use to many in the EEG research community.

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 Michigan State University Human Research Protections Program IRB. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SS-S: data collection and annotation, coding for the Methods section, and writing. EC: data collection and annotation, helped code for the Methods section, and article review. TA: algorithm design and writing. TL: coded the experiment and provided the EEG equipment used for data collection. MG: literature review for, and design of, the models presented in the Methods section, and writing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grant DFI GR100335 from Michigan State University.

ACKNOWLEDGMENTS

The subsection 2.2.1 was heavily based on the content of Ghassemi (21), which was a thesis presented by one of the authors in the Massachusetts Institute of Technology in the year 2018. We thank Dr. Susan Ravizza at Michigan state University for lending us her expertise on EEG data collection and analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2020.608920/full#supplementary-material

REFERENCES

1. Siripornpanich V, Sampoon K, Chaithirayanon S, Kotchabhakdi N, Chutabhakdikul N. Enhancing brain maturation through a mindfulness-based education in elementary school children: a quantitative EEG study. *Mindfulness*. (2018) 9:1877–84. doi: 10.1007/s12671-018-0930-3
2. Urigüen JA, Garcia-Zapirain B. EEG artifact removal—state-of-the-art and guidelines. *J Neural Eng*. (2015) 12:031001. doi: 10.1088/1741-2560/12/3/031001
3. Shamlo NB, Kreutz-Delgado K, Kothe C, Makeig S. EyeCatch: data-mining over half a million EEG independent components to construct a fully-automated eye-component detector. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka) (2013). p. 5845–8.
4. Cichy RM, Ramirez FM, Pantazis D. Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *Neuroimage*. (2015) 121:193–204. doi: 10.1016/j.neuroimage.2015.07.011
5. Ghassemi MM, Moody BE, Lehman LH, Song C, Li Q, Sun H, et al. You snooze, you win: the PhysioNet/Computing in Cardiology challenge 2018. In: *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 (Maastricht), (2018). p. 1–4. doi: 10.22489/CinC.2018.049
6. Agarwal M, Sivakumar R. Blink: a fully automated unsupervised algorithm for eye-blink detection in EEG signals. In: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (Monticello, MN), (2019). p. 1113–21. doi: 10.1109/ALLERTON.2019.8919795
7. Shannon CE. Communication in the presence of noise. *Proc IRE*. (1949) 37:10–21. doi: 10.1109/JRPROC.1949.232969

8. Djuwari D, Kumar D, Palaniswami M. Limitations of ICA for artefact removal. *Conf Proc IEEE Eng Med Biol Soc*. (2005) 5:4685–8. doi: 10.1109/IEMBS.2005.1615516
9. Walczak T, Chokroverty S. *Electroencephalography, Electromyography, and Electro-Oculography: General Principles and Basic Technology*. Elsevier Inc. (2009).
10. Britton JW, Frey LC, Hopp JL, Korb P, Koubeissi MZ, Lievens WE, et al. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants [Internet]*. St. Louis EK, Frey LC, editors. Chicago: American Epilepsy Society (2016).
11. Petrichella S, Vollere L, Ferreri F, Guerra A, Määttä S, Könönen M, et al. Channel interpolation in TMS-EEG: a quantitative study towards an accurate topographical representation. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Florida) (2016). p. 989–92. doi: 10.1109/EMBC.2016.7590868
12. Courellis HS, Iversen JR, Poizner H, Cauwenberghs G. EEG channel interpolation using ellipsoid geodesic length. In: *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (Shanghai) (2016). p. 540–3. doi: 10.1109/BioCAS.2016.7833851
13. Saba-Sadiya S, Alhanai T, Liu T, Ghassemi M. EEG channel interpolation using deep encoder-decoder networks. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Seoul) (2020).
14. Tjepkema-Cloostermans M, van Meulen F, Meinsma G, van Putten M. A cerebral recovery index (CRI) for early prognosis in patients after cardiac arrest. *Crit Care*. (2013) 17:R252. doi: 10.1186/cc13078
15. Uhlhaas P, Singer W. Abnormal neural oscillations and synchrony in schizophrenia. *Nat Rev Neurosci*. (2010) 11:100–13. doi: 10.1038/nrn2774
16. Hipp J, Engel A, Siegel M. Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*. (2011) 69:387–96. doi: 10.1016/j.neuron.2010.12.027
17. Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D. Visual input enhances selective speech envelope tracking in auditory cortex at a "Cocktail Party". *J Neurosci*. (2013) 33:1417–26. doi: 10.1523/JNEUROSCI.3675-12.2013
18. Ang KK, Chin ZY, Zhang H, Guan C. Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs. *Pattern Recogn*. (2012) 45:2137–44. doi: 10.1016/j.patcog.2011.04.018
19. Jiang H, Sun D, Jampani V, Yang MH, Learned-Miller E, Kautz J. Super SloMo: high quality estimation of multiple intermediate frames for video interpolation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT) (2018). doi: 10.1109/CVPR.2018.00938
20. Saba-Sadiya S, Chantland E, Liu T. *Decoing EEG From Passive Viewing*. GitHub (2020). Available online at: https://github.com/sari-saba-sadiya/DEPV
21. Ghassemi MM. *Life After Death: Techniques for the Prognostication of Coma Outcomes After Cardiac Arrest*. Cambridge, MA: Massachusetts Institute of Technology (2018).
22. Shannon CE, Weaver W. *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press (1998).
23. Geocadin R, Muthuswamy J, Sherman D, Thakor N, Hanley D. Early electrophysiological and histologic changes after global cerebral ischemia in rats. *Mov Disord*. (2000) 15:14–21. doi: 10.1002/mds.870150704
24. Shin HC, Jia X, Nickl R, Geocadin RG, Thakor Ast NV. A subband-based information measure of EEG during brain injury and recovery after cardiac arrest. *IEEE Trans Biomed Eng*. (2008) 55:1985–90. doi: 10.1109/TBME.2008.921093
25. Oppenheim AV, Schafer RW. From frequency to quefrency: a history of the cepstrum. *IEEE Signal Process Mag*. (2004) 21:95–106. doi: 10.1109/MSP.2004.1328092
26. Wolf A, Swift JB, Swinney HL, Vastano JA. Determining Lyapunov exponents from a time series. *Phys D Nonlin Phen*. (1985) 16:285–317. doi: 10.1016/0167-2789(85)90011-9
27. Accardo A, Affinito M, Carrozzi M, Bouquet F. Use of the fractal dimension for the analysis of electroencephalographic time series. *Biol Cybern*. (1997) 77:339–50. doi: 10.1007/s004220050394
28. Oh SH, Lee YR, Kim HN. A novel EEG feature extraction method using Hjorth parameter. *Int J Electron Electric Eng*. (2014) 2:106–10. doi: 10.12720/ijeee.2.2.106-110
29. Hegger R, Kantz H. Improved false nearest neighbor method to detect determinism in time series data. *Phys Rev E*. (1999) 60:4970. doi: 10.1103/PhysRevE.60.4970
30. Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. Hoboken, NJ: John Wiley & Sons (2015).
31. Ross SM. *Introductory Statistics*. Cambridge, MA: Academic Press (2017).
32. Stern JM. *Atlas of EEG Patterns*. Philadelphia, PA: Lippincott Williams & Wilkins (2005).
33. Blinowska KJ, Kuś R, Kamiński M. Granger causality and information flow in multivariate processes. *Phys Rev E*. (2004) 70:050902. doi: 10.1103/PhysRevE.70.050902
34. Stam C, Nolte G, Daffertshofer A. Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. *Hum Brain Mapp*. (2007) 28:1178–93. doi: 10.1002/hbm.20346
35. Kay SM. *Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ: Prentice Hall PTR (1993).
36. Shin H, Tong S, Yamashita S, Jia X, Geocadin G, Thakor N. Quantitative EEG and effect of hypothermia on brain recovery after cardiac arrest. *IEEE Trans Biomed Eng*. (2006) 53:1016–23. doi: 10.1109/TBME.2006.873394
37. Jia X, Koenig M, Nickl R, Zhen G, Thakor N, Geocadin R. Early electrophysiologic markers predict functional outcome associated with temperature manipulation after cardiac arrest in rats. *Crit Care Med*. (2008) 36:1909. doi: 10.1097/CCM.0b013e3181760eb5
38. Hirsch LJ, LaRoche SM, Gaspard N, Gerard E, Svoronos A, Herman ST, et al. American Clinical Neurophysiology Society's standardized critical care EEG terminology: 2012 version. *J Clin Neurophysiol*. (2013) 30:1–27. doi: 10.1097/WNP.0b013e3182784729
39. Schoffelen J, Gross J. Source connectivity analysis with MEG and EEG. *Hum Brain Mapp*. (2009) 30:1857–1865. doi: 10.1002/hbm.20745
40. Zhao Y, Nasrullah Z, Li Z. PyOD: a python toolbox for scalable outlier detection. *J Mach Learn Res*. (2019) 20:1–7.
41. Goldstein M, Dengel A. *Histogram-Based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm*. Saarbrücken: German Research Center for Artificial Intelligence (2012) 59–63.
42. Breunig MM, Kriegel H, Ng RT, Sander J. LOF: identifying density-based local outliers. *SIGMOD Rec*. (2000) 29:93–104. doi: 10.1145/335191.335388
43. Kriegel H, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, NV) (2008). p. 444–52. doi: 10.1145/1401890.1401946
44. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Comput*. (2001) 13:1443–71. doi: 10.1162/089976601750264965
45. Zhao Y, Hryniewicki MK, Nasrullah Z, Li Z. LSCP: locally selective combination in parallel outlier ensembles.1 In: *SDM* (Calgary, AB) (2019). doi: 10.1137/1.9781611975673.66
46. Aggarwal CC. Outlier analysis. In: *Data Mining: The Textbook*. Cham: Springer International Publishing (2015). p. 75–9. doi: 10.1007/978-3-319-14142-8_8
47. Shyu ML, Chen SC, Sarinnapakorn K, Chang LW. *A Novel Anomaly Detection Scheme Based on Principal Component Classifier*. AD-a465 712. Coral Gables, FL: University of Miami, Department of Electrical and Computer Engineering (2003). Available online at: https://books.google.com/books?id=iXEinQAACAAJ
48. An J, Cho S. *Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability*. (2015).
49. Liu Y, Li Z, Zhou C, Jiang Y, Sun J, Wang M, et al. Generative adversarial active learning for unsupervised outlier detection. In: *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*. (2019). doi: 10.1109/TKDE.2019.2905606
50. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum Brain Mapp*. (2017) 38:5391–420. doi: 10.1002/hbm.23730
51. Halford JJ, Shiau D, Desrochers JA, Kolls BJ, Dean BC, Waters CG, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clin Neurophysiol*. (2015) 126:1661–9. doi: 10.1016/j.clinph.2014.11.008

52. Jung TP, Makeig S, Humphries C, Lee TW, McKeown M, Iragui V, et al. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*. (2000) 37:163–78. doi: 10.1111/1469-8986.3720163

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 Digital Health

ORIGINAL RESEARCH
published: 10 February 2021
doi: 10.3389/fdgth.2021.559103

Machine Learning for Localizing Epileptogenic-Zone in the Temporal Lobe: Quantifying the Value of Multimodal Clinical-Semiology and Imaging Concordance

Ali Alim-Marvasti[1,2,3,4]*, *Fernando Pérez-García*[2,3,5], *Karan Dahele*[6], *Gloria Romagnoli*[1,4], *Beate Diehl*[1,4], *Rachel Sparks*[5], *Sebastien Ourselin*[5], *Matthew J. Clarkson*[2,3] and *John S. Duncan*[1,4]

[1] Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom, [2] Department of Medical Physics and Biomedical Engineering, University College London, London, United Kingdom, [3] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), London, United Kingdom, [4] National Hospital for Neurology and Neurosurgery, London, United Kingdom, [5] School of Biomedical Engineering & Imaging Sciences (BMEIS), King's College London, London, United Kingdom, [6] University College London Medical School, London, United Kingdom

OPEN ACCESS

Edited by:
Ira L. Leeds,
Johns Hopkins University,
United States

Reviewed by:
Gregory Scott,
Imperial College London,
United Kingdom
Tyler John Loftus,
University of Florida, United States

***Correspondence:**
Ali Alim-Marvasti
a.alim-marvasti@ucl.ac.uk;
alijesus.alim-marvasti@nhs.net

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

Received: 05 May 2020
Accepted: 21 January 2021
Published: 10 February 2021

Citation:
Alim-Marvasti A, Pérez-García F, Dahele K, Romagnoli G, Diehl B, Sparks R, Ourselin S, Clarkson MJ and Duncan JS (2021) Machine Learning for Localizing Epileptogenic-Zone in the Temporal Lobe: Quantifying the Value of Multimodal Clinical-Semiology and Imaging Concordance. Front. Digit. Health 3:559103. doi: 10.3389/fdgth.2021.559103

Background: Epilepsy affects 50 million people worldwide and a third are refractory to medication. If a discrete cerebral focus or network can be identified, neurosurgical resection can be curative. Most excisions are in the temporal-lobe, and are more likely to result in seizure-freedom than extra-temporal resections. However, less than half of patients undergoing surgery become entirely seizure-free. Localizing the epileptogenic-zone and individualized outcome predictions are difficult, requiring detailed evaluations at specialist centers.

Methods: We used bespoke natural language processing to text-mine 3,800 electronic health records, from 309 epilepsy surgery patients, evaluated over a decade, of whom 126 remained entirely seizure-free. We investigated the diagnostic performances of machine learning models using set-of-semiology (SoS) with and without hippocampal sclerosis (HS) on MRI as features, using STARD criteria.

Findings: Support Vector Classifiers (SVC) and Gradient Boosted (GB) decision trees were the best performing algorithms for temporal-lobe epileptogenic zone localization (cross-validated Matthews correlation coefficient (MCC) SVC 0.73 ± 0.25, balanced accuracy 0.81 ± 0.14, AUC 0.95 ± 0.05). Models that only used seizure semiology were not always better than internal benchmarks. The combination of multimodal features, however, enhanced performance metrics including MCC and normalized mutual information (NMI) compared to either alone ($p < 0.0001$). This combination of semiology and HS on MRI increased both cross-validated MCC and NMI by over 25% (NMI, SVC SoS: 0.35 ± 0.28 vs. SVC SoS+HS: 0.61 ± 0.27).

Interpretation: Machine learning models using only the set of seizure semiology (SoS) cannot unequivocally perform better than benchmarks in temporal epileptogenic-zone localization. However, the combination of SoS with an imaging feature (HS)

enhance epileptogenic lobe localization. We quantified this added NMI value to be 25% in absolute terms. Despite good performance in localization, no model was able to predict seizure-freedom better than benchmarks. The methods used are widely applicable, and the performance enhancements by combining other clinical, imaging and neurophysiological features could be similarly quantified. Multicenter studies are required to confirm generalizability.

Funding: Wellcome/EPSRC Center for Interventional and Surgical Sciences (WEISS) (203145Z/16/Z).

Keywords: epilepsy surgery, machine learning, semiology, hippocampal sclerosis, epileptogenic zone, temporal lobe epilepsy, gradient boost classifier, linear support vector classifier

INTRODUCTION

Fifty million people have epilepsy world-wide, and one third are refractory to two or more appropriate antiepileptic drugs, with recurrent seizures and impairment of quality of life. Neurosurgical resections in focal epilepsy may be curative and have been shown to improve health status (1–3). The Epileptogenic Zone (EZ) is defined as the region that when resected, renders the patient seizure-free. Understanding the symptoms, signs and semiology (chronological clinical seizure manifestations) at the onset of seizures is key to determining the site of seizure onset in the brain; but this may be imprecise (4). Despite an extensive literature on semiology, imaging and electroencephalographic (EEG) features for EZ-localization, no definitive method exists to determine the EZ (5). Concordance is sought with brain imaging: MRI, functional imaging (SPECT, FDG-PET); scalp EEG video-telemetry and neuropsychology. The results are discussed in a multidisciplinary team (MDT) conference, to localize the EZ and minimize risks, prior to consideration of resection. Despite this, many patients do not become seizure-free after surgery (6).

The value of any particular clinical feature or investigation result in contributing to a patient's differential diagnosis depends on its overall univariate association with the EZ (prior) and any other factors which may interact with it. Clinical judgement and acumen arise through experience, when there may not be objective data. Although one can assess the value of clinical features through Bayesian-belief elicitation, in the absence of grounded-objectives, responses would be capturing subjective clinical values (7). Well-designed machine learning methods using ground-truth target labels and all relevant features perform well in capturing data patterns to predict targets, akin to clinical intuition. The so-called "AI chasm" notes that algorithms are only clinically useful if they improve clinical outcomes, not just diagnostic accuracy (8).

A study in 2015 evaluated 830 patients and the value of semiology in predicting the EZ (9). Conditional inference trees' localization accuracy among five ictal onset areas was 56.1%. Accuracy for binary mesial temporal lobe epilepsy (mTLE) or lateral temporal-EZ was 71% (unquoted naïve accuracy of 63%) (9). Despite the large numbers, the supervised learning method suffered from inadequate ground-truth labels: the EZ was often labeled by clinicians on the presence or absence of a particular semiology, making the evaluation logic circular and results were reported without cross-validation or test sets, compromising generalizability. A review in 2017 showed algorithmic identification of EZ brain networks and the propagation of seizures remains an open issue. Combinations of multimodal features have not been used on large-scale high-quality patient data (10). Currently there are no clinically utilized algorithms to augment EZ-localization or quantify the value of multimodal features presented in MDTs.

In this study, we set out to objectively assess the value of combining clinical features for temporal-lobe (TL) epileptogenic zone localization – the most common form of drug refractory epilepsy with the best surgical outcomes. We investigated set of seizure semiology (SoS, devoid of sequence information) and hippocampal sclerosis (HS), as this imaging finding is specific to the TL, is the most frequent imaging finding, and provides a good univariate benchmark. HS is a scar in the medial temporal lobe and the most common pathology underlying drug-resistant TL epilepsy. These features are important in clinical evaluations and can be extracted from electronic health record texts. We used machine learning models with strong ground-truths and also assessed values in predicting surgical outcomes.

METHODS

Study Design and Participants

Our objective was to determine the value of clinical-semiology, hippocampal sclerosis and their combination for the binary localization of the EZ to the temporal or extratemporal brain. The value of combining these features was quantified for both relative diagnostic performance (Step 1) and subsequently using the model from Step 1 for post-surgical prognosis (Step 2) as well as training independent models for the direct prediction of surgical outcomes (Step 3).

Retrospective text analysis of 3,800 mixed data-type electronic health records (EHRs) pertaining to adults with refractory focal epilepsy admitted for presurgical assessment for epilepsy surgery from 2001 to 2011 was undertaken at the National Hospital for

TABLE 1 | Frequency of Features and Targets.

Variable	Frequency in seizure-free patients ($n = 126$) (%)	Frequency in all operated patients ($n = 309$) (%)
Temporal-EZ (target)	112 (89%)	256 (mix of seizure-free and not seizure-free) (83%)
Dialeptic/loss of awareness (LOA)	92 (73%)	223 (72%)
Tonic-clonic	84 (67%)	224 (72%)
Hippocampal sclerosis (imaging feature)	70 (56%)	147 (48%)
Oral automatisms	58 (46%)	140 (45%)
Other automatism (unspecified)	57 (45%)	138 (45%)
Olfactory-gustatory	56 (44%)	141 (46%)
Upper limb automatism	49 (39%)	108 (35%)
Tonic	47 (37%)	126 (41%)
Aphasia	46 (37%)	100 (32%)
Fear-Anxiety	37 (29%)	91 (29%)
Head Turn	30 (24%)	73 (24%)
Clonic	30 (24%)	77 (25%)
Epigastric	28 (22%)	61 (20%)
Autonomous-vegetative	26 (21%)	66 (21%)
Psychic	23 (18%)	57 (18%)
Non-specific aura	22 (17%)	52 (17%)
Dysphasia	21 (17%)	71 (23%)
LOC	17 (13%)	46 (15%)
Astatic	15 (12%)	38 (12%)
Other simple motor	14 (11%)	32 (10%)
Vocalization	13 (10%)	33 (11%)
Somatosensory	12 (10%)	39 (13%)
Nose-wiping	10 (8%)	18 (6%)
Dystonic	10 (8%)	26 (8%)
Head version	10 (8%)	27 (9%)
Grimace	10 (8%)	19 (6%)
Blink	9 (7%)	27 (9%)
Hypermotor	8 (6%)	19 (6%)
Dacrystic	8 (6%)	14 (5%)
Vestibular	7 (6%)	26 (8%)
Other complex motor	6 (5%)	13 (4%)
Auditory	4 (3%)	10 (3%)
Gelastic	4 (3%)	7 (2%)
Eye Version	3 (2%)	8 (3%)
Hypomotor (behavioral arrest)	3 (2%)	11 (4%)
Visual	3 (2%)	12 (4%)
Coprolalia	3 (2%)	3 (1%)
Figure of 4	2 (2%)	5 (2%)
Atonic	2 (2%)	6 (2%)
Ictal pout	1 (1%)	1 (0.3%)
Myoclonic	1 (1%)	2 (1%)
Spitting	1 (1%)	7 (2%)
Asymmetric tonic	1 (1%)	4 (1%)
Fencing	0	1 (0.3%)
Lower limb automatism	0	1 (0.3%)
Palilalia	0	0
Aphemia	0	0
Drinking	0	0
Cough	0	0
Whistling	0	0

Frequency of patients with Semiology, imaging feature and temporal resections. By "hypomotor" we mean behavioral arrest during a seizure and not the semiology specific to the pediatric population.

Neurology and Neurosurgery, London. SoS, HS, and temporal-EZ features were extracted (**Table 1**). Univariate statistics were computed and machine learning models were trained to predict temporal-EZ and subsequently prognosis.

We used set-of-semiology (SoS), because these are more readily available from a clinical history than precise symptom chronology. We restricted MRI-identifiable TL pathology to HS as this represented 92% of temporal lesions ($n = 70$).

Procedures

EHRs were pseudo-anonymised, pre-processed and text-mined for the presence of 49 semiology features and a single imaging feature (HS) using regular expressions as a taxonomy replacement. This taxonomy replacement was a bespoke expansion of major semiological categories presented elsewhere (4). The anonymised keys and identifiers were stored in secure NHS systems and checks for data-mining integrity on a subsample showed <5% binary-feature error compared to manual feature-extraction by a consultant neurologist. The Pandas DataFrame was sparse and multi-one-hot encoded. EHRs were cross-referenced to a database containing EZ-localization labels (resected lobes) alongside their post-operative year-by-year ordinal score on the ILAE epilepsy surgery outcome scale, and whether they had intracranial electrode recordings, curated since 1990, as previously reported (6). Intracranial electrodes were collected only as a univariate benchmark for negative prognostic value in epilepsy surgery, as their presence is a clinical indicator of uncertain EZ.

EHRs from 870 cases were available, 335 of which underwent epilepsy-surgery after assessment. 324 cases were from unique patients, of which 309 had one resection only, excluding hemispherectomies and corpus callosotomies, consistent with previous methodology (11).

Statistical Analysis

Fisher's exact and Mann-Whitney *U*-tests were performed at three levels of uncorrected type I error (alpha = 0.05, 0.005, and 0.0005) with Bonferroni corrections for multiple comparisons for 181 tests (Fisher's: 51 for Step 1, 53 × 2 for Step 3; MWU: 24 tests)

($p < 2.76 \times 10^{-4} = *$, $p < 2.76 \times 10^{-5} = **$, $p < 2.76 \times 10^{-6} = ***$, respectively). Theil's U (asymmetric normalized mutual information, NMI) was used to check for categorical correlations and model performance.

Machine Learning

We used multivariate binary Logistic Regression (LR), Gradient Boosted Trees (GB), and Linear Support Vector Classifiers (SVC) (implemented in Scikit-learn v 0.19.2) (12) as suggested by previous studies (9, 13). We chose these specific algorithms as LR is widely used in predictive models, SVC performs well if the target can be linearly separated by a high-dimensional hyperplane in feature space, and GB ensemble models leverage multiple weak classifiers into a strong classifier with each individual component utilizing a different feature subset, akin to clinical MDTs. GB are more likely to succeed with more data and complexity, but are less interpretable than SVC or LR. For binary features and binary outcomes as in our study, LR without regularization can have a decision boundary that asymptotically approaches that of SVC (14), which can further help assess if the targets are linearly separable. Feature selection was performed using both univariate and recursive feature elimination with 5-fold cross-validation (RFECV) methods (15). No other hyperparameter tuning was performed.

The models were compared to benchmarks in localizing temporal-EZ (Step 1). We also made indirect assessments if improved diagnostic accuracy translated to enhanced outcome predictions (Step 2), and separately trained models to directly predict outcomes (Step 3). For Step 1, we chose a binary localization target containing the most common focal epilepsy, temporal-lobe vs. extra-temporal (ET) EZ, and models were trained on patients who were entirely seizure-free at all follow-up years (ESF). For Steps 2 and 3, outcome was assessed at two binary levels: seizure-freedom at 1-year (ILAE1), and ESF. In Step 2, the Step 1 model was used to predict outcomes on all data. In Step 3, new models were trained to predict outcomes. ILAE 2 and above were considered not seizure-free (NSF) due to residual epileptogenic tissue resulting in auras or seizures with impaired awareness.

Although we report many metrics (using 1,000 × 5 repeated stratified CV with means and standard deviations in **Table 3**, or medians and IQR), due to an unbalanced dataset, we focus on Matthews-correlation-coefficient (MCC) as one of the most suitable metrics for binary classification evaluations which can be interpreted as a discretization of Pearson's-correlation-coefficient (16, 17). NMI was used to quantify information gains between features, models, and the ground truth EZ.

Role of the Funding Source

The Wellcome/EPSRC Center for Interventional and Surgical Sciences had no role in the study design; collection, analysis or interpretation of data; writing of report; nor in the decision to submit for publication.

This study was approved by the Research Ethics Committee for UCL and UCLH (20/LO/0149).

TABLE 2 | Benchmarks for Step 1 Temporal-EZ Localization.

Feature	Number with TL-EZ/number with feature ($n = 126$)	Number with TL-EZ/number with feature ($n = 309$)	Odds ratios ($n = 126$, $n = 309$)	p-values ($n = 126$, $n = 309$)
Temporal-EZ features				
Hippocampal sclerosis	70/70	144/147	DBZ**, 21***	4.2×10^{-6}**, 6.3×10^{-13}***
All Automatisms (combined)	82/84	186/206	16.4*, 4.4***	3.0×10^{-5}*, 2.2×10^{-6}***
Oral automatisms	58/58	131/140	DBZ*, 5.1**	9.7×10^{-5}*, 3.5×10^{-6}**
Other automatism (unspecified)	55/57	127/138	5.8, 3.8*	0.020, 0.00012*
Upper limb automatism	49/49	100/108	DBZ, 3.6	0.00082, 0.00077
Fear-anxiety	37/37	84/91	DBZ, 3.2	0.010, 0.0045
Dialeptic/LOA	85/92	195/223	3.1, 2.9	0.054, 0.0012
Epigastric	NS	58/61	NS, 4.9	NS, 0.0039
Aphasia	NS	90/100	NS, 2.3	NS, 0.024
Extratemporal-EZ features				
Intracranial electrodes	NS	50/89	NS, 0.09	NS, 7.1×10^{-4}
Hypomotor (behavioral arrest)	0/3	6/11	0, 0.16	0.0011, 0.0045
Somatosensory	8/12	25/39	0.19, 0.30	0.029, 0.0024
Clonic	23/30	57/77	0.26, 0.47	0.040, 0.023
Head version	NS	16/27	NS, 0.25	NS, 0.0021
Eye version	NS	3/8	NS, 0.11	NS, 0.0046
Asymmetric tonic	NS	1/4	NS, 0.07	NS, 0.017

*Fisher's exact test for Step 1 Temporal-EZ localization in postoperative seizure-free patients (n = 126, strong ground truths) and all operated patients (n = 309, 256 weakly labeled as temporal, 53 as extratemporal). All features with p < 0.05 are shown; *Represents significance at alpha 5% after Bonferroni correction. **at 0.5% after Bonferroni correction. ***at 0.05% after Bonferroni correction. DBZ, Division By Zero. NS: p > 0.05.*

TABLE 3 | Machine Learning Models for Temporal EZ-Localization (Step 1).

Model-RFECV / 5-CV metric +/-std (refit)	Naïve benchmark	Automotor semiology univariate benchmark	HS imaging univariate benchmark	LR SoS	LR SoS+HS	Linear support vector classifier SoS	Linear support vector classifier SoS+HS	GB SoS	GB SoS+HS
# of features (min equivalent)	N/A	1	1	16	25 (18)	40 (30)	9	27	17
F1 average macro	N/A	0.61 ± 0.06	0.59 ± 0.06	0.68 ± 0.17 (0.88)	0.75 ± 0.16 (0.88)	0.72 ± 0.16 (0.88)	0.85 ± 0.14 (0.91)	0.66 ± 0.15	0.81 ± 0.14 (0.98)
Balanced accuracy	0.5	0.67 ± 0.07	0.75 ± 0.04	0.65 ± 0.13 (0.82)	0.72 ± 0.15 (0.82)	0.70 ± 0.15 (0.82)	0.81 ± 0.14 (0.86)	0.65 ± 0.14	0.80 ± 0.15 (0.96)
Accuracy	0.83 ± 0.04	0.71 ± 0.05	0.63 ± 0.05	0.92 ± 0.03 (0.96)	0.93 ± 0.03 (0.96)	0.92 ± 0.04 (0.96)	0.96 ± 0.03 (0.97)	0.89 ± 0.05	0.93 ± 0.04 (0.99)
Sensitivity/recall	1	0.73 ± 0.06	0.56 ± 0.06	1.0 ± 0.004	0.995 ± 0.015	0.98 ± 0.03	1.0 ± 0.006 (1.0)	0.96 ± 0.04	0.97 ± 0.04 (1.0)
Specificity	0	0.62 ± 0.14	0.94 ± 0.06	0.30 ± 0.26 (0.64)	0.44 ± 0.29 (0.64)	0.42 ± 0.29 (0.64)	0.61 ± 0.28 (0.71)	0.35 ± 0.27	0.62 ± 0.29 (0.93)
PPV	0.83 ± 0.04	0.90 ± 0.04	0.98 ± 0.02	0.92 ± 0.03 (0.96)	0.94 ± 0.03 (0.96)	0.93 ± 0.03 (0.96)	0.95 ± 0.03 (0.97)	0.92 ± 0.03	0.95 ± 0.03 (1.0)
NPV	0	0.32 ± 0.09	0.31 ± 0.07	0.64 ± 0.48 (1.0)	0.77 ± 0.39 (1.0)	0.67 ± 0.40 (1.0)	0.93 ± 0.25 (1.0)	0.51 ± 0.39	0.76 ± 0.31 (1.0)
AUROC	N/A	N/A	N/A	0.89 ± 0.11	0.95 ± 0.06	0.83 ± 0.14	0.95 ± 0.05	0.81 ± 0.14	0.95 ± 0.07
Average Precision	N/A	N/A	N/A	0.98 ± 0.02	0.99 ± 0.01	0.97 ± 0.03	0.99 ± 0.01	0.97 ± 0.03	0.99 ± 0.01
MCC [bootstrap refit]	0	[0.28± 0.12]	[0.38± 0.08]	0.41 ± 0.33 [0.76 ± 0.22] (0.78)	0.55 ± 0.31 [0.76 ± 0.22] (0.78)	0.48 ± 0.32 [0.76 ± 0.22] (0.78)	0.73 ± 0.25 [0.81 ± 0.19] (0.83)	0.36 ± 0.30	0.64 ± 0.27 [0.96 ± 0.09] (0.96)
NMI symmetric [asymmetric bootstrap refit]	0	[0.10± 0.07]	[0.21± 0.08] (0.28)	0.31 ± 0.26 [0.57 ± 0.29] (0.53)	0.42 ± 0.28 [0.57 ± 0.29] (0.53)	0.35 ± 0.28 [0.57 ± 0.29] (0.53)	0.61 ± 0.27 [0.65 ± 0.29] (0.604)	0.23 ± 0.23	0.48 ± 0.29 [0.91 ± 0.19] (0.87)

Step 1 CV performance metrics. Mean and standard deviation of 1,000 × 5 CV scores. Benchmark std given by bootstrapping 2,000 × 5 CV. Brackets represent model-refit (training) scores. Square brackets show bootstrapped refit results. CV, cross-validation; RFECV, Recursive Feature Elimination with CV; std, standard deviation; PPV/NPV, Positive/Negative Predictive Value; AUROC, Area under receiver operating curve; MCC, Matthews Correlation Coefficient; NMI, Normalized Mutual Information. See ***Supplementary Materials*** *for expanded table and distribution of MCC and NMI scores.*

RESULTS

Patients and Outcomes

Of the 309 patients, 126 (41%) were ESF at all follow-up years (median follow-up 7 years, IQR = 5–10, **Supplementary Figure 9**), indicating correct EZ-resections. Labels were unbalanced; 112/126 (88.9%) were temporal-EZ, and 14 extratemporal.

Features

Forty-two semiology features were present in the ESF-set. Automatisms (oral, manual and other) were merged to a single category, leaving 40 SoS features. There were 76 temporal-lobe lesions in the ESF group and HS as the single imaging feature constituted 92% (70/76) of these. In addition, there were three cavernomas, one dysembryoplastic neuroepithelial tumor, one cyst and one focal cortical dysplasia in the temporal lobes.

Table 1 shows frequency of occurrences in the 126-ESF-set and all 309 operated patients.

Table 2 shows univariate benchmarks for features associated with temporal-EZ. The statistically significant features after multiple-comparisons correction on two-by-two Fisher's exact tests were seizures with automatisms and HS. The highest odds-ratios were for presence of HS, automatisms, and fear-anxiety. The performance metrics of the best univariate features, as benchmarks, are summarized in **Table 3**.

Step 1: EZ Cross-Validated Results

The learning curves for the GB and SVC models show overfitting for SoS features alone that improved with combined SoS+HS features (**Figure 1**). **Table 3** shows semiology and imaging enhanced performance above that of benchmarks using the best features obtained from RFECV (**Figures 2**, **3**), most of which were found in the univariate analysis (**Table 2**). **Figure 4** shows that combined features also enhance training-set performance.

GB betters SVC when refit to the ESF-set (**Figure 4**); whereas cross-validated results (**Figure 1**, **Table 3**) show the models perform more similarly: mean and median MCC with and without the imaging feature are:

- Best benchmark (imaging-HS): mean = 0.38 ± 0.08, median = 0.38, IQR = 0.33–0.43
- GB-SoS: mean = 0.36 ± 0.30, median = 0.35, IQR = 0.0–0.55
- GB-SoS+HS: mean = 0.64 ± 0.27, median = 0.66, IQR = 0.55–0.80
- SVC-SoS: mean = 0.48 ± 0.32, median = 0.55, IQR = 0.34–0.69
- SVC-SoS+HS: mean = 0.73 ± 0.25, median = 0.80, IQR = 0.55–0.80.

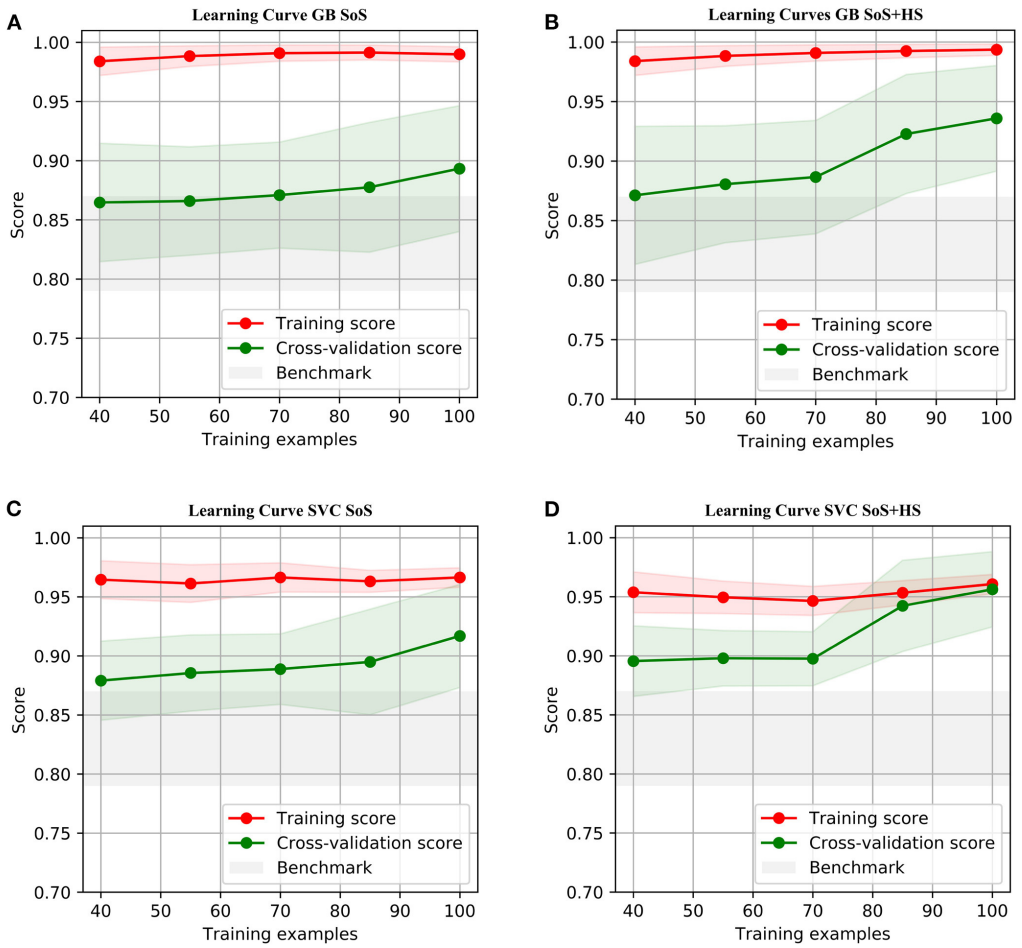


FIGURE 1 | Learning Curves using accuracy score, with standard deviations. The test-fold accuracies (in green) are more representative of model performances on prospective data, showing enhanced learning by combining semiology and HS. **(A,C)** SoS has limited test-fold learning (green) with increasing training samples. **(B,D)** SoS+HS improves test-fold accuracies after about 70 samples. See **Supplementary Materials** for comparison with logistic regression.

Comparing GB and SVC-models:

- with semiology alone, although SVC performed better, the two models performed similarly with overlap of interquartile ranges.
- with SoS+HS, there was also significant overlap between the models; the SVC-model again had a better median MCC.

Compared to SoS alone, when combining features:

- SVC mean, median, lower and upper quartiles were enhanced by between 10 to 25%. This suggests the support vectors are better defined with HS and that temporal lobe EZ are linearly separable in binary semiology-HS feature space.
- in the GB-model, there was also significant improvements in lower-quartile (55%), median (30%) and upper-quartile (25%) MCC and no overlap in interquartile ranges.
- LR (**Table 3**) shows similar improvements in metrics, except the median MCC remains at 0.55.

These affirm the value of combining multimodal features, irrespective of the model.

Step 2: Indirect Surgical Outcome Results

Of the 183 NSF patients, 144 had temporal resections (54 ILAE 1 at 1-year, median of patient ILAE outcome medians = 2, IQR = 1–4) and 39 extratemporal resections (seven ILAE 1 at 1-year, median = 4, IQR = 2–4). Temporal resections were associated with better outcomes at 1-year post-resection (ILAE 1, OR = 2.7, $p = 0.035$) and better median ILAE outcomes (Mann-Whitney $U = 2{,}057$, $p = 0.004$). None of the machine learning models' congruent predictions with actual resections were significant in improving upon this naïve benchmark (**Supplementary Figures 10–13**).

Step 3: Direct Surgical Outcome Results

Although direct ($n = 309$) benchmarks for ESF included having had a temporal-resection (OR = 2.2, $p = 0.02$), having been seizure-free-at-1-year, presence of HS (OR = 1.7, $p = 0.02$),

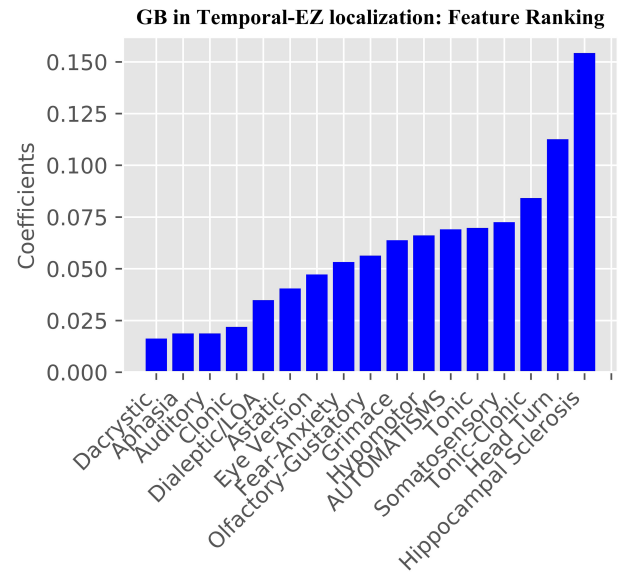


FIGURE 2 | Gradient Boosting Classifier GB SoS+HS Feature Importance. From the 41 combined features, RFECV was used to determine the most relevant features for the model.

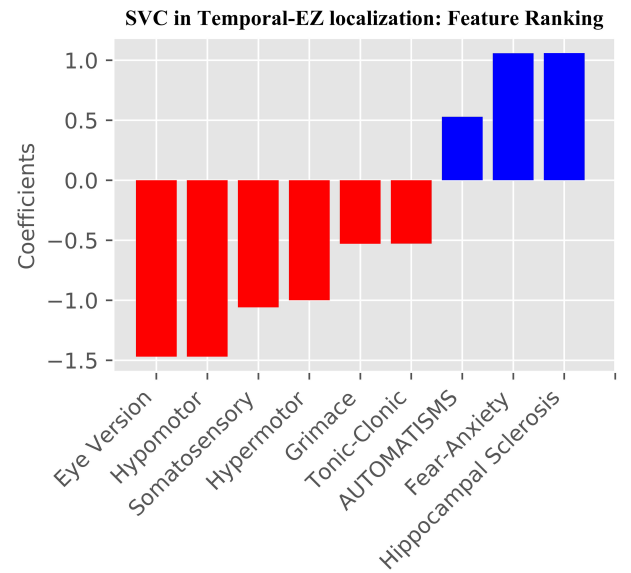


FIGURE 3 | Support Vector Classifier SVC SoS+HS feature ranking using RFECV. In blue are features which predict temporal, and in red extratemporal EZ. All SVC features are also used by the GB model, except "Hypermotor" semiology.

and dysphasia (OR = 0.53, $p = 0.039$), and benchmarks for predicting seizure-freedom at 1-year included presence of HS (OR = 1.9, RR = 1.29, $p = 0.005$), temporal-lobe-resection (OR = 2.8, $p = 0.001$) and presence of intracranial EEG (OR = 0.46, $p = 0.003$), only seizure-freedom-at-1-year as a predictor of ESF was statistically significant after multiple comparisons correction (Theil's $U = 0.43$). No model was able to exceed naïve or feature benchmarks on any metric.

DISCUSSION

Our main findings were that models localized the epileptogenic-zone to the temporal lobe when using multimodal semiology and MRI report of HS, and were better than semiology, HS or other benchmarks in isolation. Support vector machines had a slight edge over Gradient Boosted trees, but there was considerable overlap in performances (Step 1). No method was able to predict seizure-freedom at 1-year or ESF better than benchmarks (Steps 2 and 3). Multicenter case records are required to confirm generalizability, and expanded features are necessary to determine if epilepsy surgical outcomes can be predicted at all.

EZ-Localization Algorithms (Step 1)

Our study addresses a subset of the open issue of algorithmic identification of EZ networks (10), namely temporal-EZ, and provides univariate and algorithmic benchmarks with single (SoS) or two-modalities (SoS and HS). Models with multimodal features outperform semiology-only models (**Figure 1**) and univariate benchmarks (**Table 3**) using features that are significant on univariate analysis (**Table 2**) and those that are not (**Figures 2**, **3**). The strength of the GB model lies in its ability to combine an ensemble of weak-learners, and out-perform individual univariate benchmarks, including the strongest, HS, as assessed on both training-set (**Figure 4**) and CV-folds (**Table 3**). SVC strength lies in classifying temporal-EZ by defining borderline cases as class-dividing support vectors. Support vectors are the feature-states of the cases which lie at the margins of the optimum hyperplane separating the temporal vs. extratemporal EZs. The SVC-model has 26 support vectors which determine the classifiers hyperplane. Alterations to any of these cases, but not others, can result in a different SVC classifier altogether. This makes the algorithm more robust to slight sample changes during cross-validation. The coefficients in **Figure 3** represent the projections of a vector orthogonal to the classifying hyperplane onto each feature (15).

Clinical Features of Temporal-EZ (Step 1)

The following cardinal semiologies of temporal lobe seizures have been described: (18)

- Prodromes
- Auras
- Altered Consciousness (dialeptic)
- Amnesia
- Automatisms (oral, manual, dacrystic, gelastic, and leaving-behaviors).

Hippocampal sclerosis is present in more than 80% of surgically treated TLE. The published semiologies in mTLE, commonly associated with HS include:

- Rising epigastric sensation

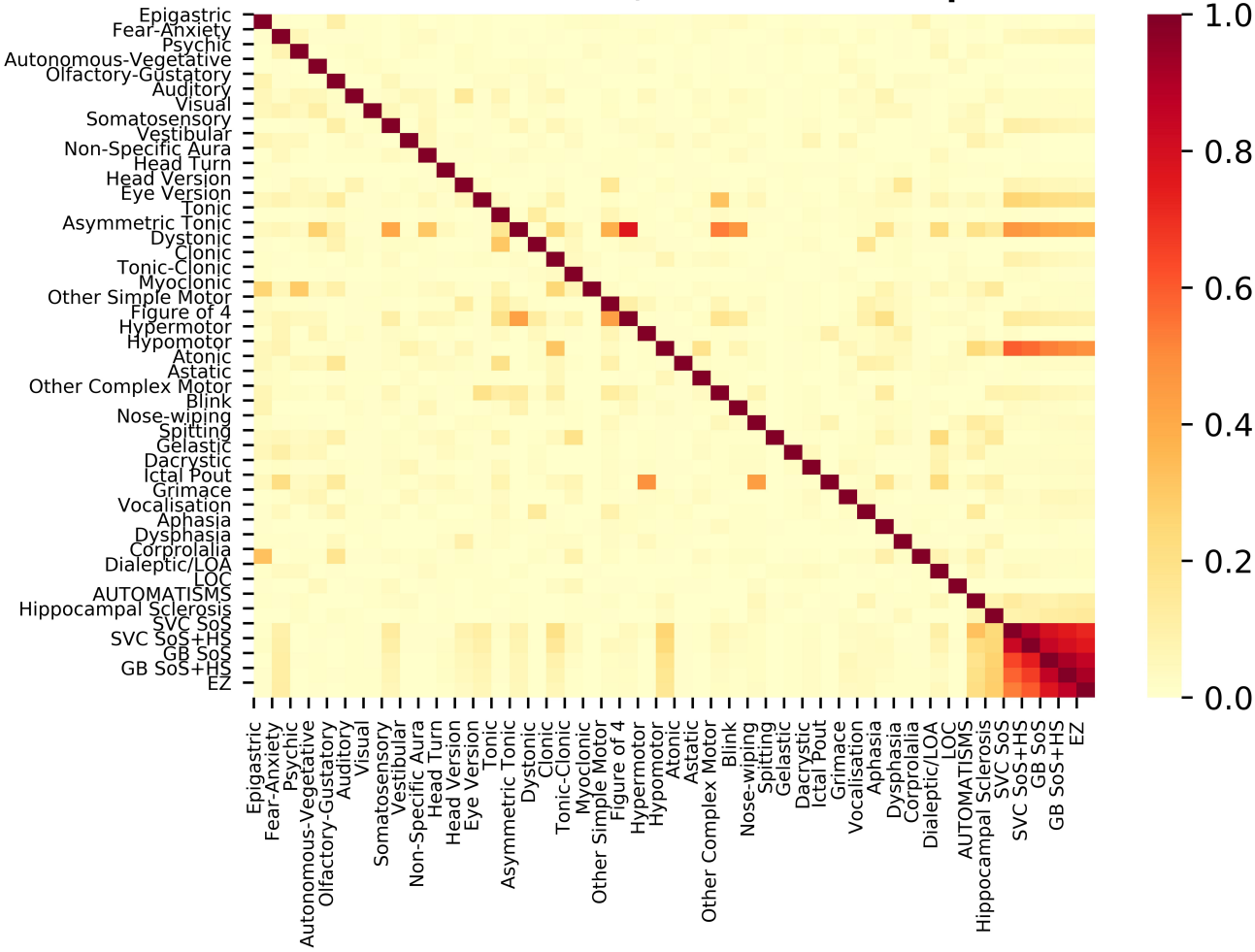


FIGURE 4 | Theil's U for features, models and temporal-EZ localization. The columns represent the known (Bayesian prior) variable, and the rows the target entropy coefficient. For example, "Figure of 4" semiology in the column is a subset of "Asymmetric Tonic" in the row with high coefficient, whereas the reverse association is smaller in magnitude. The naïve algorithm would show zero association with all the variables with a column of zeros (and is undefined in row due to division by zero). All four models are more predictive of a temporal-EZ than any of the univariate features, with 1 (intense red) representing 100% of the information in the target being predicted by the column. With the addition of Imaging HS to Semiology SoS ($n = 126$, refit to training set) SVC and GB show a graded improvement in the proportion EZ-localization entropy accounted for.

- Affective (fear)
- Experiential (including déjà vu)
- Automatisms
- Head Turns
- Autonomic phenomenon.

These semiologies are confirmed by univariate analysis (**Table 2**), and from the 17 retained features post-RFECV (**Figure 2**). A notable exception is rising epigastric sensation. Epigastric sensation is non-significant for the ESF patients used to train the data (**Table 2**) and not present as a feature after RFECV for either the SVC or GB models (**Figures 2**, **3**).

There are conflicts and overconfidence in reporting the localizing values of semiology in the literature, using small samples of clinical cases and often no ground-truths to objectively assess labels or effects on surgical outcomes. The localizing values of semiologies may be stated without measuring confidence or variation e.g., postictal cough localizing to the temporal lobe (18), unilateral upper-limb automatisms reported to both have an ipsilateral seizure onset (19, 20) and no lateralizing value in isolation (21). Such discrepancies may arise due to lack of ground-truths, small numbers, ignoring time to onset of the semiology and excluding relevant features. When value is assessed, this is usually performed in a univariate manner, e.g., in one example series the trend that hypermotor seizures occur earlier in frontal lobe epilepsy than extra-frontal epilepsies was assessed by univariate Fisher's exact test, showing that chronology is valuable for EZ-localization; but did not reach significance and only 17 surgical patients were seizure-free (ground-truth labels), limiting the power of the analysis (22). The GB algorithm (**Figure 2**) shares all the SVC-model features

(**Figure 3**) except hypermotor, which only features in the SVC-model, potentially making the SVC model more capable of identifying frontal-lobe (extratemporal) seizures.

Quantifying Value of Multimodal Features

Although studies that look at single modality data can quantify the value of semiology compared to naïve benchmarks, they cannot assess the value of multimodal features, as are utilized clinically in MDTs (9). Clinical, demographic, imaging and neurophysiological features applied in machine learning have been purported to be capable of predicting mTLE outcomes (with or without HS), but this value has not been quantified nor applied to EZ-localization (13). Multimodal features of EEG and semiology enhance EZ-lateralisation accuracy (23), and although it is known that integration of clinical data also enhance EZ-localization (20), datamining studies have not quantified the incremental value of multimodal data (13).

Different methods may be used to assess incremental multimodal value; for any given model, the convergence rate of the learning curve, choice of performance metric, and maximum or average performance. We highlighted the value of semiology and imaging using all of these methods, and used suitable summary metrics in unbalanced datasets, MCC and NMI (**Table 3**). In both the GB SoS+HS and SVC SoS+HS models, multimodal features improve MCC and NMI average scores by over 25% compared to the best univariate benchmark of HS, and compared to the SoS-only models. Therefore, although SoS is not more valuable than univariate markers, when combined with the imaging feature (HS) it enhances epileptogenic lobe localization.

Outcome Prediction (Steps 2 and 3)

In Step 2 we evaluated model performance in indirectly predicting outcomes on the 183 non-seizure free patients. We assessed the veracity of these EZ-labels using the model as the predictor of true labels. The null hypothesis was that if there was a mismatch between the actual resection (weakly labelled EZ) and prediction, the ILAE outcomes should not be significantly different to when there is congruence of prediction. A naïve benchmark which predicts all resections to be temporal outperforms models from Step 1, therefore the EZ-localization performance does not translate to better outcomes.

Step 3 directly used all 309 patients' features to predict seizure-freedom, and the training curves showed overfitting as the models performed much better on the training set, but were no better than benchmarks on cross-validation folds (**Supplementary Figure 12**). Features which could localize temporal-EZ within the context of the above algorithms are thus insufficient for outcome prediction, which limits their clinical utility (8). Many other factors besides the EZ may determine outcomes, including whether there are indicators of multifocal epilepsy, unaccounted clinical (24) and genetic features, lesion histology (25), EEG patterns, and extent of surgical resection (11, 26–29). Our model did not account for these, nor the precise structures within the temporal lobe that were resected.

Table 2 suggests that invasive EEG is more likely to be used in extra-temporal-EZ, but is not associated with better outcomes, reflecting selection bias, in that invasive EEG would only be used if localization was unclear on non-invasive investigations.

We were not able to predict outcomes with our chosen features using GB, SVC, or other models, as reported previously (30). However, other studies have purported to be capable of predicting mTLE binary post-surgical outcomes using various models and features in cross-validated studies: naïve-Bayes and SVC (max accuracy 95%) (13), neural networks and wide manual data abstraction (accuracy 92%); neural networks and diffusion-tensor imaging (PPV of 88 $\pm$ 7%) (31, 32). The smaller studies are likely to be overfitting the data and not generalizable, and even accurate prognostication does not help improve clinical outcomes (33).

Limitations

The mean CV score is considered an unbiased estimate of performance. The standard deviation estimates for the CV scores are however not unbiased (34); these are particularly large due to different training samples within each fold (e.g., SVC is sensitive to the support vector cases), and some folds predicting no extratemporal EZs due to class imbalance, resulting in larger variances for NPV and specificity (**Table 3**). As we tuned the number of features using RFECV, the mean CV score is also biased, therefore multicenter prospective data is required to assess generalizability and ascertain which model is inherently more suited to localizing temporal-EZ. The learning curves also suggest further data may enhance results.

We used the complete set of available ictal symptoms and not only the semiology presenting at seizure-onset or a sequential Markov model, which together with omitted imaging, electrophysiological and neurophysiological features may yield better results.

We did not model propagation networks in which similarly located lesions may differentially straddle inherent brain networks. Dichotomous assumption of temporal vs. extra-temporal lobe epilepsy may be only good insofar as the majority of resections are anterior temporal resections. Our labels do not differentiate between lateral or mesial temporal-lobe EZ or indeed the extent of resection.

The PPV and specificity of both semiology and HS are higher than the models in predicting temporal-EZ, although the training-scores are comparable. The GB SoS+HS model has a more balanced metric profile, as reflected in F1-macro, MCC and NMI scores (**Table 3**).

A strength of our study is the inclusion of only patients who remained ESF for epileptogenic zone localization, despite the good results for localization, this doesn't translate to better outcomes, the so-called AI chasm is thus not surmounted.

Further work is required to validate this localization model prospectively. Expanding the number of training samples and features in a multicenter approach may allow the use of these models to localize epileptogenic networks to a greater level of detail, and allow investigation of the extent that surgical outcomes can or cannot be predicted with all available multimodal data.

DATA AVAILABILITY STATEMENT

Due to patient confidentiality, the datasets are not publicly available, but anonymised versions can be made available upon reasonable request.

ETHICS STATEMENT

This study was approved by the Research Ethics Committee for UCL and UCLH (20/LO/0149).

AUTHOR CONTRIBUTIONS

AA-M designed the study, wrote the code for health record pre-processing, data mining and data analysis, performed the statistics, trained machine learning models, made inferences, obtained funding, and wrote the manuscript. FP-G edited the manuscript and checked results. KD trained machine learning models and checked results. GR, SO, and RS edited the manuscript. BD was involved in devising semiology features and edited the manuscript. MC edited the manuscript and supervised the study. JD conceived the research programme, designed the study, obtained funding, edited the manuscript, and supervised the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Wellcome/EPSRC Center for Interventional and Surgical Sciences (WEISS) (203145Z/16/Z).

ACKNOWLEDGMENTS

We would like to thank Jane de Tisi, Prof. Parashkev Nachev, and the Multidisciplinary epilepsy surgery team at NHNN Queen Square since 1990.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2021.559103/full#supplementary-material

REFERENCES

1. Téllez-Zenteno JF, Dhar R, Hernandez-Ronquillo L, Wiebe S. Long-term outcomes in epilepsy surgery: antiepileptic drugs, mortality, cognitive and psychosocial aspects. *Brain.* (2006) 130:334–45. doi: 10.1093/brain/awl316
2. Wiebe S, Blume WT, Girvin JP, Eliasziw M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. *N Engl J Med.* (2001) 345:311–8. doi: 10.1056/NEJM200108023450501
3. Engel J, McDermott MP, Wiebe S, Langfitt JT, Stern JM, Dewar S, et al. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. *JAMA.* (2012) 307:922–30. doi: 10.1001/jama.2012.220
4. Tufenkjian K, Lüders HOJJOCN. Seizure semiology: its value and limitations in localizing the epileptogenic zone. *J Clin Neurol.* (2012) 8:243–50. doi: 10.3988/jcn.2012.8.4.243
5. Luders HO. *Textbook of Epilepsy Surgery*. London, UK: CRC Press (2008). doi: 10.3109/9780203091708
6. De Tisi J, Bell GS, Peacock JL, McEvoy AW, Harkness WF, Sander JW, et al. The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: a cohort study. *Lancet.* (2011) 378:1388–95. doi: 10.1016/S0140-6736(11)60890-8
7. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. *J Clin Epidemiol.* (2010) 63:355–69. doi: 10.1016/j.jclinepi.2009.06.003
8. Topol EJJNM. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* (2019) 25:44. doi: 10.1038/s41591-018-0300-7
9. Kim DW, Jung KY, Chu K, Park SH, Lee SY, Lee SK. Localization value of seizure semiology analyzed by the conditional inference tree method. *Epilepsy Res.* (2015) 115:81–7. doi: 10.1016/j.eplepsyres.2015.05.012
10. Ahmedt-Aristizabal D, Fookes C, Dionisio S, Nguyen K, Cunha JPS, Sridharan S. Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: a focused survey. *Epilepsia.* (2017) 58:1817–31. doi: 10.1111/epi.13907
11. Jeha LE, Najm I, Bingaman W, Dinner D, Widdess-Walsh P, Lüders H. Surgical outcome and prognostic factors of frontal lobe epilepsy surgery. *Brain.* (2007) 130:574–84. doi: 10.1093/brain/awl364
12. 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.
13. Memarian N, Kim S, Dewar S, Engel J Jr, Staba RJ. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Comp Biol Med.* (2015) 64:67–78. doi: 10.1016/j.compbiomed.2015.06.008
14. Alim-Marvasti A. *Converging Support Vector Classifiers and Logistic Regression.* (2020). Available online at: https://towardsdatascience.com/support-vector-classifiers-and-logistic-regression-similarity-97ff06aa6ec3 (accessed January 29, 2021).
15. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* (2002) 46:389–422. doi: 10.1023/A:1012487302797
16. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Protein Struct.* (1975) 405:442–51. doi: 10.1016/0005-2795(75)90109-9
17. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* (2020) 21:6. doi: 10.1186/s12864-019-6413-7
18. Blair RD. Temporal lobe epilepsy semiology. *Epilepsy Res Treat.* (2012) 2012:751510. doi: 10.1155/2012/751510
19. Marks WJ Jr, Laxer KD. Semiology of temporal lobe seizures: value in lateralizing the seizure focus. *Epilepsia.* (1998) 39:721–6. doi: 10.1111/j.1528-1157.1998.tb01157.x
20. So EL. Value and limitations of seizure semiology in localizing seizure onset. *J Clin Neurophysiol.* (2006) 23:353–7. doi: 10.1097/01.wnp.0000228498.71365.7b
21. Bleasel A, Kotagal P, Kankirawatana P, Rybicki L. Lateralizing value and semiology of ictal limb posturing and version in temporal lobe and extratemporal epilepsy. *Epilepsia.* (1997) 38:168–74. doi: 10.1111/j.1528-1157.1997.tb01093.x
22. Alqadi K, Sankaraneni R, Thome U, Kotagal P. Semiology of hypermotor (hyperkinetic) seizures. *Epilepsy Behav.* (2016) 54:137–41. doi: 10.1016/j.yebeh.2015.11.017
23. Serles W, Caramanos Z, Lindinger G, Pataraia E, Baumgartner C. Combining ictal surface-electroencephalography and seizure semiology improves patient lateralization in temporal lobe epilepsy. *Epilepsia.* (2000) 41:1567–73. doi: 10.1111/j.1499-1654.2000.001567.x
24. Englot DJ, Lee AT, Tsai C, Halabi C, Barbaro NM, Auguste KI, et al. Seizure types and frequency in patients who "fail"

temporal lobectomy for intractable epilepsy. *Neurosurgery*. (2013) 73:838–44. doi: 10.1227/NEU.0000000000000120
25. Blume WT, Ganapathy GR, Munoz D, Lee DH. Indices of resective surgery effectiveness for intractable nonlesional focal epilepsy. *Epilepsia*. (2004) 45:46–53. doi: 10.1111/j.0013-9580.2004.11203.x
26. Elsharkawy AE, Alabbasi AH, Pannek H, Schulz R, Hoppe M, Pahs G, et al. Outcome of frontal lobe epilepsy surgery in adults. *Epilepsy Res.* (2008) 81:97–106. doi: 10.1016/j.eplepsyres.2008.04.017
27. Dugan P, Carlson C, Jette N, Wiebe S, Bunch M, Kuzniecky R, et al. Derivation and initial validation of a surgical grading scale for the preliminary evaluation of adult patients with drug-resistant focal epilepsy. *Epilepsia.* (2017) 58:792–800. doi: 10.1111/epi.13730
28. Yun CH, Lee SK, Lee SY, Kim KK, Jeong SW, Chung CK. Prognostic factors in neocortical epilepsy surgery: multivariate analysis. *Epilepsia.* (2006) 47:574–9. doi: 10.1111/j.1528-1167.2006.00470.x
29. Lee SK, Lee SY, Kim KK, Hong KS, Lee DS, Chung CK. Surgical outcome and prognostic factors of cryptogenic neocortical epilepsy. *Ann Neurol.* (2005) 58:525–32. doi: 10.1002/ana.20569
30. Goldenholz DM, Jow A, Khan OI, Bagić A, Sato S, Auh S, et al. Preoperative prediction of temporal lobe epilepsy surgery outcome. *Epilepsy research.* (2016) 127:331–8. doi: 10.1016/j.eplepsyres.2016.09.015
31. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Brewster Smith W. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia.* (1998) 39:61–6. doi: 10.1111/j.1528-1157.1998.tb01275.x
32. Gleichgerrcht E, Munsell B, Bhatia S, Vandergrift WA III, Rorden C, McDonald C, et al. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia.* (2018) 59:1643–54. doi: 10.1111/epi.14528
33. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman ML, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg.* (2018) 109:476–86. e1. doi: 10.1016/j.wneu.2017.09.149
34. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. *J Mach Learn Res.* (2004) 5:1089–105.

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 GS declared a past collaboration with one of the authors AA-M.

frontiers in Digital Health

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fdgth.2021.640685

Deep Multi-Modal Transfer Learning for Augmented Patient Acuity Assessment in the Intelligent ICU

Benjamin Shickel[1,2], *Anis Davoudi*[2,3], *Tezcan Ozrazgat-Baslanti*[2,4], *Matthew Ruppert*[2,4], *Azra Bihorac*[2,4] *and Parisa Rashidi*[1,2,3]*

[1] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States, [2] Precision and Intelligent Systems in Medicine (PRISMAP), University of Florida, Gainesville, FL, United States, [3] Department of Biomedical Engineering, University of Florida, Gainesville, FL, United States, [4] Department of Medicine, University of Florida, Gainesville, FL, United States

Accurate prediction and monitoring of patient health in the intensive care unit can inform shared decisions regarding appropriateness of care delivery, risk-reduction strategies, and intensive care resource use. Traditionally, algorithmic solutions for patient outcome prediction rely solely on data available from electronic health records (EHR). In this pilot study, we explore the benefits of augmenting existing EHR data with novel measurements from wrist-worn activity sensors as part of a clinical environment known as the Intelligent ICU. We implemented temporal deep learning models based on two distinct sources of patient data: (1) routinely measured vital signs from electronic health records, and (2) activity data collected from wearable sensors. As a proxy for illness severity, our models predicted whether patients leaving the intensive care unit would be successfully or unsuccessfully discharged from the hospital. We overcome the challenge of small sample size in our prospective cohort by applying deep transfer learning using EHR data from a much larger cohort of traditional ICU patients. Our experiments quantify added utility of non-traditional measurements for predicting patient health, especially when applying a transfer learning procedure to small novel Intelligent ICU cohorts of critically ill patients.

Keywords: machine learning, deep learning, transfer learning, intensive care unit, electronic health records, intelligent ICU

OPEN ACCESS

Edited by:
Ira L. Leeds, Johns Hopkins University, United States

Reviewed by:
*Keng-Hwee Chiam, Bioinformatics Institute (A*STAR), Singapore*
Paraskevi Papadopoulou, American College of Greece, Greece

***Correspondence:**
Parisa Rashidi
parisa.rashidi@ufl.edu

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

Received: *11 December 2020*
Accepted: *02 February 2021*
Published: *22 February 2021*

Citation:
Shickel B, Davoudi A, Ozrazgat-Baslanti T, Ruppert M, Bihorac A and Rashidi P (2021) Deep Multi-Modal Transfer Learning for Augmented Patient Acuity Assessment in the Intelligent ICU. Front. Digit. Health 3:640685. doi: 10.3389/fdgth.2021.640685

1. INTRODUCTION

Patients admitted to a hospital's intensive care unit (ICU) have life-threatening conditions or the propensity to develop them at any moment. An estimated 5.7 million adults are admitted to ICUs in the United States annually, and their precarious and often rapidly-changing state of health necessitates increased monitoring and hospital resources that costs the U.S. healthcare system more than 67 billion dollars every year (1).

A typical ICU stay occurs in an environment of high-frequency patient monitoring involving a wide variety of physiological measurements such as vital sign tracking, bedside nursing assessments, and laboratory test results. These clinical data points serve as a window into patient illness severity, and taken over time can indicate improving or worsening physiological health. The robust clinical data generated during an ICU stay can aid caregivers in diagnosis

and influence clinical decision-making regarding medication administration, appropriateness of clinical procedures and surgery, and duration and resource requirement of intensive care.

The rich data associated with a typical ICU stay is routinely captured in modern electronic health record (EHR) systems. As of 2017, more than 99% of U.S. hospitals use some form of EHR (2). These longitudinal systems store a large magnitude of patient information including demographics and admission information, vital signs, diagnoses and procedures, laboratory tests, prescriptions and medications, bedside assessments, clinical notes, and more. While inherently useful for care delivery and administrative hospital tasks like billing, EHR systems also function as a rich source for more automated data-driven patient monitoring applications.

Given the potential for health instability commonly associated with patients undergoing intensive care, the timely and accurate assessment of illness severity is invaluable and can inform shared decision-making among patients, families, and providers. Traditionally, overall patient acuity can be measured using a variety of manual, threshold-based scoring systems such as Sequential Organ Failure Assessment (SOFA) (3), Acute Physiology And Chronic Health Evaluation (APACHE) (4), Simplified Acute Physiology Score (SAPS) (5, 6), Modified Early Warning Score (MEWS) (7), and others. More recently, clinical informatics research has demonstrated the validity and accuracy of more automated machine learning approaches using the rich data from EHR systems (8–12). In particular, modern algorithmic techniques using deep learning have been shown to outperform traditional bedside severity scores for predicting in-hospital mortality as a proxy for real-time patient acuity (13). Automated approaches for assessing patient illness severity can help eliminate reliance on overburdened providers, improve the precision of personalized acuity estimates, and be computed in real-time when combined with streaming EHR platforms.

One potential disadvantage of automated patient monitoring solutions is that such systems are limited to physiological data that is recorded in EHR databases. This common paradigm omits important aspects of patient care, including environmental factors (such as noise, light, and sleep), facial expressions that can indicate pain, agitation, or affective state, and aspects of patient mobility and functional status.

Currently, patient pain can be measured by scoring systems such as the Non-Verbal Pain Scale (NVPS) (14) and the Defense and Veterans Pain Rating Scale (DVPRS) (15), and patient activity can be assessed by scoring systems such as the Progressive Upright Mobility Protocol (PUMP) Plus (16) and the ICU Mobility Scale (IMS) (17). However, these manual scores are much less granular than the corresponding physiological measurements and require either self-reporting or repetitive observations by ICU staff (18, 19). The reduced frequency and granularity of these types of patient data can hinder timely intervention strategies (20–25).

To overcome the limitations of current approaches to automated patient monitoring, recent studies have begun to explore the benefits of intensive care units augmented with continuous and pervasive sensing technology. In a study dubbed the Intelligent ICU, Davoudi et al. augmented traditional EHR-based data with patient-worn accelerometer sensors, room-equipped light and sound sensors, and a patient-facing camera (26) (**Figure 1**). Their initial pilot study demonstrated the positive impact of these novel clinical data streams in characterizing delirium in a small prospective cohort of ICU patients. While these non-traditional ICU data sources have shown promise for improving modeling of critically ill patients, Intelligent ICU rooms equipped with pervasive sensors are still in early stages of research.

In this study, we build upon the work of Davoudi et al. by utilizing the data generated by Intelligent ICUs for automated patient acuity assessment using deep learning techniques. In particular, we show that by augmenting existing EHR data with continuous activity measurements via wrist-worn accelerometer sensors, models are better able to capture illness severity by way of more accurate predictions of hospital discharge disposition. We overcome the issue of small sample size in the Intelligent ICU cohorts by employing transfer learning techniques, where learned knowledge and representations from a much larger cohort of EHR-only patients is used as a starting point for subsequent incorporation of the non-traditional data streams. By combining transfer learning with augmented ICU monitoring, our work demonstrates the utility, efficacy, and future promise for using Intelligent ICUs for more personalized and accurate illness severity assessments.

2. MATERIALS AND METHODS

2.1. Study Aims

The primary goal of our study is to characterize the effectiveness of augmenting traditional EHR patient data with a novel Intelligent ICU data source as it pertains to patient acuity assessment using machine learning techniques. Specifically, we combine datasets consisting of several common vital signs with continuous measurements from a wrist-worn activity sensor, and use these augmented datasets to make predictions of a patient's eventual successful or unsuccessful hospital discharge as a proxy for illness severity. In this study, we consider a discharge to home or rehabilitation facility as successful, with in-hospital mortality or transfer to another hospital or hospice being considered unsuccessful.

Our second aim is the evaluation of transfer learning as a solution to cope with the issue of small sample size in our prospective Intelligent ICU patient cohort. We hypothesized that building upon algorithmic patient representations from a much larger cohort of traditional ICU stays would result in improved predictive performance in the smaller cohort of interest.

2.2. Study Cohorts

Our primary cohort of interest, which we refer to as the Intelligent ICU cohort, includes 51 distinct ICU admissions at University of Florida Health between September 2015 and February 2020. These intensive care episodes were made up of 51 unique patients undergoing 51 unique hospital encounters, and occurred within specialized intensive care units outfitted with several unconventional monitoring systems (**Figure 1**). The

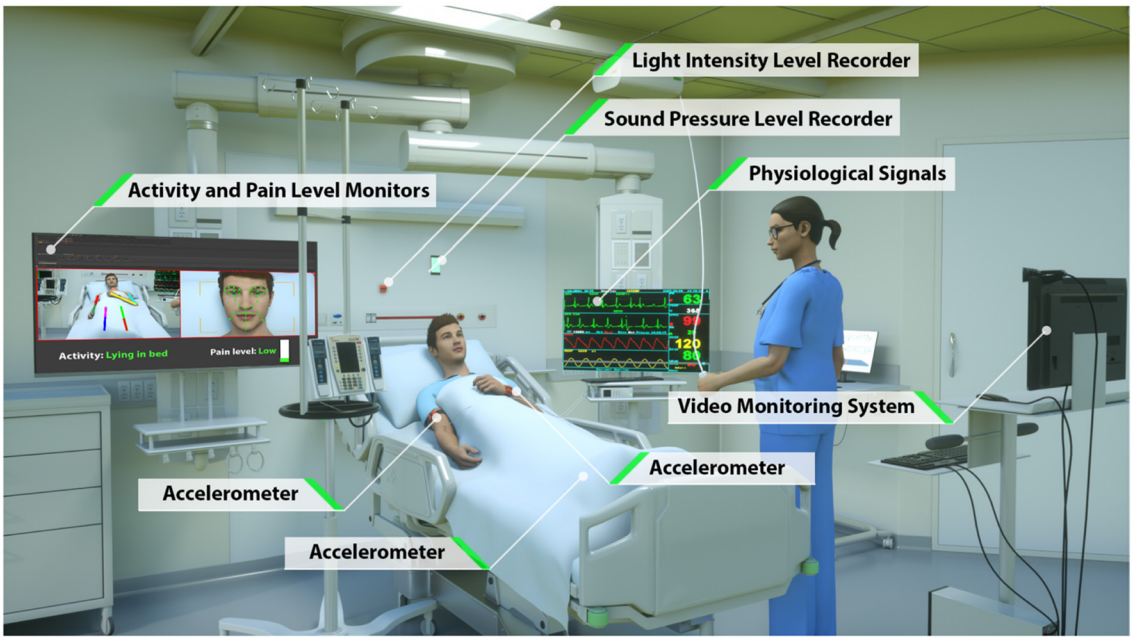


FIGURE 1 | Intelligent ICU room introduced by Davoudi et al. (26). In this study, we augment traditional vital signs from electronic health records with novel activity data from wrist-worn accelerometer sensors.

Intelligent ICU cohort included 33 successful discharges (64.7%) and 18 unsuccessful discharges (35.3%).

For transfer learning experiments, we constructed a much larger second cohort of 48,400 distinct ICU admissions occurring at University of Florida Health between January 2011 and July 2019. We refer to these admissions as the Conventional ICU cohort, as it comes from standard intensive care units that contain only the data available in typical EHR systems. These ICU admissions included 32,184 patients undergoing 45,147 unique hospital encounters. The Conventional ICU cohort included 36,392 successful discharges (75.2%) and 12,008 unsuccessful discharges (24.8%).

This study was approved by University of Florida Institutional Review Board by IRB 201900546. A summary and comparison of admission and demographic descriptors for each cohort is shown in **Table 1**.

2.3. Data Extraction and Processing

2.3.1. Traditional EHR Data

Whether receiving care in an Intelligent ICU or conventional ICU room, all patients have the same set of data recorded into their electronic health records. In this study, for both cohorts we extracted all ICU measurements of six commonly recorded vital signs: diastolic blood pressure, systolic blood pressure, heart rate, respiratory rate, oxygen saturation (SpO2), and temperature.

A multivariate time series of vital signs was constructed for each ICU stay by temporally ordering measurements and resampling to a fixed 1-h frequency, where the mean value was taken if multiple measurements existed in the same 1-h window. We extracted measurements from the entirety of each ICU stay, thus each vital sign sequence was variable length based on the number of hours a patient was in the ICU.

2.3.2. Intelligent ICU Data

The novel environmental and pervasive sensing technology was unique to the 51 ICU stays occurring in our Intelligent ICU cohort. Among all available non-traditional data sources (**Figure 1**), in this pilot study we opted to explore the added utility of wrist-worn activity sensors. Since this is the first study of its kind, we intentionally chose to limit the inclusion of novel data sources as a starting point for exploring and discussing the potential benefits of Intelligent ICU rooms for enhanced patient acuity assessment. While, we provide a brief summary of the technology and data streams contained within Intelligent ICUs, we refer interested readers to the work of Davoudi et al. (26) for a more comprehensive overview.

Patient activity data was collected from an Actigraph GT3X sensor (ActiGraph, LLC. Pensacola, Florida) placed on the patient's dominant wrist when possible, and on the opposite wrist when medical devices prevented ideal placement. These sensors generate activity based on magnitude of wrist motion (27) and sample at a frequency of 100 Hz. In this study, we aggregated accelerometer data into 24-h intervals, and extracted nine statistical features from each consecutive 24-h window after ICU admission. These features included minimum, maximum, mean, variance, standard deviation, immobile count, interquartile range (IQR), root mean square of successive differences (RMSSD), and standard deviation of RMSSD.

TABLE 1 | Summary of Intelligent ICU and Conventional ICU cohorts.

Descriptor	Intelligent ICU (*n* = 51)	Conventional ICU (*n* = 48,400)
Patients, *n*	51	32,184
Hospital encounters, *n*	51	45,147
Hospital length of stay (days), median (25th, 75th)	14.9 (9.0, 21.7)	7.3 (4.2, 12.9)
Successful hospital discharge, *n* (%)	33 (64.7)	36,392 (75.2)
Unsuccessful hospital discharge, *n* (%)	18 (35.3)	12,008 (24.8)
ICU stays, *n*	51	48,400
ICU length of stay (days), median (25th, 75th)	10.3 (6.4, 13.9)	3.0 (1.6, 6.0)
Age (years), median (25th, 75th)	63.2 (43.3, 73.0)	61.2 (48.8, 70.9)
Body mass index, median (25th, 75th)	27.3 (22.9, 33.2)	27.1 (23.1, 32.1)
Charlson comorbidity index, median (25th, 75th)	2.0 (0.0, 4.0)	2.0 (0.0, 4.0)
Sex		
Female, *n* (%)	18 (35.3)	20,188 (44.7)
Male, *n* (%)	33 (64.7)	24,959 (55.3)
Race		
White, *n* (%)	44 (86.3)	34,702 (76.9)
Black, *n* (%)	5 (9.8)	7,615 (16.9)
Other, *n* (%)	2 (3.9)	2,830 (6.2)
Ethnicity		
Hispanic, *n* (%)	1 (2.0)	1,677 (3.8)
Not Hispanic, *n* (%)	50 (98.0)	42,989 (96.2)
Language		
English, *n* (%)	51 (100.0)	44,396 (98.3)
Non-English, *n* (%)	0 (0.0)	751 (1.7)
Marital status		
Married, *n* (%)	24 (55.8)	20,513 (48.3)
Single, *n* (%)	14 (32.6)	13,606 (32.1)
Divorced, *n* (%)	2 (4.7)	4,149 (9.8)
Widowed, *n* (%)	1 (2.3)	3,341 (7.9)
Separated, *n* (%)	2 (4.7)	545 (1.3)
Life partner, *n* (%)	0 (0.0)	292 (0.7)
Provider		
Medicare, *n* (%)	27 (57.5)	23,203 (53.9)
Private insurance, *n* (%)	13 (27.7)	10,707 (24.9)
Medicaid, *n* (%)	4 (8.5)	6,612 (15.4)
Uninsured, *n* (%)	3 (6.4)	2,550 (5.9)
Smoking status		
Smoker, *n* (%)	7 (15.6)	8,514 (21.1)
Former smoker, *n* (%)	21 (46.7)	15,779 (39.1)
Never smoker, *n* (%)	17 (37.8)	16,060 (39.8)

A summary of all features used in our models for both the Intelligent ICU and Conventional ICU cohorts is shown in **Table 2**.

2.3.3. Final Data Preprocessing

For both sequences of patient data, outliers were capped at the 1st and 99th percentiles, with cutoff points determined by the development set of each individual experiment. Any missing extracted feature values in the resulting sequences were imputed with the previous sequence value, if it existed, otherwise with the feature median based on each experiment's development set. Finally, each feature was standardized to zero mean and unit variance based on values from the development set of each experiment.

TABLE 2 | Summary of features used in our experiments.

Feature	Intelligent ICU (*n* = 51) Median (25th, 75th)	Conventional ICU (*n* = 48,400) Median (25th, 75th)
Vital signs		
Diastolic blood pressure, mmHg	62.3 (53.0, 73.0)	63.0 (54.0, 73.0)
Systolic blood pressure, mmHg	123.0 (110.0, 139.0)	121.0 (107.0, 137.5)
Heart rate, beats/min	91.0 (80.0, 103.0)	85.0 (74.0, 97.0)
Respiratory rate, breaths/min	18.3 (15.0, 22.5)	18.0 (15.0, 21.0)
Oxygen saturation (SpO2), %	98.0 (95.0, 100.0)	97.0 (95.0, 99.0)
Temperature, °C	37.0 (36.7, 37.5)	36.9 (36.7, 37.3)
Wrist activity, action counts		
Minimum	0.0 (0.0, 0.0)	N/A
Maximum	62.7 (39.5, 97.1)	N/A
Mean	2.9 (1.2, 6.3)	N/A
Variance	58.6 (16.3, 135.5)	N/A
Standard deviation	7.7 (4.0, 11.6)	N/A
IQR	1.8 (0.0, 8.0)	N/A
RMSSD	7.1 (4.2, 11.6)	N/A
RMSSD standard deviation	1.0 (0.9, 1.1)	N/A
Number immobile	0.6 (0.4, 0.8)	N/A

2.4. Models

In this study, we employ single-layer recurrent neural networks (RNN), a class of deep learning algorithms that are well-suited to processing sequential data and have been validated in literature as accurate clinical models for patient acuity assessment (13). In particular, our RNN models utilize gated recurrent units (GRU) and a linear prediction layer that is used to make a discharge prediction after processing each 24-h data window (**Figure 2**). As each sequential window's features are made available, the model learns a real-time cumulative representation of patient state that is used to predict patient illness severity.

Our study involved the training of two distinct families of recurrent neural networks that were designed to handle either only traditional ICU data, or traditional data augmented with the multi-modal Intelligent ICU data. When using the augmented dataset of both EHR and Intelligent ICU data, we utilized a parallel RNN architecture comprised of two recurrent neural networks that independently processed each data source on separate time scales, with the concatenation of hidden representations passed to the linear prediction layer for assessing final predicted hospital discharge status.

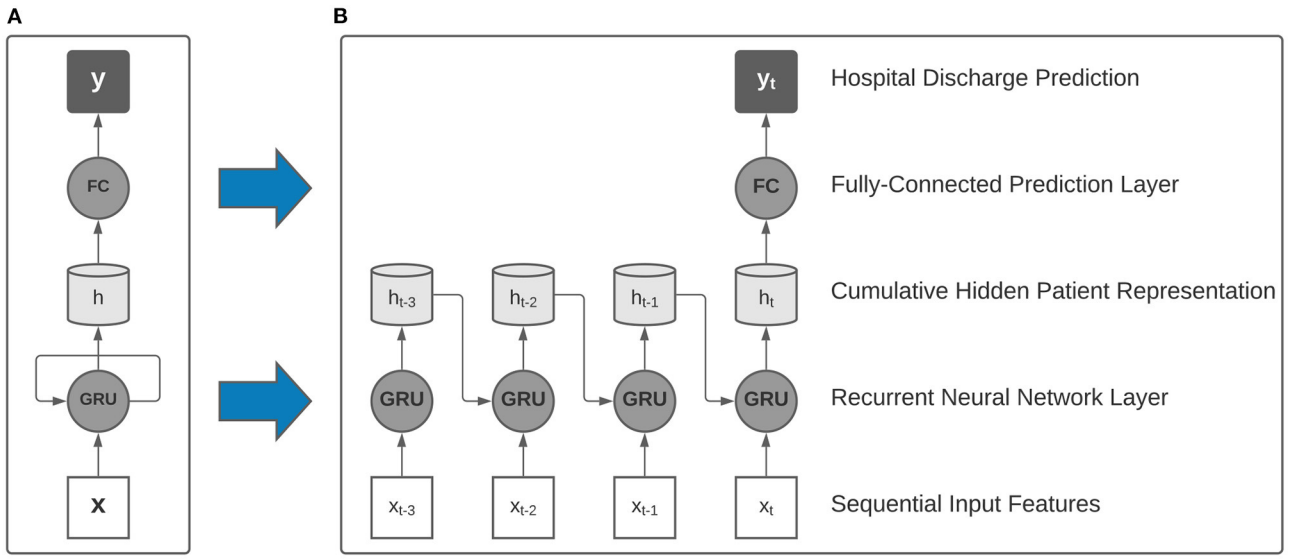


FIGURE 2 | Collapsed **(A)** and expanded **(B)** view of a single-layer recurrent neural network with gated recurrent units (GRU) used as building blocks in our model. After processing multi-resolution sequences of vital signs and activity sensor data, a final hospital discharge prediction was made using the final hidden representation.

2.5. Experiments

Corresponding to our aims in section 2.1, we sought to evaluate the effectiveness of augmenting traditional EHR data with Intelligent ICU data for making predictions of eventual successful or unsuccessful hospital discharge in our cohort of patients undergoing care in Intelligent ICU rooms. Given the small sample size of our Intelligent ICU cohort (n = 51), we also sought to explore the potential benefits of applying the technique of transfer learning, whereby a source model, typically trained on a larger dataset, is used to initialize a smaller model that is subsequently fine-tuned on the smaller dataset of interest. In our transfer learning experiments, we first trained a recurrent neural network on the Conventional ICU cohort (n = 48,400), and transferred its internal RNN weights and biases to a separate model for predicting illness severity in the Intelligent ICU cohort. This transfer learning process is shown in **Figure 3**.

This study includes four experimental variants designed to evaluate our study aims, all using the same discharge disposition targets. All results are reported on the target cohort of 51 ICU encounters occurring in Intelligent ICU rooms.

First, we sought to evaluate predictive performance in the target cohort without the application of transfer learning. The first of these experiments involved the training of a single RNN model on only the EHR data available in the target cohort. Next, we performed a similar experiment using a parallel RNN model with both the EHR and Intelligent ICU data available in the target cohort. These two experimental settings were designed to characterize potential benefits of augmenting traditional EHR data with more novel Intelligent ICU data streams.

We then repeated the above two experiments in conjunction with a transfer learning procedure. In each of these two transfer learning experiments, we first trained a single RNN model on the EHR data from the large Conventional ICU cohort of 48,400 ICU stays. Upon completion of training this source model, we initialized the RNN weights and biases in the EHR portion of the Intelligent ICU models using the final trained RNN weights and biases from the Conventional ICU models (**Figure 3**). The Intelligent ICU models were then trained as normal using the data available in the target Intelligent ICU cohort, in a process known in transfer learning literature as fine-tuning. In both transfer learning experiments, only the final RNN designed to process EHR data was initialized with pre-trained weights, as the Conventional ICU cohort did not contain any novel data sources. Consequently, the final RNN for processing the novel data sources was always trained starting with randomly initialized values.

All experiments on the target Intelligent ICU cohort were performed using 100 repeated trials of randomized five-fold cross-validation stratified by discharge target labels. Within each of the 100 cross-validation experiments, we retained the mean area under the receiver operating characteristic curve (AUROC) across all five validation set folds. 95% confidence intervals were obtained based on percentiles from these 100 averaged AUROC results. When training the large source model on the Conventional ICU cohort, we used the final chronological 20% of ICU stays as validation data, and obtain confidence intervals via 100 bootstrapped iterations based on validation set predictions.

When training a deep learning model, we used a random 20% of the development set for early stopping. Our deep learning models used hidden units of 128 dimensions across all layers, and were trained in batches of 32 samples with an Adam optimizer with learning rate 10^{-3} and $L2$ weight decay of 10^{-3}. All layers used 25% dropout.

3. RESULTS

Training the single RNN model on 80% of the large Conventional ICU cohort and evaluating on the remaining 20% validation set

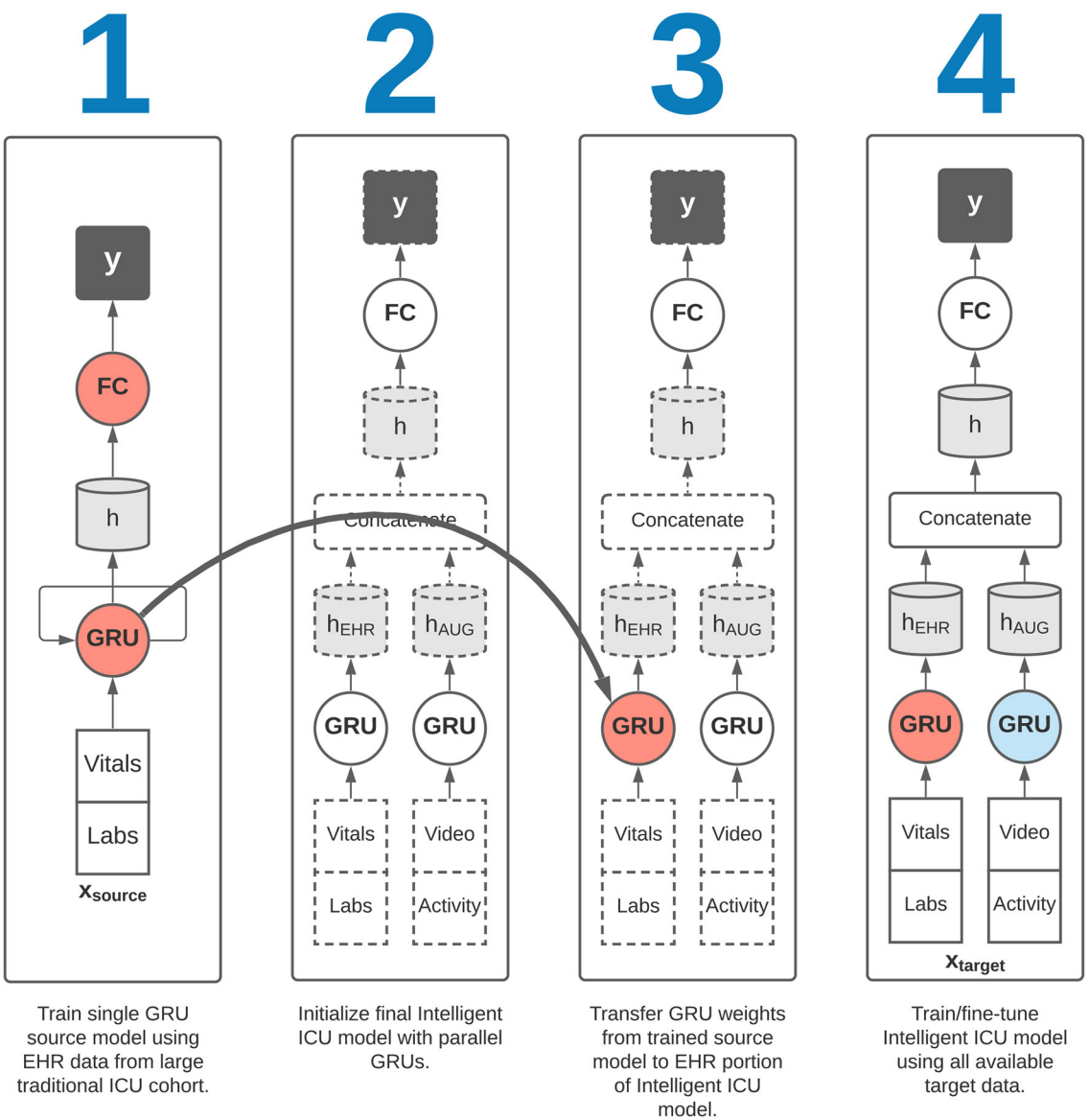


FIGURE 3 | Overview of transfer learning procedure. Our final Intelligent ICU model incorporates pre-trained representational knowledge from a source deep learning model trained only on electronic health records data from a large cohort of traditional ICU stays (n = 48,400).

resulted in an AUROC of 0.752 (95% CI: 0.743–0.763). This trained model was used in all later transfer learning experiments, where the recurrent weights and biases were transferred to the final Intelligent ICU model as shown in **Figure 3**.

The single-cohort and single-RNN Intelligent ICU model using EHR data alone resulted in an AUROC of 0.734 (95% CI: 0.622–0.830). Augmenting the input data with both novel Intelligent ICU data sources and combining with the parallel RNN model resulted in an AUROC of 0.743 (95% CI: 0.644–0.842).

After the application of transfer learning using the model trained on the Conventional ICU cohort, the single-RNN model using only EHR data from the Intelligent ICU cohort resulted in an AUROC of 0.828 (95% CI: 0.557–0.951). The transfer learning model using the augmented dataset of EHR and Intelligent ICU data sources resulted in an AUROC of 0.915 (95% CI: 0.772–0.975).

Results for all experimental settings are summarized in **Table 3**.

4. DISCUSSION

In this study, we have provided the first attempts at incorporating cutting-edge pervasive sensing technology for patient monitoring and precise acuity assessments in the intensive care unit. Based on data from the Intelligent ICU environment of Davoudi et al.

TABLE 3 | Hospital discharge prediction results for all experimental settings.

Target cohort	Input data	Training scheme	AUROC (95% CI)
Conventional ICU	EHR data	Single cohort	0.752 (0.743–0.763)
Intelligent ICU	EHR data	Single cohort	0.734 (0.622–0.830)
Intelligent ICU	EHR + Intelligent data	Single cohort	0.743 (0.644–0.842)
Intelligent ICU	EHR data	Transfer learning	0.828 (0.557–0.951)
Intelligent ICU	EHR + Intelligent data	Transfer learning	0.915 (0.772–0.975)

(26), we explored the performance impact of augmenting deep learning models with two novel data streams for the prediction of successful vs. unsuccessful hospital discharge as a measure of patient illness severity.

Several important takeaways can be gleaned from the performance results summarized in **Table 3**. When comparing single-cohort models trained on EHR data alone, the model trained on the larger Conventional ICU cohort of 48,400 ICU stays relatively outperformed a similar model trained on the much smaller Intelligent ICU cohort of 51 ICU stays (AUROC: 0.752 [95% CI: 0.743–0.763] vs. 0.734 [95% CI: 0.622–0.830]). While not unexpected given the large disparity in cohort sample sizes, the relatively small magnitude of difference between the cohorts is an interesting outcome, as one might expect an even larger discrepancy in model accuracy. While potentially attributable to a variety of factors, these results might suggest clear input patterns associated with improving or worsening health condition that yield diminishing returns as the sample size exponentially increases.

Given the results in **Table 3**, it is also clear that augmenting traditional EHR data with novel activity features in our single-cohort Intelligent ICU model marginally improved its predictive performance (AUROC: 0.743 [95% CI: 0.644–0.842] vs. 0.734 [95% CI: 0.622–0.830]).

Model accuracy was greatly improved using both input dataset variants after the application of transfer learning. When considering EHR data alone, transfer learning increased model accuracy from an AUROC of 0.734 (95% CI: 0.622–0.830) to an AUROC of 0.828 (95% CI: 0.557–0.951). Compared with the results yielded by the single-cohort model in the large Conventional ICU cohort (AUROC: 0.752 [95% CI: 0.743–0.763]), the final accuracy of the Intelligent ICU cohort was much higher. We speculate that these performance improvements point to the power of proper weight initialization in deep learning models, especially for clinical applications using relatively small patient cohorts. We note that although transfer learning with EHR data alone resulted in substantial gains in model accuracy over the model trained on the large Conventional ICU cohort, the prediction confidence interval in the small Intelligent ICU cohort was much wider (95% CI: 0.557–0.951 vs. 0.743–0.763), highlighting the large variability among the cross-validation repetitions. We speculate that this instability was due to the small size of the prediction cohort (n = 51). Given that this is a pilot study demonstrating transfer learning feasibility, we place less emphasis on the fact that absolute accuracy in the smaller cohort was greater than in the larger Conventional ICU cohort, which we partially attribute to sample size disparities. Instead, we focus on the relative performance increase in the same Intelligent ICU cohort, which clearly show the benefits of transfer learning in clinical situations where samples are not readily available.

Maximum overall performance was achieved when combining traditional EHR data with the novel Intelligent ICU data and a transfer learning approach (AUROC: 0.915 [95% CI: 0.772–0.975]). These results indicate the utility of augmenting traditional EHR data with pervasive sensing, and suggest that further research and incorporation of even more novel data streams could be beneficial to the real-time acuity estimation of critically ill patients. These results indicate the power of applying transfer learning in clinical settings with small patient cohorts. It was only when using transfer learning that the predictive benefits of augmented patient data truly became apparent. Similar to the experiments using only EHR data, we focus on the relative performance increase compared with the same augmented dataset in the Intelligent ICU cohort, which show clear benefits for using transfer learning to properly initialize model weights corresponding to electronic health record data from a much larger cohort of conventional ICU patients.

In all experiments using our target Intelligent ICU cohort of 51 ICU stays, the wide AUROC confidence intervals underscore the large variability among the repeated applications of cross-validation. This was not unexpected given the very small size of the Intelligent ICU cohort, especially when used with complex deep learning model architectures. However, when averaged over 100 repeated cross-validation trials, a more clear picture begins to emerge: predictive power is increased both when augmenting traditional vital signs with activity data, and when applying transfer learning, with optimal results achieved after implementing both techniques. We present these results as a pilot study indicating the feasibility of applying transfer learning to small cohorts of patients monitored with non-traditional data streams. While the small sample size of our target Intelligent ICU cohort is less than ideal, we speculate that relative performance increases within the same cohort show future promise for more extensive studies once more Intelligent ICU data becomes available.

Intelligent ICU rooms such as those used in our study are unfortunately rare in practice. However, we feel that pervasive sensing could play an important role in developing a more comprehensive and personalized representation of patient health, and we expect additional types of novel patient monitoring to become more common in future automated patient monitoring applications. Our preliminary results in predicting successful or unsuccessful hospital discharge using a subset of available Intelligent ICU data streams demonstrate the power of non-traditional patient data. As these novel clinical environments become more prevalent, our results also show the necessity of transfer learning approaches to jump-start models using these small augmented cohorts.

Non-traditional patient monitoring data that is not routinely measured in electronic health records as part of

a typical hospital encounter provides a unique opportunity for enhancing clinical decision-making. As we have shown, the accuracy of automated methods for assessing illness severity can be improved when considering such types of novel sensing data. As pervasive sensing becomes more common in traditional intensive care settings, modern machine learning approaches can begin to better understand inherent patterns of data such as patient activity, facial expressions, environmental factors, and more. Augmented patient data can improve clinical decisions such as allocation of clinical resources, altering the characteristics of the ICU room environment, and can help provide objective measures of a patient's affective state and activity that can better inform clinical caregivers regarding appropriateness of medications or procedures.

This study was limited by the use of data from a single institution. Additionally, only a subset of EHR data and Intelligent ICU data was used in this preliminary study. Future work will incorporate all available novel sensing and EHR data, and will focus on even more granular illness severity estimations using higher frequency sensor measurements without aggregation to generate predictions on hourly or sub-hourly time scales. As temporal deep learning techniques continue to evolve, we believe their application to a wide array of both conventional EHR and sensor-based patient health data will lead to large improvements in clinical decision-making and patient outcomes as health trajectories become more accurately predicted and monitored using a more complete perspective on patient health.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: UFHealth cohort data are available from the University of Florida Institutional Data Access/Ethics Committee for researchers who meet the criteria for access to confidential data and may require additional IRB approval. Requests to access these datasets should be directed to https://www.ctsi.ufl.edu/about/research-initiatives/integrated-data-repository/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Florida Institutional Data Access/Ethics Committee.

AUTHOR CONTRIBUTIONS

BS, AD, TO-B, MR, and AB had full access to the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study design performed by BS, AD, and PR. Conception of and data collection from the Intelligent ICU performed by AD, TO-B, AB, and PR. Analyses were performed by BS. Manuscript was drafted by BS and PR. Study supervision was performed by PR and AB. All authors contributed to the acquisition, analysis, and interpretation of data. All authors contributed to critical revision of the manuscript for important intellectual content.

FUNDING

AB, PR, and TO-B were supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01 GM110240. AB and PR were supported by the National Institute of Biomedical Imaging and Bioengineering under Grant 1R21EB027344. PR was supported in part by the NSF CAREER under Grant 1750192. TO-B was supported by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health Grant K01 DK120784.

ACKNOWLEDGMENTS

The authors acknowledge Gigi Lipori, MBA for assistance with data retrieval, and the University of Florida Integrated Data Repository (IDR) and the UF Health Office of the Chief Data Officer for providing the analytic data set for this project. The Titan X GPU partially used for this research was donated by the NVIDIA Corporation.

REFERENCES

1. Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med*. (2010) 38:65–71. doi: 10.1097/CCM.0b013e3181b090d0
2. Pedersen CA, Schneider PJ, Scheckelhoff DJ. ASHP national survey of pharmacy practice in hospital settings: prescribing and transcribing - 2016. *Am J Health Syst Pharm*. (2017) 74:1336–52. doi: 10.2146/ajhp170228
3. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intens Care Med*. (1996) 22:707–10. doi: 10.1007/BF01709751
4. Zimmerman JE, Kramer Aa, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. (2006) 34:1297–310. doi: 10.1097/01.CCM.0000215112.84523.F0
5. Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intens Care Med*. (2005) 31:1336–44. doi: 10.1007/s00134-005-2762-6
6. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. *Intens Care Med*. (2005) 31:1345–55. doi: 10.1007/s00134-005-2763-5
7. Love N, Wrightson J, Walsh S, Keeling N. The value of Modified Early Warning Score (MEWS) in surgical in-patients : a prospective observational study. *Ann R Coll Surg Engl*. (2006) 88:571–5. doi: 10.1308/003588406X130615
8. Shickel B, Tighe PJ, Bihorac A. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. (2018) 22:1589–604. doi: 10.1109/JBHI.2017.2767063
9. Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. In: 4*th International Conference on Learning Representations*. San Juan (2016).

10. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc*. (2017) 24:361–70. doi: 10.1093/jamia/ocw112
11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. *NPJ Digit Med*. (2018) 1:18. doi: 10.1038/s41746-018-0029-1
12. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. *Crit Care*. (2019) 23:1–10. doi: 10.1186/s13054-019-2561-z
13. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci Rep*. (2019) 9:1879. doi: 10.1038/s41598-019-38491-0
14. Payen JF, Gélinas C. Measuring pain in non-verbal critically ill patients: which pain instrument? *Crit Care*. (2014) 18:554. doi: 10.1186/s13054-014-0554-5
15. Buckenmaier III CC, Galloway KT, Polomano RC, McDuffie M, Kwon N, Gallagher RM. Preliminary validation of the defense and veterans pain rating scale (DVPRS) in a military population. *Pain Med*. (2013) 14:110–23. doi: 10.1111/j.1526-4637.2012.01516.x
16. Titsworth WL, Hester J, Correia T, Reed R, Guin P, Archibald L, et al. The effect of increased mobility on morbidity in the neurointensive care unit. *J Neurosurg*. (2012) 116:1379–88. doi: 10.3171/2012.2.JNS111881
17. Tipping CJ, Bailey MJ, Bellomo R, Berney S, Buhr H, Denehy L, et al. The ICU mobility scale has construct and predictive validity and is responsive. a multicenter observational study. *Ann Am Thorac Soc*. (2016) 13:887–93. doi: 10.1513/AnnalsATS.201510-717OC
18. Parry SM, Granger CL. Assessment of impairment and activity limitations in the critically ill: a systematic review of measurement instruments and their clinimetric properties. *Intens Care Med*. (2015) 41:744–62. doi: 10.1007/s00134-015-3672-x
19. Thrush A, Rozek M, Dekerlegand JL. The clinical utility of the functional status score for the intensive care acute care hospital: a prospective cohort study. *Phys Therapy*. (2012) 92:1536–45. doi: 10.2522/ptj.20110412
20. Brown H, Terrence J, Vasquez P, Bates DW. Continuous monitoring in an inpatient medical-surgical unit: a controlled clinical trial. *Am J Med*. (2014) 127:226–32. doi: 10.1016/j.amjmed.2013.12.004
21. Kipnis E, Ramsingh D, Bhargava M, Dincer E, Cannesson M, Broccard A, et al. Monitoring in the intensive care. *Crit Care Res Pract*. (2012) 2012:473507. doi: 10.1155/2012/473507
22. To KB. Common complications in the critically ill patient. *Surg Clin*. (2012) 92:1519–57. doi: 10.1016/j.suc.2012.08.018
23. Wollschlager CM, Conrad AR. Common complications in critically ill patients. *Disease Month*. (1988) 34:225–93. doi: 10.1016/0011-5029(88)90009-0
24. Rubins HB, Moskowlte MMA. Complications of Care in a Medical Intensive Care Unit. *J Gen Intern Med*. (1990) 5:104–9. doi: 10.1007/BF02600508
25. Desai SV, Law TJ, Needham DM. Long-term complications of critical care. *Crit Care Med*. (2011) 39:371–9. doi: 10.1097/CCM.0b013e3181fd66e5
26. Davoudi A, Malhotra KR, Shickel B, Siegel S, Williams S, Ruppert M, et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Sci Rep*. (2019) 9:8020. doi: 10.1038/s41598-019-44004-w
27. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. *J Sci Med Sport*. (2011) 14:411–6. doi: 10.1016/j.jsams.2011.04.003

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.

frontiers in Digital Health

BRIEF RESEARCH REPORT
published: 25 February 2021
doi: 10.3389/fdgth.2020.584555

Identifying Heart Failure in ECG Data With Artificial Intelligence—A Meta-Analysis

*Dimitri Grün[1], Felix Rudolph[1], Nils Gumpfer[2], Jennifer Hannig[2], Laura K. Elsner[1], Beatrice von Jeinsen[3], Christian W. Hamm[1,3], Andreas Rieth[3], Michael Guckert[2,4] and Till Keller[1,3]**

[1] Department of Internal Medicine I, Cardiology, Justus-Liebig University Giessen, Giessen, Germany, [2] Cognitive Information Systems, KITE - Kompetenzzentrum für Informationstechnologie, Technische Hochschule Mittelhessen - University of Applied Sciences, Friedberg, Germany, [3] Department of Cardiology, Kerckhoff Heart and Thorax Center, Bad Nauheim, Germany, [4] Department of MND - Mathematik, Naturwissenschaften und Datenverarbeitung, Technische Hochschule Mittelhessen - University of Applied Sciences, Friedberg, Germany

OPEN ACCESS

Edited by:
Amanda Christine Filiberto, University of Florida, United States

Reviewed by:
Tyler John Loftus, University of Florida, United States
Shameer Khader, AstraZeneca, United States

***Correspondence:**
Till Keller
keller@chestpain.de

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

Received: 17 July 2020
Accepted: 29 December 2020
Published: 25 February 2021

Citation:
Grün D, Rudolph F, Gumpfer N, Hannig J, Elsner LK, von Jeinsen B, Hamm CW, Rieth A, Guckert M and Keller T (2021) Identifying Heart Failure in ECG Data With Artificial Intelligence—A Meta-Analysis. Front. Digit. Health 2:584555. doi: 10.3389/fdgth.2020.584555

Introduction: Electrocardiography (ECG) is a quick and easily accessible method for diagnosis and screening of cardiovascular diseases including heart failure (HF). Artificial intelligence (AI) can be used for semi-automated ECG analysis. The aim of this evaluation was to provide an overview of AI use in HF detection from ECG signals and to perform a meta-analysis of available studies.

Methods and Results: An independent comprehensive search of the PubMed and Google Scholar database was conducted for articles dealing with the ability of AI to predict HF based on ECG signals. Only original articles published in peer-reviewed journals were considered. A total of five reports including 57,027 patients and 579,134 ECG datasets were identified including two sets of patient-level data and three with ECG-based datasets. The AI-processed ECG data yielded areas under the receiver operator characteristics curves between 0.92 and 0.99 to identify HF with higher values in ECG-based datasets. Applying a random-effects model, an sROC of 0.987 was calculated. Using the contingency tables led to diagnostic odds ratios ranging from 3.44 [95% confidence interval (CI) = 3.12–3.76] to 13.61 (95% CI = 13.14–14.08) also with lower values in patient-level datasets. The meta-analysis diagnostic odds ratio was 7.59 (95% CI = 5.85–9.34).

Conclusions: The present meta-analysis confirms the ability of AI to predict HF from standard 12-lead ECG signals underlining the potential of such an approach. The observed overestimation of the diagnostic ability in artificial ECG databases compared to patient-level data stipulate the need for robust prospective studies.

Keywords: artificial intelligence, heart failure, diagnosis, ECG, meta-analysis

INTRODUCTION

Heart failure (HF) is a common, yet unfavorable, cardiac condition. Up to 20% of all individuals in developed countries develop HF within their lifetime, and a large proportion of patients hospitalized for HF dies within 1 year of diagnosis (1).

Evaluation of symptoms suggestive of HF currently demands physicians to valuate various parameters including imaging and laboratory data and the electrocardiogram (ECG). Besides a standard examination that includes an ECG, imaging information, such as echocardiography or magnetic resonance imaging, is seen as gold standard in diagnosis of HF (2). Nevertheless, an adequate use of such imaging data is associated with relevant technical infrastructure and medical expertise. The ECG is a well-established, quick, and easily accessible method for diagnosis and screening of various cardiovascular diseases. It provides specific features that indicate presence of HF or prognosis in HF patients especially to rule out HF in case of a normal ECG (3, 4). However, use of an ECG as primary diagnostic instrument often only yields insufficient diagnostic specificity (5). Further, general practitioner–based ECG reporting has varying results, introducing further diagnostic uncertainty (6).

Devices providing medically relevant information generated directly by individuals outside the healthcare system such as smartphones with health applications or wearables including smartwatches are an emerging trend. This development promises that a growing number of, e.g., ECG data generated at home will be available for a diagnostic screening. Such data have already shown potential in computer-aided decision support systems to warn patients of rhythmic abnormalities (7). Management of this quantity of data, however, might be a challenge for the individual healthcare professional, as well as for the healthcare system itself. The potentially beneficial use of artificial intelligence (AI) in cardiology in general has been discussed already, e.g., as a tool for clinicians that could facilitate precision in daily practice and even might improve patient outcomes (8). AI might also be able to help in interpretation of ECG signals and could therefore be used to analyze ECG data in specific cases and on a large scale for early identification of cardiovascular diseases such as HF (9). Few studies have performed analyses of AI systems to detect HF from ECG data. In these studies, the methods and patient numbers vary strongly. The aim of the present evaluation was to perform a meta-analysis on these studies and thereby give an overview on the current possibilities of the use of AI in automated HF detection from ECG signals.

METHODS

A comprehensive literature search for original articles on the ability of AI to predict HF based on ECG signals was conducted using the databases PubMed and Google Scholar on May 13, 2020. These two databases were searched using the following keyword combinations as search query: ("heart failure" OR "ejection fraction" OR "systolic dysfunction" OR "diastolic dysfunction") AND ("computer-aided diagnosis" OR "ai" OR "artificial intelligence" OR "deep learning" OR "machine learning" OR "neural network") AND ("ecg" OR "ekg" OR "electrocardiogram" OR "electrocardiography"). The term "computer-aided" was added to the query to not miss articles that use a more general title potentially not revealing an AI approach as basis for a computer-based classification algorithm. This search query led to a list of 118 titles that were further screened and selected by three of the authors (D.G., F.R., and T.K.). As primary endpoints, the criteria congestive HF and reduced left ventricular ejection fraction [left ventricular ejection fraction (LVEF) $\leq$40%] were used. Identification of this endpoint had to be based on ECG time-series data as input by an AI approach. Artificial neural networks, support vector machines, random forest classifiers, and k-nearest neighbor algorithms qualified as an AI approach in this context. The screening and selection process was carried out in three steps: first a title, then an abstract, and finally a full text screening and selection. Evaluation of studies within the first and second steps was conducted by the three mentioned investigators independently. A study was selected for evaluation within the next step if at least two of the three investigators selected the individual study. After abstract classification, a total of 23 studies were selected for full text assessment. The subsequent third step was conducted by the same three investigators independently, followed by a discussion within the investigator team and a consensual selection of the articles to be evaluated within the meta-analysis. Within this third step, the quality of the studies was assessed oriented on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (10). Further, data availability of the needed information, e.g., reporting of a confusion matrix, was checked. The final set of studies consisted of five articles that fulfilled the defined criteria and provided sufficient information for the subsequent data extraction enabling the meta-analysis. This selection process including the applied criteria is also depicted with a flowchart as **Figure 1**.

To assess the heterogeneity between the selected studies, the DerSimonian-Laird estimator (τ^2) and I^2 statistics were used (11, 12). Within the meta-analysis, principal measurement of effect size was the diagnostic odds ratio (DOR) after natural logarithmic transformation (lnDOR) with 95% confidence interval (CI). For univariate analyses, a random-effects model was used. For the bivariate analyses, a summary receiver operating characteristics (sROC) curve was constructed, and a summary area under the ROC curve was calculated. For descriptive reasons, for the studies that did not provide these data, an AUC was estimated based on the respective contingency table (13–15). All statistical analyses were carried out using R3.6.0 with the *meta* (V4.12-0) and the *mada* (V0.5.10) packages (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

The five evaluated studies comprise a total of 57,027 patients and 579,134 ECG datasets. Two of these studies, both published by Attia et al. are based on patient-level data with large cohort sizes of 3,874 and of 52,870 individuals, reflecting a clinical application of an AI-based diagnostic approach (16, 17). These cohorts

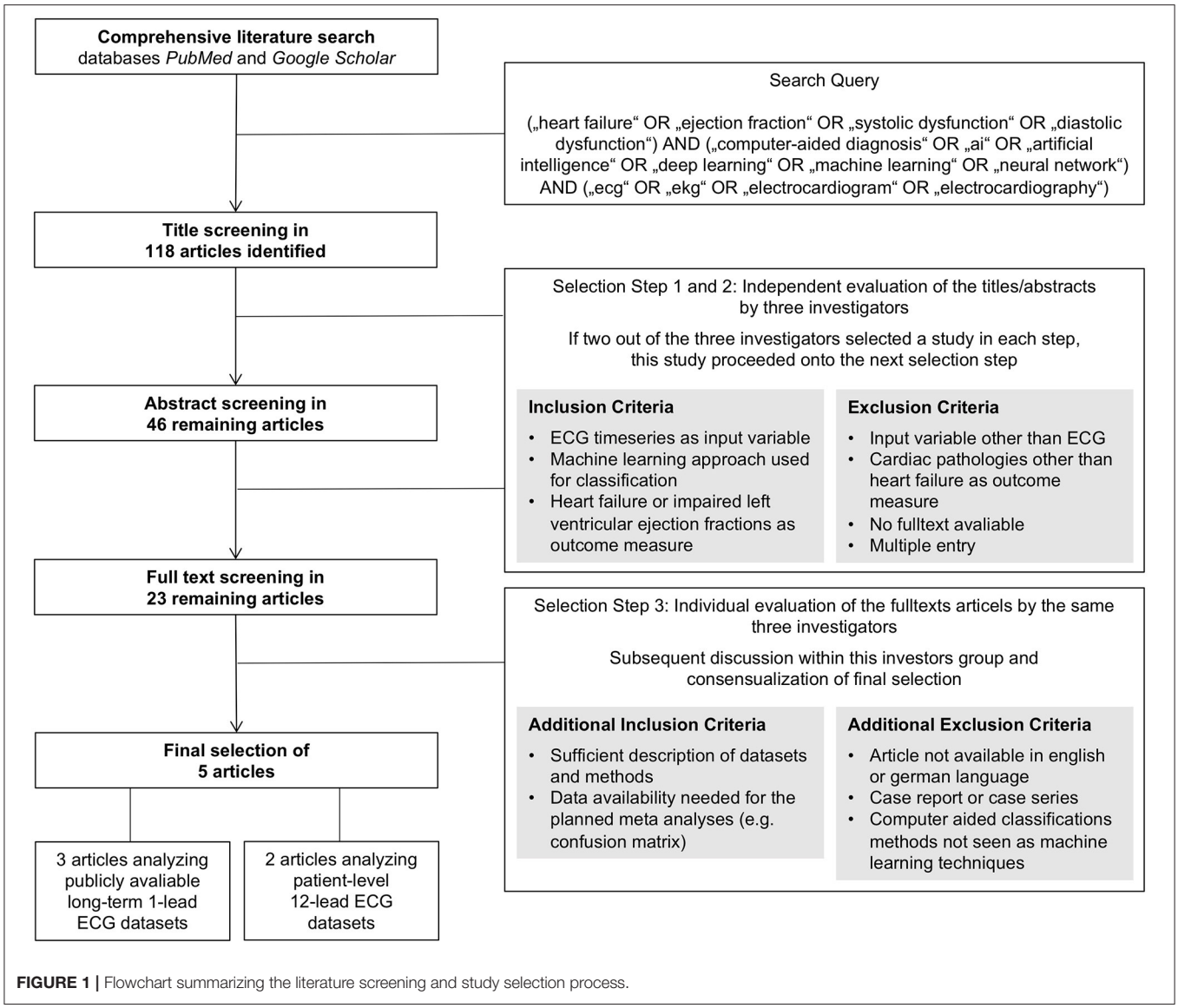


FIGURE 1 | Flowchart summarizing the literature screening and study selection process.

comprised unselected patients who underwent routine ECG and available echocardiographic data with the endpoint LVEF ≤35%. The other three studies used large numbers of ECG datasets as basis stemming from only a small number of individuals (33–107). These ECG datasets were taken from different existing databases such as the publicly available Fantasia or BIDMC database used in all three evaluated publications (18–20). Here, endpoint was the classification as congestive HF provided within these databases.

Four studies used the raw ECG time-series data as input with 500 to 12 × 1,000 features comprising the input of the respective algorithms (14–17), whereas one study used five extracted features as input (13). The proposed respective computer-aided diagnostic algorithms used a convolutional neural network (CNN) in three publications (14, 16, 17), a CNN plus long short-term memory network in one publication (15), and a dual-tree complex wavelet transform (DTCWT) model in one publication (13). The latter was accepted as an AI approach for this meta-analysis as all other criteria were fulfilled even if DTCWT itself would not qualify according to the predefined AI methods.

The algorithms of the five evaluated studies were associated with sensitivities ranging from 83 to 100% and specificities ranging from 86 to 100% identifying HF with higher values in ECG dataset–based studies. **Table 1** provides an overview of the five evaluated studies.

As meta-analysis, we calculated a combined DOR of 7.59 (95% CI = 5.85–9.34) after log transformation. This high lnDOR reflects the lnDORs of the individual studies starting from 3.44 (95% CI = 3.12–3.76) up to 13.61 (95% CI = 13.14–14.08) with lower diagnostic performance in patient-level datasets (**Figure 2**). For the bivariate analysis, an sROC curve was calculated, leading to a combined area under the curve of 0.987. Again, the diagnostic performance was lower in patient-level studies with an area under the curve of 0.92 and 0.93 compared to 0.96, 0.99, 0.99,

TABLE 1 | Summary of the studies included in the meta-analysis.

Study	Classification method	Input features	Outcome measure	No. of patients	No. of ECGs	Classification performance
Sudarshan et al. (13)	DTCWT	Five features based on 2-s segments of one-lead long-term ECG recordings	CHF	Set1: 55 Set2: 33	Set1: 82,427 Set2: 84,952	Sens (1): 1.00 (95% CI = 1.00–1.00) Spec (1): 1.00 (95% CI = 1.00–1.00) Sens (2): 0.97 (95% CI = 0.97–0.97) Spec (2): 0.99 (95% CI = 0.99–0.99)
Acharya et al. (14)	CNN	500 features based on 2-s segments of one-lead long-term ECG recordings	CHF	Set1: 33 Set2: 55	Set1: 100,308 Set2: 140,000	Sens (1): 0.97 (95% CI0.96–0.97) Spec (1): 0.96 (95% CI = 0.96–0.96) Sens (2): 0.99 (95% CI = 0.99–0.99) Spec (2): 0.99 (95% CI = 0.99–0.99)
Attia et al. (17)	CNN	12 × 1,000 features (zero-padded to 1,024) based on a 2-s segment from 10-s 12-lead ECG recordings	Low LVEF	52,870	52,870	Sens: 0.83 (95% CI = 0.78–0.87) Spec: 0.87 (95% CI = 0.86–0.88)
Attia et al. (16)	CNN	12 × 1,000 features (zero-padded to 1,024) based on a 2-s segment from 10-s 12-lead ECG recordings	Low LVEF	3,874	3,874	Sens: 0.86 (95% CI = 0.85–0.87) Spec: 0.86 (95% CI = 0.85–0.86)
Lih et al. (15)	CNN-LSTM	2,000 features based on 2-s segments of one-lead long-term ECG recordings	CHF (+ MI, CAD)	107	114,703	Sens: 0.99 (95% CI = 0.99–0.99) Spec: 0.98 (95% CI = 0.98–0.98)

DTCWT, dual-tree complex wavelet transform; CNN, convolutional neural network; LSTM, long short-term memory; CHF, congestive heart failure; LVEF, left ventricular ejection fraction; MI, myocardial infarction; CAD, coronary artery disease; Sens, sensitivity; Spec, specificity.

0.98, and 0.99 (**Figure 3**). This observed heterogeneity between the individual studies is reflected by a τ^2 of 5.52 and I^2 of 100% ($p < 0.001$).

DISCUSSION AND CONCLUSIONS

The observed diagnostic information of an AI approach using ECG data to identify HF in our meta-analysis confirms the potential of computer-aided decision-making using ECG data in diagnoses other than arrhythmias. Our analysis further shows a relevant heterogeneity between studies based on ECG data and studies based on patient-level datasets suggesting that a meta-analysis incorporating both study types might not be as meaningful as desired. Further limitation for a meta-analysis of these five studies is the varying endpoint. Still, the individual results of the studies itself all show promising results pointing in the same direction supporting the information of the meta-analysis.

Three publications of our meta-analysis are based on cases from one-lead long-term ECG recordings of the BIDMC congestive HF database, which consists of only 15 patients (13–15). Those recordings were segmented into short 2-s intervals to artificially increase the number of datasets.

In contrast, the studies of Attia et al. used 2-s segments stemming from standard 12-lead ECGs with a length of 10 s obtained in 3,874 and 52,870 individual patients, respectively (16, 17). These datasets might better depict real-life data as analyses of the segmented ECGs seem to overestimate the ability of AI to detect HF in comparison. These patient-based datasets still show a clinically relevant diagnostic information with an AUC of > 0.8. This assumption is further supported by a study by Kwon et al. who reported comparable patient-based dataset AUCs of 0.843 and 0.889 for two datasets (3,378 and 5,901 patients) (21). Interestingly, the used datasets, here patient-based vs. ECG-based, had a larger impact on the model performance compared to a difference in input features. Using ECG datasets, the study by Sudarshan et al. (13) with only five features, yielded a comparable classification performance to the studies by Acharya et al. (14) with 500 input features, and Lih et al. (15) with 2,000 input features.

ECG characteristics are known to vary according to ethnicity, possibly impacting the accuracy of an AI algorithm that was

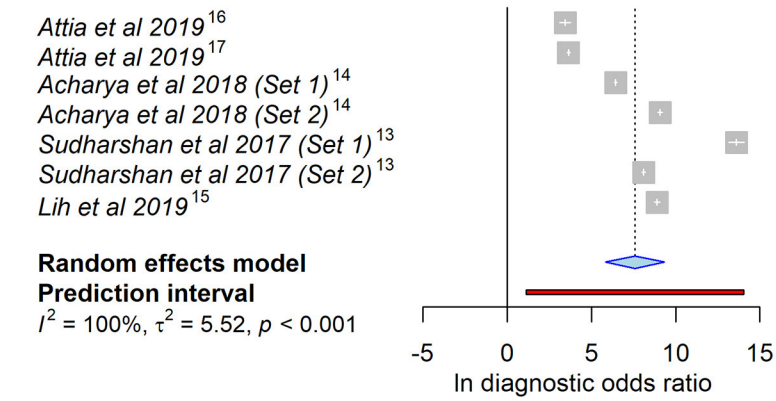


FIGURE 2 | Forest plot of the selected studies showing the ability to identify heart failure using artificial intelligence–processed ECG data. Data presented as a univariate analysis using a random-effects model with diagnostic odds ratio after natural logarithmic transformation (lnDOR) with respective confidence interval (CI).

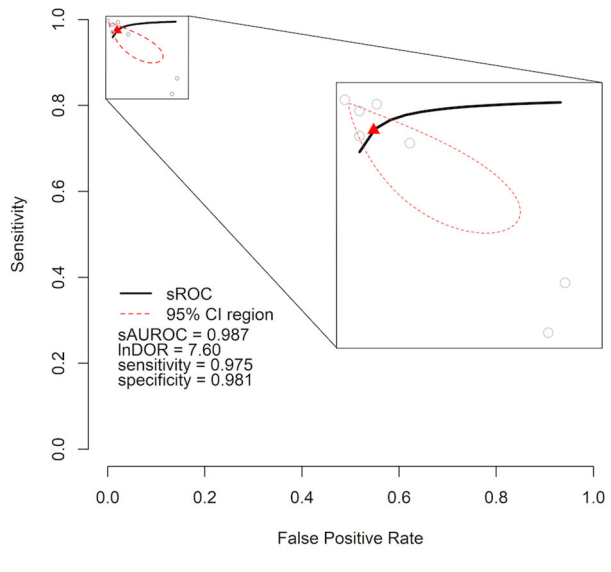


FIGURE 3 | Cumulative summary receiver operating characteristic curve (sROC) of an artificial intelligence–processed ECG approach to detect heart failure. Individual studies are shown as gray circles. Summary point is shown as red triangle. The area of interest is magnified on the right side. lnDOR denotes diagnostic odds ratio after natural logarithmic transformation, sAUROC denotes area under the sROC curve; CI, confidence interval.

trained with datasets stemming from specific geographical regions. Using the same dataset as Attia et al. (16, 17), Noseworthy et al. found that, while varying accuracies between ethnic groups are present, their network performed consistently across multiple ethnicities (22).

Besides ECG data, other information available after a recommended clinical diagnostic workup (2) might also be a valid input for an AI approach. Here, the use of data stemming from classical imaging techniques such as chest X-rays (23) or from the gold-standard imaging method of echocardiography (24) has shown a relevant potential. Also, traditional diagnostic methods, not relying on a complex infrastructure, like the evaluation of heart sound via a computer-aided approach (25), might be of use in the evaluation of HF patients. Further, combination of such different modalities as input features compared to a single diagnostic method might increase model precision in a real-world setting. Such an idea is supported by data showing that various information taken from electronic health records within a machine learning approach is able to predict HF before it is clinically obvious (26). With the inhomogeneous nature regarding features as well as outcome measures in AI-aided HF diagnosis, this analysis focuses on ECG time series as input variable. Nevertheless, other input parameters and the combination of different modalities have to be addressed by future studies.

The present meta-analysis, as well as the published data, underlines the need for robust large patient-level data–based studies to better appraise the value of AI in ECG interpretation in the context of HF. Here, the ongoing ECG AI-Guided Screening for Low Ejection Fraction (EAGLE) cluster randomized trial (NCT04000087) will provide useful prospective insights representing a real-life setting (27, 28).

Recently, technology and acceptance of wearables, smart-health devices, and applications have widely improved. The growing processing power and system memory will diminish technical limitations. Especially, one-lead ECG assessment has been implemented as feature into several devices. Supporting our observations regarding different types of ECG input, promising data on the transferability of a neural network trained with 12-lead ECGs to a one-lead ECG–enabled device have been presented at the annual meeting of the American Heart Association in 2019 underlining the potential of such an approach (29).

To conclude, the data of this meta-analysis confirm a substantial ability of AI to predict HF or a reduced LVEF from

standard ECG signals. With the current advances of mobile devices capable of ECG recording, AI might be a powerful future tool in screening for HF or even diagnosis of other diseases of the heart.

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.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Conception and design of the work was done by DG, FR, and TK. Data was collected by DG, FR, and TK. Data analyses were done by DG, NG, and JH. DG and FR visualized the data. The draft of the manuscript was created by DG, FR, and TK. MG and TK supervised the project. NG, JH, LE, BJ, CH, AR, and MG contributed to the interpretation of the result and critically revised the manuscript. All authors gave approval of the final version of the manuscript.

REFERENCES

1. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu Sh, et al. Heart failure: preventing disease and death worldwide. *ESC Hear Fail.* (2014) 1:4–25. doi: 10.1002/ehf2.12005
2. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC guidelines for the diagnosis treatment of acute chronic heart failure: The task force for the diagnosis treatment of acute chronic heart failure of the European society of cardiology (ESC) developed with the special contribution of the heart failure association (HFA) of the ESC. *Eur Heart J.* (2016) 37:2129–200. doi: 10.1093/eurheartj/ehw128
3. Lucena F, Barros AK, Ohnishi N. The performance of short-term heart rate variability in the detection of congestive heart failure. *Biomed Res Int.* (2016) 2016:1675785. doi: 10.1155/2016/1675785
4. Sadeghi R, Dabbagh VR, Tayyebi M, Zakavi SR, Ayati N. Diagnostic value of fragmented QRS complex in myocardial scar detection: systematic review and meta-analysis of the literature. *Kardiol Pol.* (2016) 74:331–7. doi: 10.5603/KP.a2015.0193
5. Davenport C, Cheng EYL, Kwok YTT, Lai AHO, Wakabayashi T, Hyde C, et al. Assessing the diagnostic test accuracy of natriuretic peptides and ECG in the diagnosis of left ventricular systolic dysfunction: a systematic review and meta-analysis. *Br J Gen Pract.* (2006) 56:48–56.
6. Goudie BM, Jarvis RI, Donnan PT, Sullivan FM, Pringle SD, Jeyaseelan S, et al. Screening for left ventricular systolic dysfunction using GP-reported ECGs. *Br J Gen Pract.* (2007) 57:191–5.
7. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med.* (2019) 381:1909–17. doi: 10.1056/NEJMoa1901183
8. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. *J Am Coll Cardiol.* (2018) 71:2668–79. doi: 10.1016/j.jacc.2018.03.521
9. Jahmunah V, Oh SL, Wei JKE, Ciaccio EJ, Chua K, San TR, et al. Computer-aided diagnosis of congestive heart failure using ECG signals - a review. *Phys Medica Eur J Med Phys.* (2019) 62:95–104. doi: 10.1016/j.ejmp.2019.05.004
10. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *BMJ.* (2009) 339:b2700. doi: 10.1016/j.jclinepi.2009.06.006
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* (1986) 7:177–88. doi: 10.1016/0197-2456(86)90046-2
12. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* (2002) 21:1539–58. doi: 10.1002/sim.1186
13. Sudarshan VK, Acharya UR, Oh SL, Adam M, Tan JH, Chua CK, et al. Automated diagnosis of congestive heart failure using dual tree complex wavelet transform and statistical features extracted from 2 s of ECG signals. *Comput Biol Med.* (2017) 83:48–58. doi: 10.1016/j.compbiomed.2017.01.019
14. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M, et al. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Appl Intell.* (2019) 49:16–27. doi: 10.1007/s10489-018-1179-1
15. Lih OS, Jahmunah V, San TR, Ciaccio EJ, Yamakawa T, Tanabe M, et al. Comprehensive electrocardiographic diagnosis based on deep learning. *Artif Intell Med.* (2020) 103:101789. doi: 10.1016/j.artmed.2019.101789
16. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol.* (2019) 30:668–74. doi: 10.1111/jce.13889
17. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med.* (2019) 25:70–74. doi: 10.1038/s41591-018-0240-2
18. Baim DS, Colucci WS, Monrad ES, Smith HS, Wright RF, Lanoue A, et al. Survival of patients with severe congestive heart failure treated with oral milrinone. *J Am Coll Cardiol.* (1986) 7:661–70. doi: 10.1016/s0735-1097(86)80478-8
19. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation.* (2000) 101:E215–20. doi: 10.1161/01.cir.101.23.e215
20. Iyengar N, Peng CK, Morin R, Goldberger AL, Lipsitz LA. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *Am J Physiol.* (1996) 271:R1078–84. doi: 10.1152/ajpregu.1996.271.4.R1078
21. Kwon JM, Kim KH, Jeon KH, Kim HM, Kim MJ, Lim SM, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. *Korean Circ J.* (2019) 49:629–39. doi: 10.4070/kcj.2018.0446
22. Noseworthy PA, Attia ZI, Brewer LPC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythmia Electrophysiol.* (2020) 13:e007988. doi: 10.1161/CIRCEP.119.007988
23. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology.* (2019) 290:514–22. doi: 10.1148/radiol.2018180887
24. Tabassian M, Sunderji I, Erdei T, Sanchez-Martinez S, Degiovanni A, Marino P, et al. Diagnosis of heart failure with preserved ejection fraction: machine learning of spatiotemporal variations in left ventricular deformation. *J Am Soc Echocardiogr.* (2018) 31:1272–84.e9. doi: 10.1016/j.echo.2018.07.013
25. Zheng Y, Guo X, Qin J, Xiao S. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. *Comput Methods Programs Biomed.* (2015) 122:372–83. doi: 10.1016/j.cmpb.2015.09.001
26. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data. *Med Care.* (2010) 48:S106–13. doi: 10.1097/mlr.0b013e3181de9e17

27. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken EM, et al. Clinical trial design data for electrocardiogram artificial intelligence-guided screening for low ejection fraction (EAGLE). *Data Br.* (2020) 28:104894. doi: 10.1016/j.dib.2019.104894
28. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken EM, et al. ECG AI-guided screening for low ejection fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. *Am Heart J.* (2020) 219:31–36. doi: 10.1016/j.ahj.2019.10.007
29. Attia ZI, Dugan J, Maidens J, Rideout A, Lopez-Jimenez F, Noseworthy PA, et al. Abstract 13447: prospective analysis of utility of signals from an Ecg-enabled stethoscope to automatically detect a low ejection fraction using neural network techniques trained from the standard 12-lead Ecg. *Circulation.* (2019) 140:A13447. doi: 10.1161/circ.140.suppl_1.13447

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 Digital Health

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fdgth.2021.608893

Patient-Specific Sedation Management via Deep Reinforcement Learning

Niloufar Eghbali[1], *Tuka Alhanai*[2] *and Mohammad M. Ghassemi*[1]*

[1] *Human Augmentation and Artificial Intelligence Laboratory, Department of Computer Science, Michigan State University, East Lansing, MI, United States,* [2] *Laboratory for Computer-Human Intelligence, Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates*

Introduction: Developing reliable medication dosing guidelines is challenging because individual dose–response relationships are mitigated by both static (e. g., demographic) and dynamic factors (e.g., kidney function). In recent years, several data-driven medication dosing models have been proposed for sedatives, but these approaches have been limited in their ability to assess interindividual differences and compute individualized doses.

Objective: The primary objective of this study is to develop an individualized framework for sedative–hypnotics dosing.

Method: Using publicly available data (1,757 patients) from the MIMIC IV intensive care unit database, we developed a sedation management agent using deep reinforcement learning. More specifically, we modeled the sedative dosing problem as a Markov Decision Process and developed an RL agent based on a deep deterministic policy gradient approach with a prioritized experience replay buffer to find the optimal policy. We assessed our method's ability to jointly learn an optimal personalized policy for propofol and fentanyl, which are among commonly prescribed sedative–hypnotics for intensive care unit sedation. We compared our model's medication performance against the recorded behavior of clinicians on unseen data.

Results: Experimental results demonstrate that our proposed model would assist clinicians in making the right decision based on patients' evolving clinical phenotype. The RL agent was 8% better at managing sedation and 26% better at managing mean arterial compared to the clinicians' policy; a two-sample t-test validated that these performance improvements were statistically significant ($p < 0.05$).

Conclusion: The results validate that our model had better performance in maintaining control variables within their target range, thereby jointly maintaining patients' health conditions and managing their sedation.

Keywords: medication dosing, personalized medicine, deep reinforcement learning, propofol, sedation management

OPEN ACCESS

Edited by:
Amanda Christine Filiberto, University of Florida, United States

Reviewed by:
Matthieu Komorowski, Imperial College London, United Kingdom
Ira L. Leeds, Johns Hopkins University, United States
Mengling Feng, National University of Singapore, Singapore

***Correspondence:**
Mohammad M. Ghassemi
ghassem3@msu.edu

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

Received: *21 September 2020*
Accepted: *04 February 2021*
Published: *31 March 2021*

Citation:
Eghbali N, Alhanai T and Ghassemi MM (2021) Patient-Specific Sedation Management via Deep Reinforcement Learning. Front. Digit. Health 3:608893. doi: 10.3389/fdgth.2021.608893

INTRODUCTION

Intensive care units (ICUs) serve patients with severe health issues who need continuous medical care and monitoring (1). In the course of their treatment within ICUs, patients generate a wide variety of data that are stored in electronic health record systems including computed tomography scans, care-provider free-text notes, clinician treatment decisions, and patient demographics. The task of a clinician is to carefully consider these data to infer the latent disease *state* of their patients and (given this state) apply an optimal treatment *policy* (a set of *actions*) that will maximize the odds of short-term patient survival and longer-term patient recovery. This sequential inference process used by clinicians during care is one instance of a greater class of problems referred to as reinforcement learning (RL) in the artificial intelligence community.

Interest in the applications of RL to healthcare has grown steadily over the last decade. Within the last few years, numerous works have demonstrated the potential of RL methods to help manage sensitive treatment decisions in sepsis (1–5), sedation regulation (6, 7), mechanical ventilation (1, 8), and medication dosing (9–11). Refer to the works of Liu and Prescott (12) and Yu et al. (1) for a recent systematic review of RL models in critical care and healthcare. In this article, we demonstrate the use of deep RL for the regulation of patient sedation. Sedation is essential for invasive therapies such as endotracheal intubation, ventilation, suction, and hemodialysis, all of which may result in patient pain or discomfort when conducted without the assistance of sedatives (13, 14); it follows that sedation management is an important component of effective patient treatment in critical care environments.

Sedation management is particularly challenging because ICU patients enter treatment for a variety of health reasons (often with incomplete medical records) and may require prolonged periods of sedation as they recover (15, 16). Overdosing sedatives has been associated with several negative health outcomes including longer recovery times, increased need for radiological evaluation, increased odds of long-term brain dysfunction, and death (7, 17). Conversely, underdosing sedatives may result in untreated pain, anxiety, and agitation, which have been associated with patient immunomodulation and posttraumatic stress disorder (13). Hence, great care must be taken in the delicate process of sedation management (14), where patients may exhibit unique pharmacological responses for the same dose of a given medication. This results in pharmacokinetic or pharmacodynamic variations for the same drug administered with the same frequency in different individuals (18, 19). In order to address this issue, a growing number of clinical studies have proposed automated methods based on patients' evolving clinical phenotypes to deliver safe and effective sedation regulation (6, 16, 20).

RL is a promising methodological framework for sedation regulation because it can learn nuanced dosing policies that consider variation in disease intensity, drug responsiveness, and personal patient characteristics (1, 20). In the past few decades, several RL-based models have been proposed to regulate sedation in the ICU (6, 7, 21–29). However, most sedation management methods exhibit one or more of the following limitations: (1) incomplete physiological context or patient response variability, (2) use of simulated data for validation, (3) failure to account for common clinical practices such as attempts to minimize the total dosage of sedatives (17), and (4) assumption of discrete state and action spaces resulting in sensitivities to heuristic choices of discretization levels (5). Lastly, most of the prior work has focused on a specific medication—propofol—which has no intrinsic analgesic effect and must be coadministered with an opioid or other analgesic for ICU patients (30).

Our work herein extends previous studies by employing an RL framework with continuous state-action spaces to identify an optimal dosing policy for *both* a common sedative and opioid medication *together* (propofol and fentanyl). Our proposed model considers interindividual differences to reach the target level of sedation as measured by the Riker Sedation–Agitation Scale (SAS), while also minimizing the total sedative amount administered. Although our sedation measure is based on patient behaviors, which do not directly reflect the brain, they are useful as an optimization target for both their reliability and ease of collection (31); the SAS is a progressive sedation–agitation indicator with excellent interrater reliability (32).

MATERIALS AND METHODS

In this section, the critical care data set and our preprocessing approach are introduced. The decision-making framework and its associated RL components are discussed afterward.

Data

Database

All data for this study were collected from the Medical Information Mart for Intensive Care (MIMIC-IV), a freely accessible ICU data resource that contains de-identified data associated with more than 60,000 patients admitted to an ICU or the emergency department between 2008 and 2019 (33, 34).

Key Variables

We extracted 1,757 patients from MIMIC who received a commonly used sedative (propofol) and opioid (fentanyl) during their ICU stay; for each of these patients, we also extracted a time series of sedation level according to SAS. SAS is a 7-point ordinal scale that describes patient agitation: 1 indicates "unarousable," 4 indicates "calm and cooperative," and 7 indicates "dangerous agitation" levels. SAS serves as our therapeutic target for this work; it has been shown previously that optimization of patients' level of sedation is associated with decreased negative outcomes, such as time spent on mechanical ventilation (17). We note that our study population excluded all patients diagnosed with severe respiratory failure, intracranial hypertension, status epilepticus traumatic brain injury, acute respiratory distress syndrome, and severe acute brain injury (including severe traumatic brain injury, poor-grade subarachnoid hemorrhage, severe ischemic/hemorrhagic stroke, comatose cardiac arrest, status epilepticus) because sedation management approaches for such patients are idiosyncratic (35, 36).

TABLE 1 | Summary of data set.

Gender	% Survivors	Mean age (y)	Mean hours in ICU	No. of patients
Female	100	75	157	806
Male	100	65	146	1,301
Total population	**100**	**69**	**149**	**1,757**

TABLE 2 | Summary statistics of selected features based on different levels of sedation [Riker Sedation–Agitation Scale (SAS)]. Last row presents the proportion of data in each level.

SAS / Features	SAS = 1 Unarousable	SAS = 2 Very sedated	SAS = 3 Sedated	SAS = 4 Calm, cooperative	SAS = 5 Agitated	SAS = 6 Very agitated	SAS = 7 Dangerous agitation
Noninvasive blood pressure mean	74 ± 17	72 ± 16	74 ± 17	76 ± 71	79 ± 18	79 ± 19	81 ± 17
Diastolic blood pressure	59 ± 15	60 ± 19	60 ± 23	64 ± 418	69 ± 625	65 ± 18	66 ± 15
Heart rate	86 ± 21	88 ± 19	89 ± 477	88 ± 213	91 ± 21	94 ± 18	94 ± 19
Respiration rate	21 ± 7	21 ± 38	20 ± 8	20 ± 9	21 ± 6	21 ± 6	22 ± 7
Arterial PH	7 ± 0	7 ± 0	7 ± 0	7 ± 0	7 ± 0	7 ± 0	7 ± 0
Positive end-expiratory pressure set	7 ± 4	9 ± 5	7 ± 3	5 ± 3	5 ± 3	6 ± 3	6 ± 2
Oxygen saturation pulse oximetry (Spo_2)	96 ± 7	96 ± 6	97 ± 5	97 ± 40	97 ± 3	96 ± 6	97 ± 3
Inspired oxygen fraction (Fio_2)	52 ± 18	54 ± 17	47 ± 13	46 ± 70	46 ± 15	47 ± 16	55 ± 21
Arterial oxygen partial pressure	137 ± 69	126 ± 65	123 ± 57	120 ± 53	120 ± 58	122 ± 57	117 ± 44
Plateau pressure	21 ± 6	23 ± 8	20 ± 6	18 ± 4	19 ± 5	20 ± 6	19 ± 3
Average airway pressure	12 ± 5	14 ± 6	11 ± 12	7 ± 3	9 ± 13	9 ± 4	8 ± 3
Mean arterial pressure (MAP)	80 ± 20	79 ± 25	83 ± 74	88 ± 42	89 ± 41	100 ± 63	85 ± 29
Proportion of data %	**3.32**	**6.37**	**20.47**	**53.15**	**5.94**	**0.45**	**0.06**

Measures Utilized

According to the American Society of Anesthesiologists, current recommendations for monitoring sedation include blood pressure (diastolic blood pressure and mean noninvasive blood pressure), respiration rate, heart rate, and oxygen saturation pulse oximetry (Spo_2) (37); we utilized these measures in our modeling approach. Additionally, we utilized measures based on studies conducted by Yu et al. (1) and Jagannatha et al. (38), including arterial pH, positive end-expiratory pressure (PEEP), inspired oxygen fraction (Fio_2), arterial oxygen partial pressure, plateau pressure, average airway pressure, mean arterial pressure (MAP), age, and gender.

A total of 14 features were used to describe patients in our data: diastolic blood pressure, mean noninvasive blood pressure, respiration rate, heart rate, Spo_2, arterial pH, PEEP, Fio_2, arterial oxygen partial pressure, plateau pressure, average airway pressure, MAP, age, and gender (dichotomized, with male coded as 0). Prior to modeling, all continuous measures were zero-mean variance normalized.

Table 1 presents summary information about the final data set, which contained a total of 1,757 subjects, with a 100% survival rate, a mean age of 68.5 years, and a mean ICU stay of 149.8 h. **Table 2** provides summary statistics of the measures based on different levels of sedation defined by SAS. The final row presents the proportion of data available in each level, which exhibits a Gaussian distribution with the mean at SAS level 4 out of 7 (calm and cooperative).

Preprocessing and Time Windowing

For each patient, we divided the ICU stay duration into hourly contiguous windows. A given window may contain multiple recordings of a given measure. In windows with more than one recording, the mean of the recording was used. To address missing data, we removed entries where data for all measures, or the SAS outcome, were missing and applied the sample-and-hold interpolation technique. We imputed any remaining missing values with the mean value of the missing measure calculated across the training data.

Training, Validation, and Testing Set Partition

We partitioned our data at the subject level into a training (60%, 1,055 subjects, 156,303 time windows), validation (20%, 351 subjects, 49,997 time windows), and test set (20%, 351 subjects, 55,493 time windows). The training data set was used to identify model parameters; the validation set was used to identify model hyperparameters, and the testing set was used to evaluate the model's ability to generalize to data unseen during training.

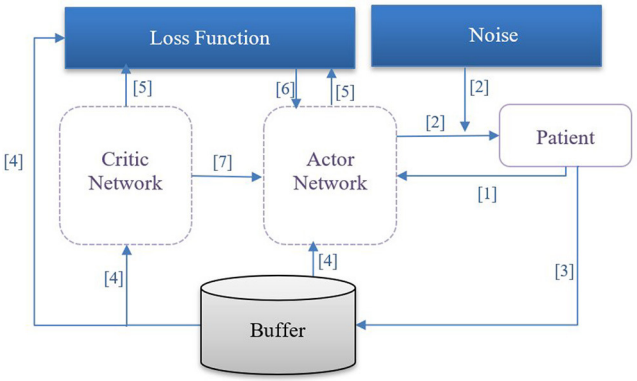


FIGURE 1 | DDPG procedure: [1] The agent observes patient's state $\mathbf{s_t}$ and transfers it to the actor network. [2] The actor network receives $\mathbf{s_t}$ as an input and outputs the dosage amount plus a small noise (action); the purpose of the noise is to promote exploration of the action space. [3] The agent observes a reward $\mathbf{r_t}$, and patient's next clinical state $\mathbf{s_{t+1}}$; the tuple of $< \mathbf{s_t}, \mathbf{a_t}, \mathbf{r_t}, \mathbf{s_{t+1}} >$ is retained in an experience pool. [4] From the experience pool, a batch of N tuples will be selected to learn the optimal policy. [5] The temporal difference loss function is computed. [6] The critic network is updated by minimizing the temporal difference loss. [7] The actor network is updated using the deterministic policy gradient theorem.

Model Architecture

The sedation dosing problem can be cast as a Markov Decision Process (MDP) where the purpose is to find an optimal dosing policy that, given the patient's state, specifies the most effective dosing action (1, 9). Our RL model is based on a deep deterministic policy gradient (DDPG) approach introduced by (39). DDPGs benefit from the advantages of deterministic policy gradients (DPGs) (40) and deep Q networks (41), which robustly solve problems in continuous action spaces. In order to learn the optimal policy, we used an *off-policy* RL algorithm that studied the success (and failures) of the clinicians' policies in our data set. In the following sections, the proposed method is elaborated.

Policy

We modeled the sedation management problem as an MDP described by the tuple (S, A, P, R), in which

- $\mathbf{s_t} \in S$ is the patient state containing the 14 dimensional feature vector described above in a given hourly window t;
- $\mathbf{a_t} \in A$ is a two-dimentional action vector corresponding to the quantity of propofol and fentanyl administered in a given hourly window.
- $P(\mathbf{s_{t+1}} | s_t, a_t)$ is the probability of the next state vector given the current state vector and the action taken.
- $r(\mathbf{s_t}, a_t) \in R$ is the observed reward following a state transition at time window t that is related to how closely the SAS and blood pressure of the patient match the optimal value (discussed in *Reward*).

Given our formulation of the sedation management problem, we trained an RL agent that (1) observes the current patient state $\mathbf{s_t}$, (2) updates the medication doses with an optimal action $\mathbf{a_t}$, and (3) receives a corresponding reward $r(\mathbf{s_t}, a_t)$ before moving to the next state $\mathbf{s_{t+1}}$ and continuing the process. For the agent to maximize its cumulative reward over several state-action pairs, it must learn a policy π–a function that maps states (patient's state) to actions (drug dosages): $a = \pi(s)$. In training, the RL agent uses a sequence of observed state-action pairs (s_t, a_t), called a trajectory (τ), to learn the optimal policy π^* by maximizing the following objective function:

$$\mathbf{J}(\pi) = \mathbb{E}\left[\mathbf{R}(\tau)\right] = \mathbb{E}_{\mathbf{s}}[\int_{\mathbf{a}} \mathbf{p}(\tau|\pi)\mathbf{R}(\tau)\mathbf{d}\tau] \quad (1)$$

where $\mathbf{R}(\tau) = \mathbf{r_t} + \gamma\mathbf{r_{t+1}} + \gamma^2\mathbf{r_{t+2}} + \gamma^3\mathbf{r_{t+3}} + \ldots + \gamma^{\mathbf{T}}\mathbf{r_{t+T}}$ is a sum of discounted rewards, γ is a discount factor that determines the relative weight of immediate vs. long-term rewards, and θ denotes the set of model parameters learned during RL training. If γ is close to 0, the agent is biased toward short-term rewards; if γ is close to 1, the agent is biased toward longer-term rewards. In our case, the value of γ was 1E-3 and was determined by exploring several values of γ and retaining the value that maximized the model's performance on the validation set.

In our case, the specific formulation of π^* is determined via DDPG, which employs four neural networks to ultimately learn the optimal policy from the trajectories: a Q network (critic), a deterministic policy network (ac), a target Q network, and a target policy network. The "critic" estimates the value function, while the "actor" updates the policy distribution in the direction suggested by the critic (for example, with policy gradients). The target networks are time-delayed copies of their original networks that slowly track the learned networks and greatly improve the stability of learning. Similar to deep Q learning, DDPG utilizes a replay buffer to https://www.powerthesaurus.org/collect/synonymscollect experiences for updating neural network parameters. During each trajectory, all the experience tuples (state, action, reward, next state) will be stored in a finite-sized cache called "replay buffer." At each time window, the actor and critic are updated by sampling a minibatch from the buffer. The replay buffer allows the algorithm to benefit from learning across a set of uncorrelated transitions. Instead of sampling experiences uniformly from replay buffer, we have used prioritized experience reply (42) to replay important transitions more frequently, thereby learning more efficiently. In our case, the next state $\mathbf{s_{t+1}}$, is computed by a neural network consisting of three fully connected layers with ReLu activation functions in the first two layers and a linear activation in the final layer. Batch normalization was used during training. Models were implemented in Pytorch 1.6.0 and used Adam optimization (43). We illustrate the procedure of DDPG for finding the optimal policy for medication dosing in **Figure 1** and describe the procedure below:

(1) The agent observes the patient's state $\mathbf{s_t}$ and transfers it to the actor network.
(2) The actor network receives $\mathbf{s_t}$ as an input and outputs the dosage amounts plus a small noise (actions); the purpose of the noise is to promote exploration of the action space.
(3) The agent observes a reward $\mathbf{r_t}$ and the patient's next clinical state $\mathbf{s_{t+1}}$. The tuple of $< \mathbf{s_t}, \mathbf{a_t}, \mathbf{r_t}, \mathbf{s_{t+1}} >$ is stored in a pool of experiences.
(4) From the pool of experiences, a batch of N tuples will be used to learn policies.

(5) The loss function [temporal difference (TD)] is then computed.
(6) The critic network is updated by minimizing the loss.
(7) The actor network is updated using the DDPG theorem.

Reward

In order to learn from the trajectories, our RL agent requires a formal definition of reward based on deviations from the control variables (SAS, MAP). Propofol administration lowers sympathetic tone and causes vasodilation, which may decrease preload and cardiac output and consequently lower the MAP and other interrelated hemodynamic parameters. Therefore, ensuring a desired range of MAP is an essential consideration of propofol infusion (7, 44). Moreover, efforts should be made to minimize the sedative dosage (17). Under these premises, the reward issued to the sedation management agent at each time window is defined with the purpose of keeping SAS and MAP measurements at the clinically acceptable and safe range while penalizing increases in dose; for our purposes, these ranges are described by the following equations:

$$\mathbf{r_{MAP}} = \frac{\mathbf{2}}{\mathbf{1}+\mathbf{e}^{-(\mathbf{MAP_t}-\mathbf{65})}} - \frac{\mathbf{2}}{\mathbf{1}+\mathbf{e}^{-(\mathbf{MAP_t}-\mathbf{85})}} - \mathbf{1} \tag{2}$$

$$\mathbf{r_{RSS}} = \frac{\mathbf{2}}{\mathbf{1}+\mathbf{e}^{-(\mathbf{SAS_t}-\mathbf{3})}} - \frac{\mathbf{2}}{\mathbf{1}+\mathbf{e}^{-(\mathbf{SAS_t}-\mathbf{4})}} - \mathbf{1} \tag{3}$$

where r_{MAP} assigns value close to 1 when MAP values fall within the therapeutic range of 65–85 mmHg and negative values elsewhere; r_{RSS} assigns value close to 1 when SAS value falls within the therapeutic range of 3–4 and negative values elsewhere. Target therapeutic ranges are selected based on Hughes et al. (17) and Padmanabhan et al. (7), respectively.

Next, let $\mathbf{D_t}$ describe deviations from the clinically acceptable and safe range of SAS and MAP in time window t with the static lower target boundary (LTB) and upper target boundary (UTB) described above:

$$\mathbf{D_t}\left(\text{control variable}\right) \begin{cases} \text{if measured value for control variable is in target range}, & 0 \\ \text{if measured value for control variable} < \text{LTB}, & \text{LTB} - \text{measured value for control variable} \\ \text{if measured value for control variable} > \text{UTB}, & \text{UTB} - \textit{measured value for control variable} \end{cases} \tag{4}$$

From this deviation, we may compute the total error in time window t from both control variables as follows:

$$\mathbf{error_t} = \mathbf{D_t}\,(\mathbf{MAP}) + \mathbf{D_t}\,(\mathbf{SAS}) \tag{5}$$

If $\mathbf{e_{t+1}}$ (deviation from target range for MAP and SAS at time window $t+1$) is $\geq \mathbf{e_t}$, then we assign $\mathbf{r_{t+1}} = 0$, which serves to penalize a "bad" action.

$$\mathbf{r_t} = \begin{cases} \mathbf{r_{SAS}} + \mathbf{r_{MAP}} - 0.02\,\mathbf{r_{dosage}} & \text{if } \mathbf{e_t} < \mathbf{e_{t-1}} \\ 0 \qquad \text{otherwise} \end{cases} \tag{6}$$

where r_{dosage} is the amount of the medications provided.

Performance Evaluation Approach

We compared the performance of our model to the recorded performance of the clinical staff with the reasonable assumption that the clinical staff intended to keep patients within the therapeutic range during their ICU stay. For this purpose, the performance error is defined for each trajectory (hours spent in ICU) as follows:

$$PE_i^c = \frac{\text{patient } i \text{ ICU duration} \;-\; \text{time control variable } c \text{ is in target range}}{\text{patient } i \text{ ICU duration}} \times 100 \tag{7}$$

Equation 7 captures the proportion of the total ICU stay hours that patient i spent outside the therapeutic range for the control variable $c \in \{SAS, MAP\}$. If the measured value falls within the target interval, the difference between the measured value and the target value will be zero; otherwise, the difference will be computed based on the target interval boundaries. More specifically, to assess the sedation management performance of the trained agent against the clinical staff, the root mean square error (RMSE), mean performance error (MPE), and median performance error (MDPE) were compared for chosen actions under both our model policy and the clinicians' policy (24). MDPE gives the control bias observed for a single patient and is computed by:

$$\text{MDPE}_\text{i}^c = \text{median}\left(\text{PE}_\text{i}^c\right) \tag{8}$$

RMSE_i^c is the RMSE for each patient and control variable, which is computed using

$$\text{RMSE}_i^c = \sqrt{\frac{\sum_{\mathbf{t=1}}^{\mathbf{N}} \left(\mathbf{D_t}\,(c)\right)^2}{N}} \tag{9}$$

where N represents ICU stay duration in hours, and t iterates over the set of hourly measurements for each patient i.

RESULTS

For assessment purposes, we applied our model to the held-out test set (351 patients, 55,493 h); patients in the test set had a mean ICU duration of 158 h.

In **Table 3**, we present the performance for both the learned sedation management policy and clinicians' policy (as reflected by the data). **Table 3** indicates that MDPE and RMSE for our model are lower than that of clinicians; this means that our learned sedation management policy may reduce the amount of time a patient spends outside the

TABLE 3 | Performance metrics for control variables SAS (Riker Sedation–Agitation Scale) and MAP (mean arterial pressure).

Performance metric	Control variables			
	Learned policy		Clinician's policy	
	MAP	SAS	MAP	SAS
MPE %	17.82 ± 9.22	8.69 ± 1.14	44.66 ± 23.18	17.43 ± 21.54
MDPE %	15.0	0	45.45	0.69
Mean RMSE	23.45	0.08	46.38	0.71
Mean Values	74.99 ± 4.47	3.42 ± 0.07	85.26 ± 28.4	3.47 ± 1.04
Mean propofol dosage	10.49 ± 60		24.23 ± 132	
Mean fentanyl dosage	15.9 ± 8.9		15.1 ± 2.3	

The MPE (mean performance error), MDPE (median performance error), and RMSE (root mean square error) values for learned policy are lower for both control variables, which means our model had a better performance in keeping these variables in their target range.

therapeutic range when compared to the clinicians. As seen in **Table 3**, the measured values for SAS and MAP are within the target range for 91.3% and 82.2% of the patient ICU duration, respectively. These results correspond to a 26% (MAP) and 8% (SAS) improvement in MPE, compared to the clinicians' policy. A two-sample *t*-test validates that the reduction of performance error and RMSE in our model is significant ($p < 0.05$) compared to the clinicians' policy; the results validate that our model had better performance in maintaining control variables within their target range, thereby jointly maintaining patients' health condition and managing their sedation.

In **Figure 2**, we compare the SAS and MAP value distributions using a boxplot; the green box corresponds to our model's results. The figure indicates that that our policy has promising results for sedation management while keeping MAP in the target range. The lower SAS values predicted by our model, as seen in **Figure 2**, are reasonable as our model suggests less medication, on average, which therefore leads to lower levels of sedation (lower SAS).

In **Table 3**, we show the mean medication amount for patients for both the learned policy and clinicians' policy. We assessed the ability of our model to lower the total amount of medication administered while maintaining the therapeutic status of patients. More specifically, for each patient trajectory, we computed the medication administered by our policy, compared to the clinicians. A two-sample *t*-test indicated a statistically significant reduction in the total amount of medication administered by our RL agent ($p < 0.03$) compared to the clinicians. Thus, we conclude that dosage amounts administered to patients following our model is lower than the clinician's prescription.

In **Figure 3**, we illustrate the RL-based closed-loop sedation scenario for three randomly selected patients. The figure shows the variation in SAS and MAP values for three randomly selected patients during ICU stay; dashed lines depict the changes when using the clinician's policy, constant lines represent our proposed policy, and the green area represents the target range. **Figure 3** illustrates the ability of our model to drive SAS values to the therapeutic range without drastic deviation from the MAP therapeutic range for these three randomly selected patients. The evaluation results confirm that the RL agent is able to maintain the SAS value and MAP value in the target ranges while lowering the medication amount.

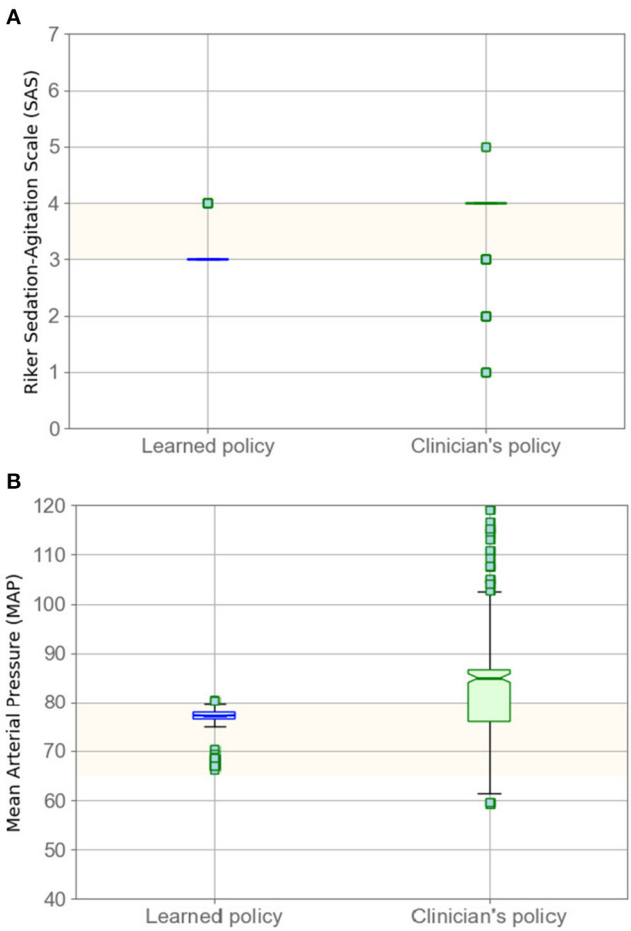


FIGURE 2 | Evaluating policy in terms of SAS **(A)** and MAP **(B)** boxplots. Left boxplots (blue) correspond to our model, whereas the right boxplots are clinicians' results. Highlighted areas identify the target range.

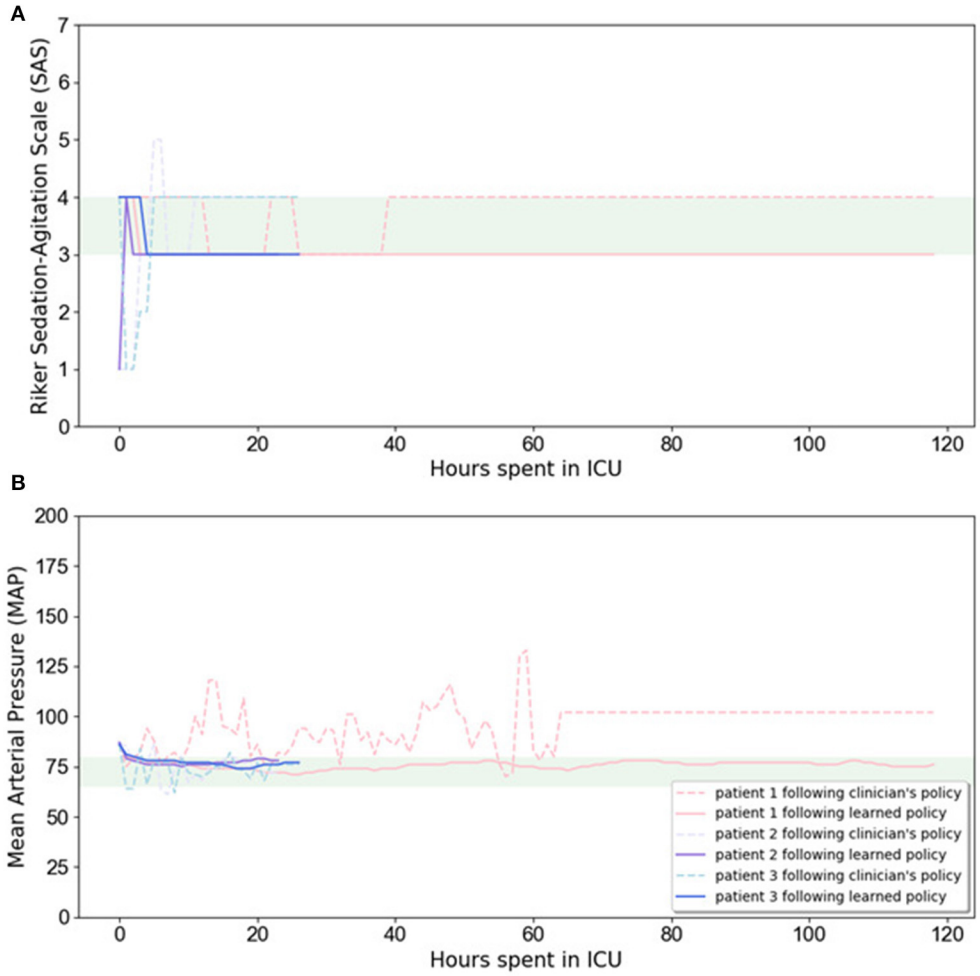


FIGURE 3 | Variation in SAS **(A)** and MAP **(B)** values for three randomly selected patients during ICU stay. Dashed lines depict the changes when using clinician's policy, while constant lines are related to learned policy, and the highlighted area is the target range.

DISCUSSION

In this study, we proposed a deep RL method based on a DDPG approach to manage propofol administration while considering the dynamic observations that were available in patient's electronic medical records. We utilized RL because it is an effective framework for deriving optimal and adaptive regulation of sedatives for patients with different responses to the same medication and is able to learn an optimal sequence of decisions from retrospective data. Moreover, RL-based methods can be practically applied to real clinical practice by taking simple steps. RL has two main components: the *environment* (patient) and the *agent* (our sedative regulator). Every time the agent performs an action (recommends dosage), the patient gives a reward to the agent, which can be positive or negative depending on how appropriate the dosage was from that specific state of a patient. The goal of the agent is to learn what dosage maximizes the reward, given every possible state of the patient. *States* are the observations that the agent receives at each step in the patients' care process. Using retrospective data from medical records, our agent will learn from the set of patient states, administered dosage, response to the doses, and the reward it gets. After initial training of the agent, it is able to generalize over the state space to recommend doses in situations it has not previously encountered. In a practical setting, the state observed by the agent may be either extracted from the electronic medical record directly or provided by the clinician through a user interface.

This work extends previous studies in a number of ways. First, our trained agent operates in a continuous action space; this distinguishes it from prior models that utilized Q learning for medication dosing with an arbitrary discretization of the action space. Second, we used the SAS to assess the patient's sedation level, which is one of the most widely used sedation scales in the ICU, but instead of merely regulating sedation level, we also trained our agent to consider hemodynamic parameters (MAP) by reflecting them in the reward function. Third, in practical clinical settings, it is common to minimize the sedative dosage, which is unaccounted for in prior works on medication dosing

using RL. To address this limitation, we penalized the increase in medication dosage while learning the optimal policy. Our test results confirm the ability of our model to manage sedation while also lowering the dosage in comparison to clinicians' prescriptions. Therefore, our policy leads to lower administration of sedatives in comparison to the clinicians' policy; the sedation level during sedative administration is close to the lower target SAS boundary, which corresponds to higher sedation.

Administration of sedatives such as propofol can have adverse effects on the hemodynamic stability of patients. Specifically, propofol causes vasodilation leading to a decrease in MAP (7). Our results indicate a notable improvement (26%) in MAP management compared to the recorded performance of clinicians. This achievement is important because if MAP drops below the therapeutic range for an extended period, end-organ manifestations such as ischemia and infarction can occur. If MAP drops significantly, blood will not perfuse cerebral tissues, which may result in loss of consciousness and anoxic injury (45).

We conclude that our sedation management agent is a promising step toward automating sedation in the ICU. Furthermore, our model parameters can be tuned to generalize to other commonly used sedatives in ICU and will work with other sedation monitoring scales such as bispectral index or Richmond Agitation and Sedation Scale.

Further efforts need to be taken in order for the method described herein to be effective enough for real-world deployment. Long-term anesthetic infusion often results in drug habituation, and hence, a patient's pharmacologic response may change over the course of their treatment (44); future approaches may need to account for the effects of habituation. Additionally, future work in this domain would benefit by accounting for other factors that confound sedation in the ICU environment including adjunct therapies such as clonidine, ketamine, volatile anesthetics, and neuromuscular blockers. We validated our model based on an assumption that clinicians were dosing patients with an intention to achieve the target sedation level (as defined by ICU protocols). However, this could be untrue in some cases; for example, some procedures performed in the ICU require a deeper sedation level, which contradicts our assumption of keeping patients in light sedation. We believe that combining our model with the *clinician-in-loop* paradigm presented by (11) may help address this issue in future works.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data can be found in: https://mimic.physionet.org/.

AUTHOR CONTRIBUTIONS

NE, TA, and MG contributed to the design and implementation of the research. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Yu C, Liu J, Nemati S. Reinforcement learning in healthcare: a survey. *arXiv preprint.* (2019) *arXiv*:1908.08796.
2. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* (2018) *24*:1716–20. doi: 10.1038/s41591-018-0213-5
3. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Lehman LWH, et al. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association (2018). p. 887.
4. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment—a deep reinforcement learning approach. *arXiv preprint.* (2017) *arXiv*:1705.08422.
5. Yu C, Ren G, Liu J. Deep inverse reinforcement learning for sepsis treatment. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. Beijing: IEEE (2019). p. 1–3.
6. Lowery C, Faisal AA. Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control. In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. San Diego, CA: IEEE (2013). p. 1414–17.
7. Padmanabhan R, Meskin N, Haddad WM. Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy. In: Taher A, editor. C*ontrol Applications for Biomedical Engineering Systems*. Academic Press (2020). p. 251–97.
8. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inform Decis Mak.* (2020) 20:1–8. doi: 10.1186/s12911-020-1120-5
9. Ghassemi MM, Alhanai T, Westover MB, Mark RG, Nemati S. Personalized medication dosing using volatile data streams. In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans, LA) (2018).
10. Lin R, Stanley MD, Ghassemi MM, Nemati S. A deep deterministic policy gradient approach to medication dosing and surveillance in the ICU. In*: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE (2018). p. 4927–31.
11. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Lake Buena Vista, FL: IEEE (2016). pp. 2978–81.
12. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. *J Med Internet Res.* (2020) 22:e18477. doi: 10.2196/18477
13. Reade MC, Finfer S. Sedation and delirium in the intensive care unit. *N Engl J Med.* (2014) 370:444–54. doi: 10.1056/NEJMra1208705
14. Haddad WM, Chellaboina V, Hui Q. *Nonnegative and Compartmental Dynamical Systems*. Princeton: Princeton University Press (2010).
15. Haddad WM, Bailey JM, Gholami B, Tannenbaum AR. Clinical decision support and closed-loop control for intensive care unit sedation. *Asian J Control.* (2013) 15:317–39. doi: 10.1002/asjc.701
16. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint.* (2017) *arXiv*:1704.06300.
17. Hughes CG, McGrane S, Pandharipande PP. Sedation in the intensive care setting. *Clin Pharmacol.* (2012) 4:53. doi: 10.2147/CPAA.S26582
18. Maheshwari R, Sharma P, Seth A, Taneja N, Tekade M, Tekade RK. Drug Disposition Considerations in Pharmaceutical Product. In: Tekade RK, editor. *Dosage Form Design Considerations*. Academic Press (2018). p. 337–69.
19. Bielinski SJ, Olson JE, Pathak J, Weinshilboum RM, Wang L, Lyke KJ, et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time—using genomic data to individualize treatment protocol. *Mayo Clin Proc.* (2014) 89:25–33. doi: 10.1016/j.mayocp.2013.10.021

20. Padmanabhan R, Meskin N, Haddad WM. Optimal adaptive control of drug dosing using integral reinforcement learning. *Math Biosci.* (2019) 309:131–42. doi: 10.1016/j.mbs.2019.01.012
21. Borera EC, Moore BL, Doufas AG, Pyeatt LD. An adaptive neural network filter for improved patient state estimation in closed-loop anesthesia control. In: *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*. Boca Raton, FL: IEEE (2011). p. 41–6.
22. Sinzinger ED, Moore B. Sedation of simulated ICU patients using reinforcement learning based control. *IJAIT.* (2005) 14:137–56. doi: 10.1142/S021821300500203X
23. Moore BL, Quasny TM, Doufas AG. Reinforcement learning versus proportional–integral–derivative control of hypnosis in a simulated intraoperative patient. *Anesth Analg.* (2011) 112:350–9. doi: 10.1213/ANE.0b013e318202cb7c
24. Moore BL, Sinzinger ED, Quasny TM, Pyeatt LD. May. Intelligent control of closed-loop sedation in simulated ICU patients. In: *Flairs Conference* (Miami Beach, FL) (2004). p. 109–14.
25. Sadati N, Aflaki A, Jahed M. Multivariable anesthesia control using reinforcement learning. In: *2006 IEEE International Conference on Systems, Man and Cybernetics.* Vol. 6). Taipei: IEEE (2006). p. 4563–8.
26. Padmanabhan R, Meskin N, Haddad WM. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed Signal Process Control.* (2015) 22:54–64. doi: 10.1016/j.bspc.2015.05.013
27. Moore BL, Doufas AG, Pyeatt LD. Reinforcement learning: a novel method for optimal control of propofol-induced hypnosis. *Anesth Analg.* (2011) *112*:360–7. doi: 10.1213/ANE.0b013e31820334a7
28. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inform Decis Mak.* (2019) 19:57. doi: 10.1186/s12911-019-0763-6
29. Sessler CN, Varney K. Patient-focused sedation and analgesia in the ICU. *Chest*. (2008) 133:552–65. doi: 10.1378/chest.07-2026
30. Barr J, Donner A. Optimal intravenous dosing strategies for sedatives and analgesics in the intensive care unit. *Crit Care Clin.* (1995) 11:827–47. doi: 10.1016/S0749-0704(18)30041-1
31. Sun H, Nagaraj SB, Akeju O, Purdon PL, Westover BM. July. Brain monitoring of sedation in the intensive care unit using a recurrent neural network. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE (2018). p. 1–4.
32. Riker RR, Fraser GL, Simmons LE, Wilkins ML. Validating the Sedation-Agitation Scale with the Bispectral Index and Visual Analog Scale in adult ICU patients after cardiac surgery. *Intens Care Med.* (2001) 27:853–8. doi: 10.1007/s001340100912
33. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 0.4). *PhysioNet*. (2020) doi: 10.13026/a3wn-hq05
34. Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. (2000) 101:e215–20. doi: 10.1161/01.CIR.101.23.e215
35. Oddo M, Crippa IA, Mehta S, Menon D, Payen JF, Taccone FS, et al. Optimizing sedation in patients with acute brain injury. *Crit Care.* (2016) 20:128. doi: 10.1186/s13054-016-1294-5
36. Hariharan U, Garg R. Sedation and Analgesia in Critical Care. *J Anesth Crit Care Open Access.* (2017) 7:00262. doi: 10.15406/jaccoa.2017.07.00262
37. Gross JB, Bailey PL, Connis RT, Coté CJ, Davis FG, Epstein BS, et al. Practice guidelines for sedation and analgesia by non-anesthesiologists. *Anesthesiology.* (2002) 96:1004–17. doi: 10.1097/00000542-200204000-00031
38. Jagannatha A, Thomas P, Yu H. *Towards high confidence off-policy reinforcement learning for clinical applications*. In: *CausalML Workshop, ICML* (Stockholm) (2018).
39. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. *arXiv preprint.* (2015). *arXiv:*1509.02971.
40. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller, M. Deterministic policy gradient algorithms. *PMLR.* (2014) 32:387–95. Available online at: http://proceedings.mlr.press/v32/silver14.html
41. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning. *arXiv preprint.* (2013) *arXiv:*1312.5602.
42. Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. *arXiv preprint.* (2015) *arXiv:*1511.05952.
43. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint.* (2014) *arXiv:*1412.6980.
44. Fan SZ, Wei Q, Shi PF, Chen YJ, Liu Q, Shieh JS. A comparison of patients' heart rate variability and blood flow variability during surgery based on the Hilbert–Huang Transform. *Biomed Signal Proces.* (2012) 7:465–73. doi: 10.1016/j.bspc.2011.11.006
45. DeMers D, Wachs D. Physiology, mean arterial pressure. In: Dulebohn S, editor. *StatPearls*. StatPearls Publishing (2020).

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 Digital Health

REVIEW
published: 25 June 2021
doi: 10.3389/fdgth.2021.645232

Accessing Artificial Intelligence for Clinical Decision-Making

Chris Giordano[1]*, *Meghan Brennan*[1], *Basma Mohamed*[1], *Parisa Rashidi*[2], *François Modave*[3] and *Patrick Tighe*[1]

[1] Department of Anesthesiology, University of Florida College of Medicine, Gainesville, FL, United States, [2] J. Clayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, United States, [3] Department of Health Outcomes & Biomedical Informatics, University of Florida College of Medicine, Gainesville, FL, United States

Advancements in computing and data from the near universal acceptance and implementation of electronic health records has been formative for the growth of personalized, automated, and immediate patient care models that were not previously possible. Artificial intelligence (AI) and its subfields of machine learning, reinforcement learning, and deep learning are well-suited to deal with such data. The authors in this paper review current applications of AI in clinical medicine and discuss the most likely future contributions that AI will provide to the healthcare industry. For instance, in response to the need to risk stratify patients, appropriately cultivated and curated data can assist decision-makers in stratifying preoperative patients into risk categories, as well as categorizing the severity of ailments and health for non-operative patients admitted to hospitals. Previous overt, traditional vital signs and laboratory values that are used to signal alarms for an acutely decompensating patient may be replaced by continuously monitoring and updating AI tools that can pick up early imperceptible patterns predicting subtle health deterioration. Furthermore, AI may help overcome challenges with multiple outcome optimization limitations or sequential decision-making protocols that limit individualized patient care. Despite these tremendously helpful advancements, the data sets that AI models train on and develop have the potential for misapplication and thereby create concerns for application bias. Subsequently, the mechanisms governing this disruptive innovation must be understood by clinical decision-makers to prevent unnecessary harm. This need will force physicians to change their educational infrastructure to facilitate understanding AI platforms, modeling, and limitations to best acclimate practice in the age of AI. By performing a thorough narrative review, this paper examines these specific AI applications, limitations, and requisites while reviewing a few examples of major data sets that are being cultivated and curated in the US.

Keywords: data curation, decision making, deep learning, artificial intelligence, electronic health record, machine learning

OPEN ACCESS

Edited by:
Ira L. Leeds,
Johns Hopkins University,
United States

Reviewed by:
Paraskevi Papadopoulou,
American College of Greece, Greece
Tyler John Loftus,
University of Florida, United States

***Correspondence:**
Chris Giordano
cgiordano@anest.ufl.du

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

Received: *22 December 2020*
Accepted: *01 June 2021*
Published: *25 June 2021*

Citation:
Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F and Tighe P (2021) Accessing Artificial Intelligence for Clinical Decision-Making. Front. Digit. Health 3:645232. doi: 10.3389/fdgth.2021.645232

INTRODUCTION

Healthcare systems around the world have rapidly and pervasively adopted electronic health record (EHR) systems. Many countries report adoption rates higher than 90%, and the US is among this group with a reported 96% use as of 2017 (1–3). Currently, nearly 80% of all US office-based physicians have also adopted an EHR system to satisfy the specifications and requirements set forth by the US Department of Health and Human Services for such systems (4). The resulting underlying databases created by EHR systems contain large heterogeneous data sets that combine structured and formatted data elements such as diagnoses (International Classification of Diseases-10), procedures (Current Procedural Terminology® code), and medications (RxNorm), but also rich unstructured data such as clinical narratives, which represent over 80% of the data in EHRs (5).

Large healthcare systems realized the importance of this data early on and created data warehouses, now used both for research purposes and guiding evidence-based clinical practice. Such data warehouses not only contain EHR data, but also are often enriched with claims data, imaging data, "omics"-type data (e.g., genetic variants associated with a disease or a specific drug response), patient-generated data such as patient-reported outcomes (Patient-Reported Outcomes Measurement Information System®) (6) and wearable-generated data (e.g., nutrition, at-home vitals monitoring, physical activity status) from smartphones and watches. One example of the warehousing of large clinical data for research is the OneFlorida Clinical Research Consortium (7), funded by the Patient-Centered Outcomes Research Institute (PCORI). The OneFlorida Clinical Research Consortium is one of nine clinical data research networks funded by PCORI and aggregates, which harmonizes clinical data from 12 healthcare organizations that care for nearly 15 million Floridians in 22 hospitals and 914 clinical practices across all 67 counties of the state of Florida. This data repository functions alongside additional data warehouses that connect to larger systems that share healthcare data across different countries. The phenomenon of data sharing in healthcare is worldwide. For instance, the European Medical Information Framework (EMIF) contains EHR data from 14 countries, harmonized into a common data model to facilitate cohort discovery and research. With virtually unlimited capacity for data storage and advances in computational power for data analysis, the bottleneck is now in the development of appropriate methods to discover new knowledge to improve care.

Artificial intelligence (AI) methods, in particular machine learning (ML), reinforcement learning, and deep learning, are particularly well-suited to deal with both the data type and looming questions in healthcare. AI can aide physicians in the complex task of risk stratifying patients for interventions, identifying those most at risk of imminent decompensation, and evaluating multiple small outcomes to optimize overall patient outcomes. Integrating physicians into model development and educating physicians in this field will be the next paradigm shift in medical education. For example, the complexity of AI methodologies varies greatly, in turn impacting the ease of physician understanding and interpretation of results. Physicians frequently use decision trees as tools; however, they are effectively tied to the initial tree structure and thus somewhat static (8). On the other hand, deep learning models such as convolution neural networks are less easily interpretable, and may make it more difficult to establish a causal link (9); thus, the development of such models requires the active involvement of clinicians (10). Neural networks commonly used to decipher images collected from patients coupled with the corresponding interpretations often require involvement from radiologists to curate appropriate imaging data for training (11). A priori discussions by AI developers and medically informed physicians are necessary to define the levels of accuracy and interpretability that are required in each clinical context.

Despite methodological, societal, and ethical concerns (12), big data methods are being broadly adopted in healthcare systems for evidence-based clinical decision-making. In this paper, we discuss some of the major opportunities for how AI can assist healthcare workers in clinical decision-making. To prepare for this disruptive innovation, certain facets of medicine will be impacted earlier and more substantially than others. In this paper, we performed an narrative review of specific aspects of healthcare that we predict will most likely be first impacted by AI and how that impact can influence everyday clinical practice. Furthermore, this review includes the potential risks incurred by adopting AI as well as the requisite educational curricula changes and knowledge base needed to avert biases and prevent unsound decision-making.

METHODOLOGY

We performed a comprehensive literature search using the databases PubMed, EMBASE, and Cochrane Review using the keywords (including alternative keywords): artificial intelligence, machine learning, deep learning, perioperative medicine, perioperative clinical decision making, preoperative risk stratification, machine learning and multi-objective optimization, machine learning and warning, machine learning and bias, and machine learning in medical education. Literature search included articles published between 2010 and 2020. Inclusion criteria were articles that focused on adult surgical patients, randomized controlled trials, observational studies, review articles, systematic reviews, and meta-analyses. Exclusion criteria were articles that focused on non-surgical encounters, editorials, letters to the editors, commentaries, books and book chapters, conference proceeding, and pediatric surgical patients. The scope of this review is perioperative clinical decision-making, including settings in the intensive care unit. In addition, we highlight the impact of AI on the future of medical education.

RESULTS

An overview of the study's methodology and results is presented in **Figure 1**. The literature search yielded 1,072 abstracts, of which 185 were duplicates. The authors screened 887 abstracts

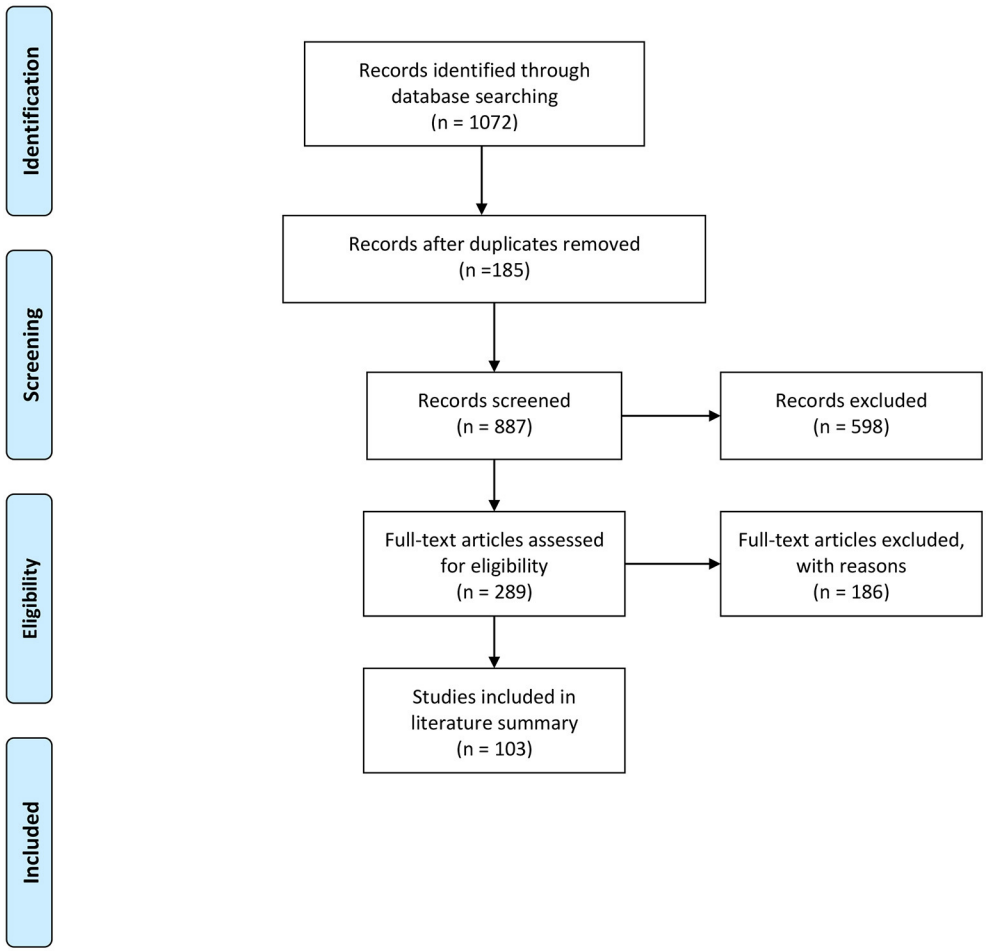


FIGURE 1 | PRISMA flow diagram of accessing artificial intelligence for clinical decision-making.

and 589 were excluded based on the above exclusion criteria. The authors reviewed 289 full articles for eligibility and 186 articles were excluded because they did not meet inclusion criteria. The literature summary focused on 103 full articles. Upon completion of the literature review, we found that there were five main themes related to the role of machine learning, artificial intelligence, and clinical decision-making. The ever-increasing applications of AI methods and tools have potential in nearly every aspect of the clinical decision-making process. In this review, the scope was narrowed to three main promising AI application areas, the potential risks of implementation, and the requisite need for additional education. Specifically, the areas of application include: (1) risk stratification, (2) patient outcome optimization, (3) early warning of acute decompensation, (4) potential bias in ML, and (5) future medical training. These five areas were chosen based on consensus among the authors, who are familiar with recent literature and currently work and research within the AI space. Additionally, these areas reflect contemporary discussion points among clinicians, scientists, engineers, and policymakers given the continued public health burdens of acute illness, as well as the readily available detailed time series data for many at-risk patients. For a more detailed and granular review of AI and deep learning application, which is outside the aims of this review, please see (13, 14).

DISCUSSION

Risk Stratification

ML models that can risk-stratify patients in preparation for surgery will help clinicians identify high-risk patients and optimize resource use and perioperative decisions. ML and AI can help clinicians, patients, and their families efficiently process all available data to generate informed, evidence-based recommendations and participate in shared decision-making to identify the best course of action. ML algorithms can be incorporated into several areas across the spectrum of care, either for disease management or in perioperative settings (15). Risk-prediction models have been used in healthcare practice to identify high-risk patients and to make appropriate subsequent clinical decisions. Appropriate risk stratification should result in proper resource use in this era of value-based care. Most risk-prediction tools are historically built

based upon statistical regression models. Examples include the Framingham risk score, QRISK3 (for coronary heart disease, ischemic stroke, and transient ischemic attack), and National Surgical Quality Improvement Program (NSQIP). Unfortunately, many of these risk stratification methods are either non-specific and lack patient-level precision or require trained clinicians to review the records and specifically assess the risk. Healthcare systems have increasingly sought to use ML to assist in risk stratification, and these ML models may outperform statistical models in calibration and discrimination. A growing nationwide effort is seeking to enhance preoperative and perioperative support for high-risk patients and high-cost populations (16, 17). Preoperative evaluation clinics focusing on evaluating high-risk patients have shown improvement in 30-day postoperative outcomes (18). However, identifying these patients is challenging because of the difficulty in timely access to patient data coupled with the lack of robust predictive models. Many traditionally used models have been created to predict postoperative complications but with limited applicability at an individual patient level. Any predictive risk score is dependent on the underlying data and the technology used to process the data. In order to create a better prediction, high-quality, continuous data from multiple domains are required. Also, advancements in health data processing, biosensors, genomics, and proteomics will help provide a complete set of data that will enable perioperative intelligence (19). Furthermore, risk stratification is not limited to the preoperative setting. Incorporating intraoperative data for early detection of complications or clinical aberrations could also prevent inflammatory reactions that exacerbate the injury or high-risk interventions that may lead to iatrogenic injuries. Therefore, clinicians can use ML technology to build proactive systems to avoid these potentially destructive processes.

Multiple ML models that risk-stratify patients with a disease or prepare patients for surgery have been recently developed and validated (16, 20–25). These ML models have been shown to better predict mortality than conventional logistic regression after liver cancer surgery, aortic aneurysm surgery, and cardiac surgery. Other ML models have also been developed and validated to predict the risk of super-utilization and plan accordingly, starting in the preoperative setting in an increased effort to enhance value-based care (17). ML models to predict perioperative risk need to be accurate, locally calibrated, and clinically accessible. Changes in patient condition throughout the perioperative period can be included to update the risk assessment. The advantage of ML models in risk prediction is its automation capability, which is less burdensome compared to current tools (e.g., NSQIP). ML models allow for continuous recalculation of risk longitudinally over time, which can act as early-warning systems alerting clinicians to sudden changes. Incorporation of intraoperative data and interventions, such as hypotension, enable further interventions that enhanced recovery after surgery pathways emphasize. Another advantage is the promise that the use of ML in medicine will facilitate an understanding of what features drive outcomes (26). In perioperative medicine, ML can maximize the benefits of technology to provide safe, timely, and affordable healthcare. The key is integration of all data-generating platforms throughout all phases of patient care with collaboration to identify risks, detect complications early, and offer timely treatment (19).

Patient Outcome Optimization

Optimization for each or the multiple potential patient outcomes is vital to the clinical decision-making process and the ensuing patient care. Typically, the requisite optimal steps, their timing, and the best sequence are determined by healthcare providers in consultation with family members. Despite best intentions, such decisions occasionally lead to suboptimal care due to the complexity of patient care, the increasing responsibilities of healthcare providers, or simply because of human error. The clinical decision-making process is often strictly based on standard guidelines and protocols that satisfy safety and accountability requirements. However, deviation from established protocols in complex care environments can be beneficial for the patient to adapt treatments for a more personalized regimen. In such dynamic settings, ML methods can be valuable tools for optimizing patient care outcomes in a data-driven manner, especially in acute care settings. ML and modern deep-learning techniques typically optimize an objective function (e.g., medication dosage) based on complex and multidimensional data (e.g., patient medical history extracted from EHRs). ML tools for optimizing care outcomes have been used in various settings, including critical care for optimizing sepsis management (27), management of chronic conditions (28), and optimizing surgical outcomes (29). Optimizing patient outcomes can be based on relatively simple yet efficient tools, such as decision trees in conjunction with the domain expertise to systematically codify accepted understanding of disease models and common treatments for patients. Although helpful in assisting with single-step decisions, these tools fail to consider the importance of sequential decision-making, which include many decisions that are dependent on previous actions.

Another more sophisticated approach is to use sequential decision-making tools that draw inspiration from related fields, such as operation research. For example, deep reinforcement learning models (30, 31) are based on well-known concepts such as the Markov decision process (MDP) (32) and Q-learning (33) adapted to neural networks. Reinforcement learning models learn to identify optimal policies based on a reward function. The policies are defined as a series of actions that culminate in the greatest reward, hence identifying the optimal policy. Recently, reinforcement learning and deep reinforcement learning have been used in several clinical settings, including optimal dosing and choice of medications, optimal timing of interventions, and optimal individual target laboratory values (34). For example, Nemati et al. used deep reinforcement learning to optimize medication dosing (35), and Prasad et al. (36) used a reinforcement learning approach to weaning mechanical ventilation in the intensive care unit. Although such tools hold great potential in optimizing the patient care process, safety and accountability is paramount. This could be complicated by the black-box nature of modern deep-learning approaches. The resulting policies may be dynamic and personalized, but their rationale may be challenging to interpret and explain. Additionally, unlike typical simulation

and gaming environments, applying reinforcement learning in clinical settings is much more challenging. It is not trivial to identify the most suitable reward structure, and the effects of treatments can be non-deterministic. In such settings, it is difficult to solve the credit assignment problem, i.e., to demonstrate that deviations from the protocol based on a reinforcement learning suggestion were beneficial for the patient. Future approaches also could examine different time scales. For example, although early interventions (e.g., early antibiotics) may not lead to immediate improvements, they could culminate in the greatest ultimate reward (e.g., higher survival rate).

Patient outcome optimization such as reinforcement learning methods can ultimately provide a tool to help standardize care at health systems of different scales. This could provide a more equitable healthcare system, especially in rural and remote settings.

Early Warning of Acute Decompensation

Acute decompensation is uncommon, but it is typically accompanied by increasing physiologic derangements and worse outcomes. Intervening early may mitigate poor outcomes; however, it is often difficult to identify this patient population before significant hemodynamic compromise with our traditional standard monitoring and commonly used early-warning scores. Six to eight hours may precede such acute patient decompensation, which can easily provide ample time for interventions to be made (37). The EHR contains a large amount of data that may be useful to identify patients at the highest risk of decompensation if the data are evaluated over time (37, 38). Multivariate regression-based models or AI-based early-warning systems have the potential to detect subtle trends in physiologic parameters over time to provide precision and reliability (38–41).

Vital sign monitoring and associated alarms were one of the earliest methods to detect patient decompensation (40). They are effective in alerting providers to discrete vital sign abnormalities in real time; however, early or isolated vital sign abnormalities also may fail to signal to providers an impending decompensation (40). Once it becomes evident that a patient is decompensating, the initial response is often directed toward correcting one or more abnormalities until an etiology is determined. The Modified Early Warning Score (MEWS), Rothman Index, Sequential Organ Failure Assessment Score (SOFA), and quick SOFA (qSOFA) were developed to incorporate multiple vital sign abnormalities to identify at-risk patients before decompensation occurs. The drawbacks to these scores are that even if they are automated and incorporated into the medical record, they rely on discrete data points of pre-existing vital sign changes and are subject to reporting error. Additionally, because of their high sensitivity but low discriminatory ability, these scores identify a large number of patients as "at risk" when the actual number is far lower (38). Furthermore, because interventions often involve their own risks, they may not be implemented until it becomes clear that a patient's condition is rapidly deteriorating. At that point, immediate and possibly emergent interventions that are themselves high risk and invasive must be performed. Preventative measures may be taken earlier and with more accuracy if AI metrics are implemented as opposed to the traditional risk-evaluation scores (40, 42). AI-based monitoring incorporated into the EHR can facilitate the use of large volumes of data for all patients more efficiently and precisely than a physician could, enabling AI to identify patients who are most at risk.

The operating room may be one of the most challenging areas for early detection, workup, and treatment of acute decompensation. The Hypotension Prediction Index (HPI; Edwards Lifesciences, Irvine, CA) is an algorithm created to aid in the early detection of intraoperative hypotension, defined as mean arterial pressure <65 mmHg for non-cardiac surgeries (41, 43, 44). It is now incorporated into the Edwards monitoring system. It was developed using an ML, logistic regression-based model analyzing components of the arterial waveform (41, 43, 44). One advantage is that in addition to early notification of hypotension, this tool also identifies some of the most likely causes for the predicted hypotensive event, e.g., vasoplegia, hypovolemia, or possibly conditions related to cardiac contractility. Initial studies, although small, single center, and not without bias, indicate that the HPI and implementation of the monitor were reasonably effective in preventing clinically significant hypotensive events. Although developed with AI, this monitor and associated alarm rely on the data that it was trained and developed on, and they do not learn and adapt with each patient.

Using AI to effectively create an early-warning score using time series data from the EHR presents many challenges. An ideal score would identify patients before an obvious decompensation. It would have excellent discriminatory ability so that physicians would have confidence implementing appropriate interventions as well as transparency to identify the sources of risk and the reasons for decompensation. Incorporating appropriate treatments and their effects on risk reduction remains a weakness of all existing early-warning systems. AI-based algorithms using time series data from the EHR are in development with strong results. Shickel et al. used a modified recurrent neural network model on temporal intensive care unit data to develop deepSOFA, a real-time mortality risk prediction score based on the traditional SOFA score (38). Its predictive ability performed well in identifying increased risk of mortality. Lauritsen et al. developed the explainable AI early-warning score (xAI-EWS). It is meant to be incorporated into the EHR and uses a temporal convolutional network and deep Taylor explanation model to provide predictions. It has demonstrated feasibility using predictions for risk of acute injury, sepsis, and acute lung injury (39).

Potential for Bias in ML

As AI becomes more pervasive in both public and personal health across diverse populations, there have been increasing concerns, and related examples, of AI solutions leading to inadvertent bias of modeling results (45–48). Broadly, such bias can originate from the data used for model training and testing, as well as the mechanics of the model itself (49). Bias originating from data can be pernicious; for instance, work by Weber et al. found that simply filtering for "complete" EHRs, a common strategy for

managing missing data, introduced a bias toward older patients who were more likely female (50).

Less pernicious examples include reference imaging datasets in which more than 80% of subjects were light-skinned individuals (51). With respect to modeling mechanics, the non-linearities, extensive interactions among variables, and difficulties interpreting how ML models arrived at their results, ML also presents many new challenges to addressing sources of inadvertent bias that differ from classifiers that enforce linear models of independent variables in smaller, more manageable datasets. Under the rubric of decision support, an unfair algorithm has been defined as "one whose decisions are skewed toward a particular group of people" (49). Verma and Rubin have clarified several definitions of algorithmic fairness, where definitions are based on objective probabilistic assessments (52). These definitions help provide a platform for promoting algorithmic fairness by creating neutral models through approaches addressing anti-classification, classification parity, and model calibration on protected attributes (53). Notably, these solutions may present their own ethical issues. McCradden et al. (54) suggest that some solutions to algorithmic fairness can instead reinforce health inequities and even exacerbate harms to vulnerable groups. Until more robust solutions to the challenges of algorithmic fairness can be identified and implemented, physicians should remain vigilant for how ML models, built on training samples from general populations, may be misapplied to their own patients. This appreciation of ML building and application will require a new level of professional development and commensurate medical education curricula, which will be discussed in the next section.

Paradigmatic Shift in Medical Training

Applying advances in biomedical informatics and ML models to patient care will require clinicians to reconsider their educational training and infrastructure. Wartman et al. noted that the practice of medicine is transitioning from the Age of Information to the Age of AI (55). Traditionally, medical curriculum has been founded on memorizing a massive curriculum, applying it to a learned clinical experience, and determining the validity of ensuing information as it becomes published. Similarly, understanding principles of normal variants of anatomy and physiology, followed by an examination of pathophysiologic variants, presents students with a model-based rubric in which to incorporate each new wave of information learned through personal experience as well as throughout the medical literature. This paradigm has also permitted physicians to extrapolate previous understanding by logic and experience to novel diagnostic reasoning and therapeutic approaches by extension of previous models. However, the amount of information has become insurmountable. The time for medical information to double was 50 years in 1950, 7 years in 1980, 3.5 years in 2010, and a staggering estimate of 73 days in 2020 (56). Humans are not only incapable of this level of exposure or retention, but the magnitude has also created substantial levels of stress-induced mental illness among learners (57). Fortunately, advances in biomedical informatics point to new approaches that can seamlessly synthesize old and new medical information. These advances will provide the foundation for AI advances to recognize patterns of patient information to help diagnose, treat, and manage patients. This transition will require the development of new knowledge, skills, and attitudes by healthcare workers. Furthermore, it will require a rethinking of the medical school curriculum, in which new data analytics methods are carefully integrated with traditional medical education. In an extremely busy curriculum and at a time of numerous other considerations, such as climate change (58), incorporating AI will present challenges.

Many of the AI subfields such as ML and deep learning use complex algorithms that generate outputs from seemingly opaque non-linear functions that most physicians likely find difficult to understand or incorporate into their existing approaches to evidence-based medicine. Subsequently, this black-box phenomena (10) will be difficult for physicians to trust, and it will also be a challenge for the doctor-patient relationship since many physicians will find themselves unable to explain the diagnosis, prediction, or therapy (59). This challenge will increase with the stakes and timeliness of the given issue; for instance, outcome assessments involving the withdrawal of care may pose heightened anxiety regardless of the model's accuracy. Therefore, physicians will need to develop a basic understanding of how input data are aggregated, analyzed, and generated into specific pathways of care for individualized patients. Furthermore, these algorithms will require physicians to have a better understanding of calculus and linear algebra, manipulation of data sets (curation, provenance, quality, integration, and governance), and model performance metrics fundamental in grading AI algorithmic decision-making. This knowledge will allow physicians to recognize when AI algorithms are being used on inappropriate patient populations, when AI tools have become outdated and need updating, or when aggregated data is biased. These new AI clinical decision-support systems have limitations in their application to patient populations, contextual changes, and therapeutic variances that will require a stronger appreciation of probabilities and confidence ratings (59). It will also be important to understand when physicians are justified in deviating from AI-inspired treatment protocols. Physicians will need to update their understanding of evidence-based medicine principles to include modern approaches to analyzing and assessing causality to ensure a robust understanding of how patients, social determinants of health, and healthcare systems interact to inform health-related outcomes. Physicians practicing in the age of AI should be competent in the effective integration and data use that emerges from an endless array of sources.

The emerging need for understanding how AI data platforms function and generate predictions is juxtaposed with the ever-important traditional need for communication skills, empathy, and teamwork. Translating the predictions from complex AI algorithms into meaningful and personalized information for patients will require strong communication skills as well as compassion. Compounding this resurgence of social skills requisition for medical practice will be the application of cognitive psychology principles. Understanding this need for social skills will help identify biases and heuristics that impact decision-making as well as help physicians frame

choices, understand context, and have neutral but meaningful conversations with patients (55).

CONCLUSIONS

The ubiquitous adoption of EHRs in healthcare systems around the world has created vast repositories of personalized data sets that are perfectly fitted for AI to examine, develop, and predict upon. The subfields of ML and deep learning networks have shown success in providing solutions to the healthcare questions of risk stratification and optimizing patient outcomes. Use of this technology will exponentially expand as it is increasingly integrated into large healthcare systems. AI capabilities will aide physicians in weighing competing healthcare goals and numerous risks by facilitating multiple outcome optimization of outcomes that are too difficult to recognize and navigate on an individual and isolated basis. Healthcare workers will be expected to comfortably work within this new AI frontier and in turn relate it to their patients. Furthermore, physicians must be able to interpret the predictions of these AI algorithms as well as deconstruct the models from which they ebb. In addition, physicians will need to recognize plausible bias and the appropriate patient population application that stems from understanding the training cohort used to create the model. This understanding will require additional medical education and professional development for current practitioners and a revamped curriculum for all new learners currently in medical school. Most importantly, physicians must maintain and cultivate emotional intelligence and compassion when relaying results and recommending interventions from these complex models to uncertain and vulnerable patients who want to make informed decisions for themselves or a family member's well-being.

AUTHOR CONTRIBUTIONS

CG, MB, BM, PR, FM, and PT contributed substantially to all aspects of the work, including conception, literature review, manuscript drafting, critical revision, and agreed to be accountable for all aspects of the work. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Donn M. Dennis, MD, Professorship in Anesthetic Innovation 386 (PT) and by NIH 5R01GM114290 (PT) and 5R01AG055337 (PT). PR was partially supported by the National Science Foundation CAREER award 1750192 and 1R21EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB).

REFERENCES

1. Becker's Health IT. *Top 10 Countries for EHR Adoption.* (2013). Available online at: https://www.beckershospitalreview.com/healthcare-information-technology/top-10-countries-for-ehr-adoption.html (accessed March 1, 2021).
2. The Health Institute for E-Health Policy. *A Glimpse at EHR Implementation Around the World: The Lessons the US Can Learn.* (2014). Available online at: https://www.e-healthpolicy.org/sites/e-healthpolicy.org/files/A_Glimpse_at_EHR_Implementation_Around_the_World1_ChrisStone.pdf (accessed March 1, 2021).
3. Office of the National Coordinator for Health Information Technology. *Non-federal Acute Care Hospital Electronic Health Record Adoption*. Health IT Quick-Stat #47 (2017). Available online at: https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php (accessed February 16, 2020).
4. Office of the National Coordinator for Health Information Technology. *Office-Based Physician Electronic Health Record Adoption*. Health IT Quick-Stat #50 (2019). Available online at: https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php (accessed February 16, 2020).
5. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* (2008) 17:128–44. doi: 10.1055/s-0038-1638592
6. HealthMeasures. *Transforming How Health Is Measured*. PROMIS®. Available online at: https://www.healthmeasures.net/explore-measurement-systems/promis (accessed December. 22, 2020).
7. Shenkman E, Hurt M, Hogan W, Carrasquillo O, Smith S, Brickman A, et al. OneFlorida Clinical Research Consortium: linking a clinical and translational science institute with a community-based distributive medical education model. *Acad Med.* (2018) 93:451–55. doi: 10.1097/ACM.0000000000002029
8. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. *Neural Comput Appl.* (2013) 23:2387–403. doi: 10.1007/s00521-012-1196-7
9. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell.* (2020) 2:369–75. doi: 10.1038/s42256-020-0197-y
10. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med.* (2020) 172:59–60. doi: 10.7326/M19-2548
11. Martín Noguerol T, Paulano-Godino F, Martín-Valdivia MT, Menias CO, Luna A. Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. *J Am Coll Radiol.* (2019) 16(Pt. B):1239–47. doi: 10.1016/j.jacr.2019.05.047
12. Prosperi, M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak.* (2018) 18:139. doi: 10.1186/s12911-018-0719-2
13. Panesar A. *Machine Learning and AI for Healthcare*. Coventry: Apress (2019).
14. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform.* (2018) 22:1589–604. doi: 10.1109/JBHI.2017.2767063
15. Debnath S, Barnaby DP, Coppa K, Makhnevich A, Ji Kim E, Chatterjee S, et al. Machine learning to assist clinical decision-making during the COVID-19 pandemic. *Bioelectron Med.* (2020) 6:14. doi: 10.1186/s42234-020-00050-8
16. Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med.* (2018) 15:e1002701. doi: 10.1371/journal.pmed.1002701
17. Hyer JM, Ejaz A, Tsilimigras DI, Paredes AZ, Mehta R, Pawlik TM. Novel machine learning approach to identify preoperative risk factors associated with super-utilization of medicare expenditure following surgery. *JAMA Surg.* (2019) 154:1014–21. doi: 10.1001/jamasurg.2019.2979

18. McDonald SR, Heflin MT, Whitson HE, Dalton TO, Lidsky ME, Liu P, et al. Association of integrated care coordination with postsurgical outcomes in high-risk older adults: the Perioperative Optimization of Senior Health (POSH) Initiative. *JAMA Surg.* (2018) 153:454–62. doi: 10.1001/jamasurg.2017.5513
19. Maheshwari K, Ruetzler K, Saugel B. Perioperative intelligence: applications of artificial intelligence in perioperative medicine. *J Clin Monit Comput.* (2020) 34:625–8. doi: 10.1007/s10877-019-00379-9
20. Tseng Y-J, Wang H-Y, Lin T-W, Lu J-J, Hsieh C-H, Liao C-T. Development of a machine learning model for survival risk stratification of patients with advanced oral cancer. *JAMA Netw Open.* (2020) 3:e2011768. doi: 10.1001/jamanetworkopen.2020.11768
21. Merath K, Hyer JM, Mehta R, Farooq A, Bagante F, Sahara K, et al. Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery. *J Gastrointest Surg.* (2020) 24:1843–51. doi: 10.1007/s11605-019-04338-2
22. Fernandes MPB, Armengol de la Hoz M, Rangasamy V, Subramaniam B. Machine learning models with preoperative risk factors and intraoperative hypotension parameters predict mortality after cardiac surgery. *J Cardiothorac Vasc Anesth.* (2020) 35:857–65. doi: 10.1053/j.jvca.2020.07.029
23. Wise ES, Hocking KM, Brophy CM. Prediction of in-hospital mortality after ruptured abdominal aortic aneurysm repair using an artificial neural network. *J Vasc Surg.* (2015) 62:8–15. doi: 10.1016/j.jvs.2015.02.038
24. Kwon J-M, Jeon K-H, Kim HM, Kim MJ, Lim S, Kim K-H et al. Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. *PLoS ONE.* (2019) 14:e0224502. doi: 10.1371/journal.pone.0224502
25. Myers PD, Scirica BM, Stultz CM. Machine learning improves risk stratification after acute coronary syndrome. *Sci Rep.* (2017) 7:12692. doi: 10.1038/s41598-017-12951-x
26. Hill BL, Brown R, Gabel E, Rakocz N, Lee C, Cannesson M, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *Br J Anaesth.* (2019) 123:877–86. doi: 10.1016/j.bja.2019.07.030
27. Tsoukalas T, Albertson T, Tagkopoulos I. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Med Inform.* (2015) 3:e11. doi: 10.2196/medinform.3445
28. Siontis KC, Yao X, Pirruccello JP, Philippakis AA, Noseworthy PA. How will machine learning inform the clinical care of atrial fibrillation? *Circ Res.* (2020) 127:155–69. doi: 10.1161/CIRCRESAHA.120.316401
29. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg.* (2018) 109:476–86.e1. doi: 10.1016/j.wneu.2017.09.149
30. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature.* (2015) 518:529–33. doi: 10.1038/nature14236
31. François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J. An introduction to deep reinforcement learning. *arXiv.* (2018) doi: 10.1561/9781680835397
32. Bellman R. A Markovian decision process. *J Math Mech.* (1957) 6:679–84. doi: 10.1512/iumj.1957.6.56038
33. Watkins CJCH. *Learning from delayed rewards* (Ph.D. thesis), University of Cambridge, Cambridge, United Kingdom (1989).
34. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. *J Med Internet Res.* (2020) 22:e18477. doi: 10.2196/18477
35. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Annu Int Conf IEEE Eng Med Biol Soc.* (2016) 2016:2978–81. doi: 10.1109/EMBC.2016.7591355
36. Prasad N, Cheng L-F, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv*. (2017) arXiv:170406300.
37. Taenzer AH, Pyke JB, McGrath SP, Blike GT. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: a before-and-after concurrence study. *Anesthesiology.* (2010) 112:282–7. doi: 10.1097/ALN.0b013e3181ca7a9b
38. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci Rep.* (2019) 9:1879. doi: 10.1038/s41598-019-38491-0
39. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jorgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun.* (2020) 11:3852. doi: 10.1038/s41467-020-17431-x
40. Vistisen, ST, Johnson AEW, Scheeren TWL. Predicting vital sign deterioration with artificial intelligence or machine learning. *J Clin Monit Comput.* (2019) 33:949–51. doi: 10.1007/s10877-019-00343-7
41. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology.* (2018) 129:663–74. doi: 10.1097/ALN.0000000000002300
42. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. *Anesth Analg.* (2020) 130:352–9. doi: 10.1213/ANE.0000000000004121
43. Edwards Lifesciences Corporation. *Acumen Hypotension Prediction Index*. Available online at: https://www.edwards.com/devices/decision-software/hpi (accessed December 22, 2020).
44. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA*. (2020) 323:1052–60. doi: 10.1001/jama.2020.0592
45. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. (2018) 378:981. doi: 10.1056/NEJMp1714229
46. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* (2018) 169:866–72. doi: 10.7326/M18-1990
47. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *Br J Anaesth*. (2020) 125:843–6. doi: 10.1016/j.bja.2020.07.040
48. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* (2019) 366:447–53. doi: 10.1126/science.aax2342
49. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *arXiv*. (2019) arXiv:190809635.
50. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data." *J Am Med Inform Assoc.* (2017) 24:1134–41. doi: 10.1093/jamia/ocx071
51. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY: PMLR (2018). p. 77–91.
52. Verma S, Rubin J. Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. New York, NY: Association for Computing Machinery (2018). p. 1–7.
53. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv*. (2018) arXiv:180800023.
54. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health.* (2020) 2:e221–3. doi: 10.1016/S2589-7500(20)30065-0
55. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med.* (2018) 93:1107–9. doi: 10.1097/ACM.0000000000002044

56. Densen P. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc.* (2011) 122: 48–58.
57. Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. *N Engl J Med.* (2017) 377:1209–11. doi: 10.1056/NEJMp1705348
58. Finkel ML. A call for action: integrating climate change into the medical school curriculum. *Perspect Med Educ.* (2019) 8:265–6. doi: 10.1007/s40037-019-00541-8
59. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ.* (2019) 5:e16048. doi: 10.2196/16048

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, TL, declared to the editor a past collaboration with the authors, and confirms the absence of ongoing collaborations at the time of the review.

frontiers in Digital Health

METHODS
published: 06 September 2021
doi: 10.3389/fdgth.2021.648190

Discovering Composite Lifestyle Biomarkers With Artificial Intelligence From Clinical Studies to Enable Smart eHealth and Digital Therapeutic Services

Sofoklis Kyriazakos[1,2]*, *Aristodemos Pnevmatikakis*[1], *Alfredo Cesario*[1,3], *Konstantina Kostopoulou*[1], *Luca Boldrini*[4], *Vincenzo Valentini*[4,5] *and Giovanni Scambia*[4]

[1] *Innovation Sprint Sprl, Brussels, Belgium,* [2] *Business Development and Technology, Aarhus University, Herning, Denmark,* [3] *Scientific Directorate, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy,* [4] *Advanced Radiation Therapy, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy,* [5] *Università Cattolica del Sacro Cuore, Rome, Italy*

OPEN ACCESS

Edited by:
Ira L. Leeds,
Johns Hopkins University,
United States

Reviewed by:
Amanda Christine Filiberto,
University of Florida, United States
Emmanouil Spanakis,
Foundation for Research and Technology Hellas (FORTH), Greece

***Correspondence:**
Sofoklis Kyriazakos
skyriazakos@innovationsprint.eu

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

Received: *31 December 2020*
Accepted: *27 July 2021*
Published: *06 September 2021*

Citation:
Kyriazakos S, Pnevmatikakis A, Cesario A, Kostopoulou K, Boldrini L, Valentini V and Scambia G (2021) Discovering Composite Lifestyle Biomarkers With Artificial Intelligence From Clinical Studies to Enable Smart eHealth and Digital Therapeutic Services. Front. Digit. Health 3:648190. doi: 10.3389/fdgth.2021.648190

Discovery of biomarkers is a continuous activity of the research community in the clinical domain that recently shifted its focus toward digital, non-traditional biomarkers that often use physiological, psychological, social, and environmental data to derive an intermediate biomarker. Such biomarkers, by triggering smart services, can be used in a clinical trial framework and eHealth or digital therapeutic services. In this work, we discuss the APACHE trial for determining the quality of life (QoL) of cervical cancer patients and demonstrate how we are discovering a biomarker for this therapeutic area that predicts significant QoL variations. To this extent, we present how real-world data can unfold a big potential for detecting the cervical cancer QoL biomarker and how it can be used for novel treatments. The presented methodology, derived in APACHE, is introduced by Healthentia eClinical solution, and it is beginning to be used in several clinical studies.

Keywords: digital biomarkers, machine learning, ai clinical trials, Healthentia, real world data, e-clinical platform

INTRODUCTION

The field of clinical research is undergoing a "data revolution." The transformation of large volumes of medical records to an electronic format, and the remarkable growth in the data collected by health registries and during clinical studies provide opportunities to make risk prediction and intervention selection more precise. This increasing availability of the so-called "Big Data" has brought about a growing interest in machine learning (ML) algorithms for extracting knowledge from observations, typically conceptualized as datasets, and for constructing personalized risk prediction models.

The concepts of real-world data (RWD) and real-world evidence (RWE) have come to be fashionable, along with those that describe outcome and experience from the perspective of the patient [Patient-Reported Outcome Measures (PROMs), Patient-Reported Experience Measures (PREMs)]. The advent of wearable technologies has made the objective measurement of lifestyle possible to an unprecedented scale of dimensionality, and the collection of subjective information about outcome and experience as PROMs and PREMs.

However, in spite of the use of RWD/RWE (and PROMs/PREMs) one important shortcoming in clinical research is that the actual outcome of trials is usually different from the expected one, and often not reproducible (1, 2). Evidence included in the access-to-market dossier of any intervention, pitched to a lesser extent with respect to the expected one, leads to an economic loss in different ways, such as: (i) the tag price agreed upon by regulators at the moment of pricing negotiations drops due to the worse intervention results; (ii) the marketed solution loses competitiveness; and (iii) the overall benefits to the citizens are reduced.

Factors like trial-protocol adherence and compliance, dropout rate, and surveillance of adverse effects have been traditionally outlined as the most significant reasons behind this outcome difference, and several types of interventions aimed at reducing their impact have been put into place (3).

To mitigate unexpected results from clinical trials, along with adherence and compliance issues, digital solutions as clinical diaries have flourished and are vastly used in the running of clinical trials to collect information on how patients are coping through the trial itself by focusing on PROMs. These digital solutions are termed ePRO.

ePROs generally do not take into account the impact of the lifestyle and habits of the patient that can be measured using wearables (the RWD as objectively measured) on the effectiveness of the intervention and focus, instead, on the PROMs and PREMs. In addition to that, lifestyle has not yet been aggregated into actionable predictors, or used to generate simulated models to derive predictors, on outcomes and experiences.

The processing of patient-centered RWD represents an innovative challenge for modern personalized medicine. Today, various patient-well-being dimensions can be satisfactorily met, using merely a multidimensional data collection approach. Data collection platforms, able to collect, manage, and interpret RWD of the patients, eventually supported by artificial intelligence (AI), are fundamental.

In this study, we attempt to establish patient lifestyle and behavior as the driving force of an effective treatment, by addressing the following hypotheses:

1. Objectively-measured RWD correlate to PROMs and PREMs and thus can predict them.
2. The impact of behavior/lifestyle, as expressed by measured RWD, on PROMs and PREMs can be simulated by utilizing biomarkers in models of patients and testing intervention on them.
3. Groups of biomarkers identified *via* simulation of trials lead to behavior/lifestyle phenotypes that can be used as clinical endpoints and eligibility criteria in clinical trials.
4. Coaching on behavior/lifestyle can complete traditional interventions in everyday practice.

This paper is organized around seven sections. Following the paper, after the abstract and introduction, the subjective and objective RWD as clinical outcomes are discussed in section Subjective and Objective RWD as Clinical Outcomes, including definitions and the role of RWD, RWE, and ePROs. section Lifestyle Behavior as a Biomarker With Clinical Value and Types of RWD presents the concept of lifestyle as a biomarker with clinical value, and in section AI Technologies for Defining, Modeling, and Simulating Lifestyle we present how AI can support the discovery and extraction of such biomarkers during clinical studies. Furthermore, in section Pilot Study to Evaluate the Hypothesis, we present elements from a series of clinical studies that utilize the Healthentia eClinical platform (4) to capture insights that can lead to advanced services, and section Expected benefits, Early Findings and Next Steps presents the expected benefits, the early findings, and discusses the next steps. Finally, the conclusions are drawn in section Conclusions.

SUBJECTIVE AND OBJECTIVE RWD AS CLINICAL OUTCOMES

New Extended Meanings for Old Medical Definitions

A biomarker (5) is the value of a quantity that characterizes the outcome (of a disease) or diagnoses a disease stage. Digital biomarkers (5) are biomarkers whose method of collection involves sensors and computational tools implemented in software or hardware. Traditionally, biomarkers have been split into direct and indirect (6). A direct biomarker is a single measurement of one of the factors or products of the disease that allows diagnosing a disease outcome or stage. Direct biomarkers are usually biochemical, measured obtrusively in a lab. An indirect biomarker is also a single measurement, obtained easily using ubiquitous devices, indirectly associated but highly correlated with a factor or product of the disease. For example, body temperature T is an indirect biomarker for flu if:

$$T - 37^{\circ}C > 0$$

Is body temperature the only biomarker for the flu? No, but it certainly is a good standalone one. Is body mass a biomarker for obesity? Consider a person with a body mass of 120 kg. That person is obese only if their height is <2 m, based on their body mass index (7). Any standalone measurement can give indications on some disease, but the full power of measurements comes into effect when they are combined together into a composite biomarker. A composite biomarker (6) is thus the usually non-linear combination of multiple measurements into a single metric used in disease diagnosis or outcome prediction. In simple cases, the combination can be done analytically, e.g., the body mass index already mentioned is a composite biomarker for obesity that non-linearly combines the height h and mass m of a person as:

$$\frac{m}{h^2} - 30 > 0$$

Such simple cases are the exceptions. Usually, there are many measurements forming a vector $\mathbf{x}$ and the biomarker non-linearly combines them into $F(\mathbf{x})$. Discovering the non-linear combination function $F(\ )$ is not done manually, resulting in an equation. Instead, it is done using an ML algorithm that learns $F(\ )$ for the measurements $\mathbf{x}$, yielding the metric to be evaluated

for disease diagnosis or outcome prediction. In ML terminology, x is the feature vector, and $F(\ )$ is the discriminant function of the classifier (8). Hence composite biomarker discovery is about training a classifier or regressor using some ML algorithm.

Finally, a contextualized composite biomarker (6) is the combination of intrinsic factors (that comprise the composite biomarkers) and extrinsic factors, that is, the environment, providing a metric (classifier or predictor output) for personalized disease management.

In sections AI Technologies for Defining, Modeling and Simulating Lifestyle and Pilot Study to Evaluate the Hypothesis of this paper, we are detailing a methodology to discover digital contextualized composite biomarkers in RWD with outcome prediction capabilities.

RWD and RWE: Definition and Usefulness in Clinical Research

The 21st century Cures Act of 2016, a harbinger of the increasing use of electronic health records (EHRs) and insurance claims data for medical research in the United States, required the Food and Drug Administration (FDA) to develop guidance on the use of RWE in the studies of medical product safety and outcomes for both postapproval studies and studies of new indications of approved drugs. Hence, FDA has issued the following definition: RWD are "*data related to patient health status and/or the delivery of health care routinely collected from EHRs, claims and billing data, data from product and disease registries, patient-generated data including home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices.*" (9). The European Medicine Agency address the same items accordingly, in Organization for Economic Co-Operation and development (OECD)/World health Organization (WHO) (10).

Real-world data are analyzed to create RWE, which is clinical evidence about "*the usage, and potential benefits or risks, of a medical product derived from analysis of RWD.*" (9).

Compared with evidence collected in randomized controlled trials, RWE better reflects the actual clinical environments, in which medical interventions are used, including patient demographics, comorbidities, adherence, and concurrent treatments.

When RWE is intended as the data generated in an observational trial, we notice that there is a significant increase in the number and quality of this type of trial with the consequence of a very significant increase of data assets. It is also well-known that when EHRs are used as a source of RWE, the vast majority of this data, up to 80%, is unstructured. Moreover, insurance claims are a rich source of information and they can cover ontologies not normally included in EHRs, such as the experience of the patients, extensive information on comorbidities (that in EHR is normally highly unstructured), etc., but they are not good for measuring disease severity, biomarkers and, in general, detailed clinical information. In this setting of increasing dimensionality and heterogeneity of data, ML methods are gaining traction as tools to analyze massive and complex datasets (11). When RWE coming from EHRs has been analyzed with ML techniques to create predictive models vs. outcome, these have outperformed traditional ones (12) even in real-time analysis executed, for example, in the emergency setting (13). Some concerns are raised, however, regarding the readiness of EHRs systems to support machine-learning methods from a data quality standpoint (12).

In general, the value of RWE has been well-understood, to the point that many initiatives (and national registries) are amassing insurance claims and EHR curated data.

Real-world evidence derived from observational clinical trials is traditionally collected, objectively as such as data stemming from experimental trials. However, in the observational trials, the setting is entirely uncontrolled and this makes the advancing of "causal learning" to identify direct causes of a certain outcome more difficult. This uncertainty can be overcome by means of modern AI techniques that have proven to be effective, as well, in the setting of synthetic data created by simulation (14). Last but not the least, AI has been proven to mitigate the issue represented by missing data that can impact the process of learning of the causal structure (risk/intervention/outcome). Missing data in real world, moreover, are a significant threat to the understanding of the inference and they are very common; however, AI techniques (in particular Bayesian Networks) do not need complete information on any single record (case/patient) to derive a response variable (15).

To the best of our knowledge, RWD stemming from lifestyle have never been used to train AI-powered simulators alone or along with RWE.

Patient Reported Outomes: Definition and Usefulness in Clinical Research

The FDA deffinition of Patient Reported Outomes (PROs) is "*any report of the status of the patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else*" (16). Indeed, PROs may be referred to symptoms related to a disease, functional statuses or multidimensional constructs such as, for example, the health-related quality of life, as defined in Revicki et al. (17).

PROs are currently used as clinical trial endpoints following a constant increase of their recognition as such over the last two decades. Along with PROs, other measurements related to the patient–reported experience and patient-reported behavior are being used more and more as endpoints in clinical trials.

Structured and validated PROs reduce significantly the heterogeneity of the responses of patients making possible, to a higher extent, the understanding of the real differences in the perception of the outcome, as compared with the information collected *via* open-end questions. In this setting, the value of PROs is not only recognized by the regulator and competent authorities but, as well, by many scientific societies.

PROs can be used as primary, secondary, or even exploratory/tertiary endpoints for the hypothesis generation. Interventional trials experimenting with a new medicinal product do have PROs, normally, as secondary endpoints, whereas palliative care trials or rehabilitation ones can have PROs as primary endpoints.

Following the lines of simplification and according to (18), the benefits of including PROs into clinical trials are:

1. Better understanding of the cost/benefits of a treatment;
2. Better understanding of the patients' experience beyond the biomedical outcomes, especially in the domains of pain, fatigue and inconvenience from any other symptom;
3. Better tools to improve methodologies of trials.

From the regulatory perspectives, the strength of the methodology PROs used in the trial could allow the achievement of the status of "PRO labeling" for approved products, which, in turn, allows PRO-supported claims.

We consider PROs, thus defined and characterized as very reliable endpoints, to measure, in conjunction with the impact on the PROs themselves, the variations in the lifestyle behaviors; these are discussed in the following paragraph under the assumption that they may be considered as biomarkers with clinical value.

LIFESTYLE BEHAVIOR AS A BIOMARKER WITH CLINICAL VALUE AND TYPES OF RWD

Lifestyle behavior includes all features that characterize the daily living of people, without considering possible diseases in a direct manner. The importance of observing lifestyle behavior lies upon the evidence that lifestyle is a health determinant and has a two-way link with the disease. A patient with a chronic disease can see the health deterioration mapped into their lifestyle behavior, whereas changes in daily living can have a significant contribution to health, besides any intervention provided.

Lifestyle and Health

There are several studies that provide solid evidence about the relation of lifestyle with health. A study related to diabetes prevention (18) suggests that lifestyle behaviors are important to the outcomes in youth and adults, with evidence that obesity in adults has risen from <5% to more than 40% in some states, and similar increases in prevalence have been seen in type 2 diabetes, a disease that has increased in prevalence over the last 20–30 years.

Another study (19) shows that good-health-promoting lifestyle behavior, especially health responsibility, physical activity, and stress management behavior are determinants of overweight and obesity, which are major risk factors for the development of cardiovascular diseases, type II diabetes, and some form of cancer.

We quantify lifestyle by obtaining RWD in four fields: physiological, psychological, social, and environmental. Information in all these fields can be objectively measured using devices, or can be subjectively reported by people by answering questionnaires. This information will form the constituents of our proprietary composite biomarker, making up the feature vector to be used as an input to the underlying ML algorithm implementing this biomarker. Not everything we are discussing in the following subsections will be used in the final biomarker. As it is presented in section Composite Lifestyle Biomarker Discovery, domain-knowledge, and feature-importance analysis of the biomarker design process will drive the selection on a per-case basis, but here we give the extensive list.

Types of RWD

We have identified several types of RWD that can be grouped into four categories. In the physiological category of RWD, we mainly encounter RWD that are mostly measured using activity trackers and/or smartphones. Activity-related features are steps walked, distance, elevation, energy dissipation, time spent in different activity intensity zones (e.g., mild, moderate, and high intensity physical activity, as it is formally defined as a function of age) and exercise activities (walking, running, cycling, etc.), and their distribution in the day. Presence indoors or outdoors is also of importance. Specialized physical activity is also measured *via* composite tests like the 6-min walk, the frailty test, or games specifically designed to measure muscular responses (taping on a mobile phone screen for Parkinson's disease or performing other exercises which are monitored and analyzed by depth cameras to measure features important in stroke or accident rehabilitation). All these tests are scripted and hence can be measured using sensors and audiovisual instructions to the people on their smartphones. Heart-related features include the continuous measurements of the heart rate variability, the time spent in different heart rate zones, and the daily resting heart rate measurement. Sleep-related features include continuous measurements on the time spent in the different sleep stages (awake in bed, light, REM, deep sleep). More physiological aspects are reported. We collected reported symptoms (e.g., headache, body temperature, blood pressure, pains, diarrhea, fatigue, nausea), including their intensity. We regularly collected reported weight and height also. Nutrition is paramount, starting at a higher level with the number of meals in the day and the main ingredient of the meal (plant vs. meat-based meal), but a more detailed analysis can also be used when available. Water, coffee, and alcohol intake are reported, and so are toilet visits. Finally, the menstrual cycle is also of importance.

Most RWD types in the psychological category are reported and include a high-level simple emotional state self-assessment or the 11 aspects of the OECD better life index (20), but when deemed necessary the collected information goes deeper using standardized reports from professional therapists who are monitoring the patients. Measurements can also be used to indirectly capture psychological aspects. Emotion can be recognized from the face video, the voice audio, or the social media text posts. Places visited (which, how diverse they are) are also an indication of the psychological state. Aspects like the weather or spending unusual time in commuting can have some importance.

Real-world data in the social category can be measured indirectly from the usage of the phone (diversity, duration, and frequency of calls) and social media (diversity, number, and frequency of interactions). More direct information can be reported using questionnaires on activities with friends, family, or co-workers.

Real-world data in the *environmental category* include environmental indicators for the assessment of the quality of life (*QoL*) that can be reported by the patients using questionnaires. Precise measurements of living or working environment quality can be obtained by integrating relevant commercial devices (e.g., for air quality analysis).

AI TECHNOLOGIES FOR DEFINING, MODELING, AND SIMULATING LIFESTYLE

Risk and outcome predictions in clinical medicine have become more precise due to the remarkable growth in the data collected, and with RWE and the growing interest in AI techniques, the construction of personalized risk and outcome prediction models is now more robust.

Composite Lifestyle Biomarker Discovery

Our digital biomarkers are composite contextual ones, in the sense, that they comprise numerous diverse (usually) indirect measurements, including environmental aspects (5). In Guthrie et al. (21) a methodology for discovering digital biomarkers is introduced that comprises choices on outcomes, features and modeling techniques, and model validation and explanation. Our biomarker discovery is a variation of this methodology. We propose a workflow of specifically designed trials where our biomarker discovery is done in three stages (definition, RWD selection, and iterative design), followed by performance assessment. Our contribution is the introduction of the iterative design where we follow the steps of Guthrie et al. (21) in validating and explaining the models, but we also use the results of this explanation to redefine our RWD selection and retrain the model in iterations.

During the *biomarker definition stage*, the domain experts select the clinically significant outcomes that need to be predicted by the biomarker(s). In most cases, the investigators are interested in whether these outcomes are reached or not by the patients. Then the biomarker is implemented as a binary classifier that predicts success or failure in reaching the outcome (21). There can also be cases where it is interesting to predict the actual values of the different outcome quantities. Then, either a classifier with discrete states predicts outcome value ranges (21), or a regressor predicts an outcome value (22). In essence, this definition stage leads to the number of biomarkers needed, and the underlying ML algorithm family (predictor—binary or multiclass, or regressor) to be employed for implementing each of them.

Other aspects that have to do with training the ML algorithms are also defined at this stage: Primarily, based on the different algorithmic needs and the clinical considerations, the ideal amount of data that needs to be collected is established by deciding on the number of trial people participants and the duration of the trial. If the biomarker is expected to be used during the trial, then the training period of the biomarker needs to be defined. During this training period information is collected to train the ML algorithm but no prediction is attempted. If the purpose of the trial is to discover the biomarker for future use, then the split of the trial population into training, validation, and testing datasets is defined. The design stage is carried out prior to the trials, and the choices made are reflected in their protocols.

In the *biomarker RWD selection stage*, domain knowledge is applied to manually narrow down the list of RWD in all four fields discussed in section Lifestyle Behavior as a Biomarker With Clinical Value and Types of RWD, into those that are relevant to the disease/condition in question. Only established irrelevant RWD are omitted at this stage, since one purpose of the biomarker discovery is to establish if some aspects of RWD discussed in section Lifestyle Behavior as a Biomarker With Clinical Value and Types of RWD do impact the condition in question. Another factor for the RWD selection is the ease of measurement. RWD that are collected unobtrusively are usually in the initial selection, since its collection does not impact the everyday life of the participant. RWD requiring manual input using complicated questionnaires needs to be adequately justified. A user-centric design of the interface of the ePRO greatly helps at this stage, since it can remove the burden of collecting certain RWD. The outcome of this stage is the identification of the initial constituents of the feature vectors to be used as input to the ML algorithms implementing the biomarkers. This stage is also carried out prior to the trials to finalize their protocols.

The core of our biomarker discovery workflow is the biomarker iterative design stage. It involves the iterative retraining of the classifier or regressor implementing the biomarker. In this loop, the ML algorithm is used to train the biomarker using the current version of the feature vector. After training, the results are analyzed to refine the feature vector and repeat the process as long as the validation results are improved. The initial training happens when the first outcomes are collected after the end of the training period. Such outcomes can be intermediate ones, or even the final ones, meaning that the biomarker discovery cannot enter phase three before the end of the trial. At the end of the process, the feature vectors of all people are collected, together with the actual outcomes for the duration of the trial. During the iterations of this design stage, the training set is used to train the ML algorithm of choice. The choice depends on the problem at hand, the most determining factors being the size of the training set and the dimensionality of the feature vector. Classifiers that are able to uncover nonlinear decision surfaces are preferred, namely subclass linear methods (23–25), Kernel methods (26), random forests (27), and (deep) neural networks (28). The validation set is used in iterations to tune the parameters of the underlying ML algorithm of the biomarker and determine which of the feature vector elements strongly contribute to the predictions of the biomarker (either toward positive or negative outcomes). Such an analysis can be done using Pearson correlation coefficients (29, 30) or, more importantly, Shapley additive explanations (SHAP) analysis (31–33). SHAP analysis is applied on every feature vector instance, yielding the effect of each feature vector element toward a positive or negative prediction. *Via* SHAP analysis, we identify those feature vector elements that are consistently not contributing to either positive or negative predictions, and those feature vector elements whose value groups (large, medium, or small) do not consistently push toward a positive or negative prediction.

Cases of SHAP values are shown as rows of point clouds in **Figure 1**. Each point cloud (row) corresponds to a feature vector element, whose importance in the overall decision is being assessed. Each point in the clouds corresponds to the corresponding element of one feature vector on which the decision is based. The color of the point indicates the value of the element (from small values in blue to large values in red).

The placement of the point on the horizontal axis corresponds to the SHAP value. Values close to zero correspond to feature vector elements with negligible effect on the decision, whereas large positive or negative values correspond to feature vector elements with large effect. The vertical displacement of the point within its row indicates how many feature vectors fall into the particular range of SHAP values. Thus, thick point cloud areas correspond to many feature vectors. The point clouds marked as (a) correspond to feature vector elements that have a large impact on decisions (either positive or negative). The point cloud marked as (b) corresponds to a feature vector element whose large, medium, or small value seems to push the decision to random directions. The point cloud marked as (c) corresponds to a feature vector element that has a small impact on decisions. Feature vector elements falling in any of the categories marked as (b) or (c) are candidates to be dropped in the next iteration of biomarker retraining.

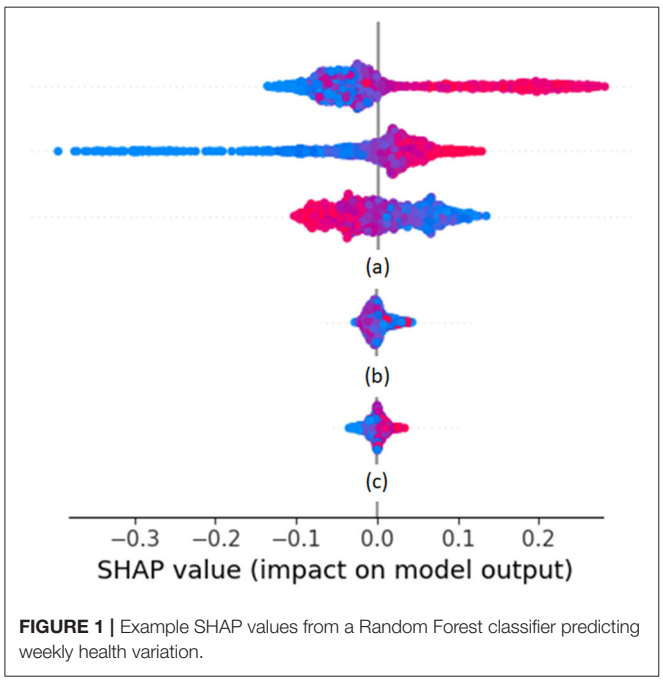


FIGURE 1 | Example SHAP values from a Random Forest classifier predicting weekly health variation.

The performance of the biomarker is assessed at the biomarker performance assessment stage that follows after the iterative process. The RWD test set is used at this stage. It needs to be noted that for biomarkers predicting final trial outcomes, there are only as many feature vectors as there are patients in the trial. When the number of participants is low, then the validation stage is done using the test set itself, or in more tight cases the process is run repeatedly employing the "leave one out for testing" method, where all features are individually used to test classifiers or estimators trained and tuned using all the rest. The three biomarker-discovery phases and the resulting biomarker assessment are summarized in our proposed workflow for biomarker discovery trials, shown in **Figure 2**.

Using the Biomarkers for Digital Therapeutics

By employing AI discovered biomarkers that successfully predict clinically significant outcomes, it is possible to drive decisions in digital therapeutics (DTx) (21, 34, 35). When a disease is considered, then the aim is to drive the intervention. The biomarker predictions indicate intervention strength (drug dosage), which usually is not one of the feature vector elements.

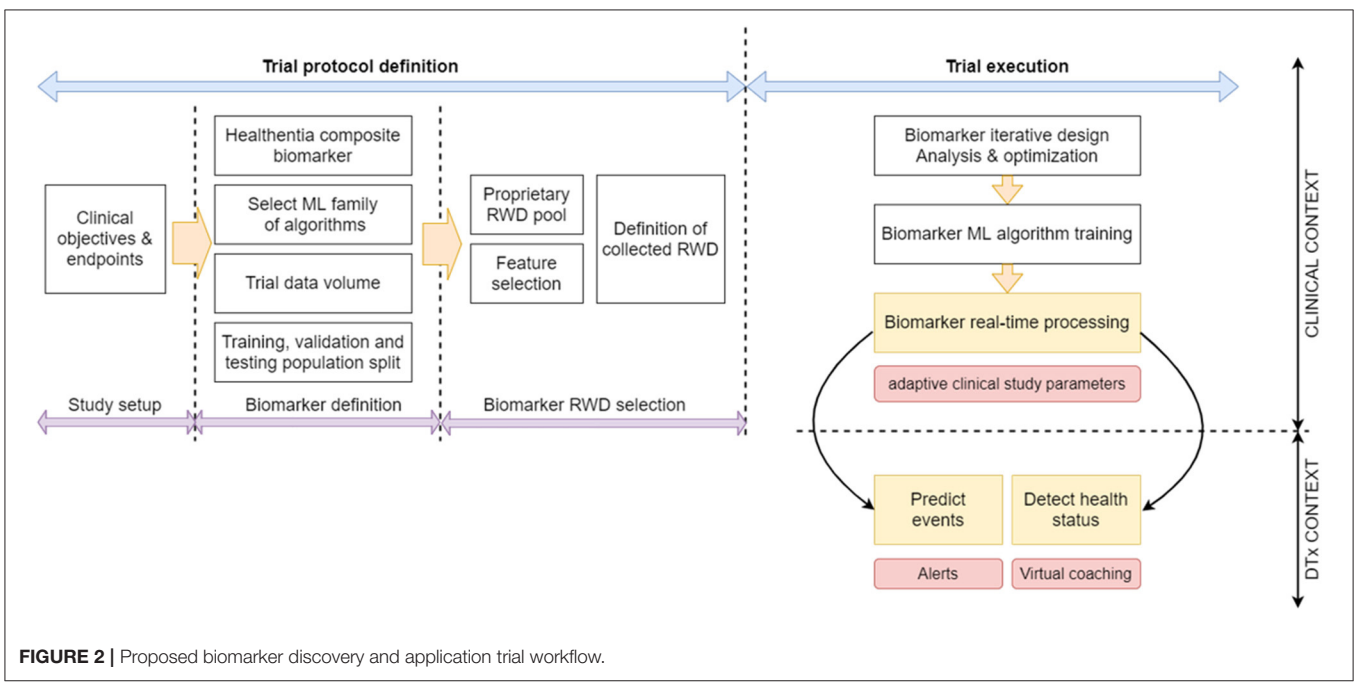


FIGURE 2 | Proposed biomarker discovery and application trial workflow.

The usage of a biomarker in DTx involves balancing the trade-off between its specificity (its ability to correctly identify those patients who do not achieve the desired clinical outcome) and its sensitivity (the ability to correctly identify those patients that achieve the desired clinical outcome). Usually, high specificity is required not to reduce the intervention to patients who actually need it. If at that high specificity the biomarker also yields high sensitivity, that is, identifies the patients who should receive less strong intervention, then the biomarker is a successful one. This is usually quantified by the area under the receiver operating characteristic (ROC) curve.

Digital therapeutics is also applied in the more general, disease agnostic, and well-being areas. In this area, a biomarker is used to drive behavioral change in a virtual coaching setup. The explainable AI elements already discussed in the previous section determine the elements of the feature vector of the particular patient who had the most positive and negative influence on the probability of a successful outcome. Then the person is coached in these elements. The virtual coach selects the feature vector elements of the strong influence that are related to the behavior (physiological, psychological, and social aspects) and to the environment. If they have a strong negative influence, it coaches the person to change behavior. If they have a strong positive influence, it encourages the person to keep up the good lifestyle in those aspects.

While recently the use of ePRO and digital monitoring devices in clinical trials is ever-increasing, and many of those aim at deriving digital biomarkers (6), to the 'knowledge of the author there are no trials that already have evaluated DTx applications. It is our hypothesis that our digital biomarkers and the SHAP analysis of their individual decisions can be applied in a DTx context by driving coaching of the patients.

PILOT STUDY TO EVALUATE THE HYPOTHESIS

Study Description

The study APACHE, which is an advanced patient monitoring and AI-supported outcomes assessment in cervical cancer using Internet of things technologies, is a cofinanced monocentric observational study using a remote monitoring device for patients affected by locally advanced cervical cancer. Patients are considered as such if staged larger than or equal to IB2 according to the FIGO staging system (an international staging system for locally advanced cervical cancer), with primary lesions larger than 4 cm. The patients undergo chemoradiotherapy (CRT) followed by either radical surgery or brachytherapy boost and are treated in Fondazione Policlinico Universitario "A. Gemelli" IRCCS of Rome, Italy. The foreseen study duration is 24 months. The study protocol foresees inclusion and exclusion criteria. The inclusion criteria require the patients to be younger than 70 years, be clinically able to use portable technologies, and be able to understand and sign informed consent. The exclusion criteria involve a major psychiatric disorder, inadequate performance status (larger than 3 according to the Eastern Cooperative Oncology Group score, that is, capable of only limited selfcare; confined to bed or chair for more than 50% of waking hours), and ongoing pregnancy or breastfeeding. Patient enrolment began in October 2020 and continues to date. A total of 50 patients are foreseen for this exploratory study. The selection procedure of patients adhering to the study protocol criteria foresees only a brief interview during the first visit to the advanced radiation therapy center of Gemelli for the initial radiotherapy treatment. During the interview, the informed consent is acquired by the attending physician, and papers describing the trial and expected role of the patient are provided. If the patient is motivated and computer literate, the whole procedure does not take more than 15–20 min. Please note that the initial inclusion criteria on cervical cancer has been widened to include other pelvic cancers, as discussed in section Expected Benefits, Early Findings, and Next Steps.

The primary objective of the study is to assess the experience of patients using Healthentia (see section Healthentia Platform) and a wearable tracker during the multimodal oncological therapies and follow-up period. The study has three secondary objectives. Firstly, to compare PROs with corresponding clinical records about toxicity, instrumental activities of daily living (IADLs), and stress/coping levels. Secondly, to profile patients based on their scores and activity, and lastly, to train models using ML on the patient-reported and monitored data.

Patients have received a state-of -the-art wearable device (Fitbit INSPIRE) during their first visit prior to CRT start, that collects at a daily basis RWD like activity (i.e., steps per day), sleep, and vital signs. During the whole observation period, patients are asked to report their weekly well-being by completing dedicated questionnaires sent to the ePRO App on their phone. A dedicated research nurse will flank the patients enrolled in the study in filling the e-questionnaires in case of need and follow up the correct flow of the questionnaires.

The following scoring systems are selected to assess specific aspects of the experience of the patient during the multimodal treatment period:

1. Early and late toxicity will be assessed using the NCI-PRO-CTCAE™ ITEMS-ITALIAN Version 1.0 for the cutaneous, gastro-intestinal, and genito-urinary sections
2. Therapy impact on instrumental daily activities will be assessed using the Lawton Brody IADL
3. QoL will be assessed using the EORTC QLQ C30
4. Nutritional status will be assessed using the malnutrition screening tool
5. Psychological status will be assessed through self-administered tests, namely the distress thermometer, DT6 for distress evaluation and the Mental Adjustment to Cancer Scale, MINI-MAC 7, Italian version for coping evaluation
6. User experience and technology acceptance will be assessed using a customized questionnaire on the Healthentia App on enjoyment, aesthetics, control, trust in technology, perceived usefulness, and intention to use.

The collected data are transferred through the Healthentia app on the smartphone of the patient to Healthentia platform. Clinicians can monitor lifestyle behavioral patterns, detect changes in the health status, and be informed on clinical

endpoints *via* the Healthentia portal web application. More information on the Healthentia eClinical solution is given in section Healthentia Platform.

The collected RWD, i.e., objective data from wearable devices and subjective data collected from questionnaires (e.g., IADLs, toxicity, QLQ, etc.) are fused together and define a multidimensional vector for each patient that consists of steps, resting heart-rate, sleep, etc., which characterizes their behavior throughout the day. After the models are trained from the RWD collected over an initial period, it is possible to predict outcomes and system scoring of the above scoring systems i.e., IADLs or QLQ, by feeding the system with the automatically collected vectors.

The RWD that is being collected is currently grouped into five lifestyle aspects described in **Table 1**. Most of the aspects can be considered generic (applicable to people in general, not just cervical cancer patients). Only the aspect on QoL, since it focusses on low toxicity adverse effects, is dedicated to the particular condition under study in APACHE. Each aspect is measured *via* a set of parameters (different measurements of questions). The parameters are concatenated into a set of scores, as described in the section on the manual RWD selection stage. It is through these scores that the five aspects are quantified.

Manual RWD Selection Stage

Each of the five lifestyle aspects (see **Table 1**) combines multiple parameters (measurements or questionnaire answers). In most cases, these parameters are aggregated into one or more scores quantifying the performance in the respective lifestyle aspect. The physiological lifestyle aspect parameters are measured on a daily basis using Fitbit activity trackers. Four scores are derived from these measurements. A sleep score is derived from the total sleep duration, and the REM and deep sleep durations, the sleep efficiency (ratio of time being asleep, over the total time in bed), sleep disturbances (count of wake-up times), and the bedtime alignment to the habits of the patient. An activity score is determined by the active vs. inactive time (excluding sleep), positive or negative deviation from habits, and auto-detected training count. Steps are used as a standalone score since they are the most usual activity tracking metric and are easy to compare against different activity trackers. Finally, the resting heart rate is another standalone score due to its clinical importance as a biomarker on body condition.

The independent lifestyle aspect is reported on a weekly basis using the IADL questionnaire. There are eight questions covering telephone use, shopping, food preparation, housekeeping, laundry, transportation, medication adherence, and finances management. Each of these parameters contributes equally to the overall independence score.

The QoL lifestyle aspect is assessed using two questionnaires. The EORTC QLQ assesses parameters on symptom experience (15 parameters, assessed weekly), body image (three parameters, assessed weekly), sexual/vaginal functioning (four parameters, assessed weekly), and sexual worry/activity/enjoyment (three parameters, assessed monthly). These four groups are the four scores from EORTC QLQ, three of them obtained weekly and one monthly.

The NCI/DRO/CTCAE questionnaire collects 45 parameters on a weekly basis, all having to do with different symptoms (their occurrence, frequency, and/or distress level). They are grouped into five categories, the gastrointestinal (16 parameters), the skin (13 parameters), the neuro (2 parameters), the sexual (2 parameters), and the urinary (8 parameters). There are also two parameters that cannot be classified in the above groups of interest and are ignored by our scoring of the APACHE outcomes. They have not been removed from the data collected to maintain the integrity of the questionnaire used. These five categories are the five scores from NCI/DRO/CTCAE, obtained weekly.

The psychological lifestyle aspect is also assessed using two questionnaires. The Mini Mac Scale assesses (every 3 months) parameters that are aggregated into scores (all parameters contributing equally to their respective scores) on the fighting spirit (16 parameters), helplessness/hopelessness (six parameters), anxious preoccupation (nine parameters), fatalism (eight parameters), and denial/avoidance (one parameter).

The distress thermometer comprises a single parameter forming a score on a weekly basis. Finally, the nutrition lifestyle aspect comprises a single parameter on malnutrition score collected on a weekly basis.

Summarizing, the biomarkers discovered in APACHE utilize 66 parameters and/or their grouping into 12 scores as a feature vector. They are being trained to predict significant variations in the three scores quantifying QoL from the EORTC QLQ questionnaire. Training is done on anonymized APACHE data using proprietary scripts built on top of well-established implementations of ML algorithms found in the Scikit Learn and Tensorflow libraries.

TABLE 1 | Lifestyle aspects grouping of the Real-World Data collected in APACHE, their parameters, and their scores.

Lifestyle aspect	Questionnaire/ measurement	Parameters/score
Physiological	Fitbit activity measurements	16 parameters, aggregated into 4 scores: sleep quality, steps, activity score, resting heart rate
Independence	Instrumental Activities of Daily Life (IADL)	8 parameters, aggregated into 1 score
QoL (Low toxicity adverse effects)	EORTC QLQ (CX24, CX30)	25 parameters aggregated into 4 scores
	NCI/DRO/CTCAE	45 parameters, aggregated into 5 scores
Psychological	Mini Mac Scale	40 parameters, aggregated into 5 scores
	Distress thermometer	1 parameter/score
Nutrition	Nutrition score	1 parameter/score

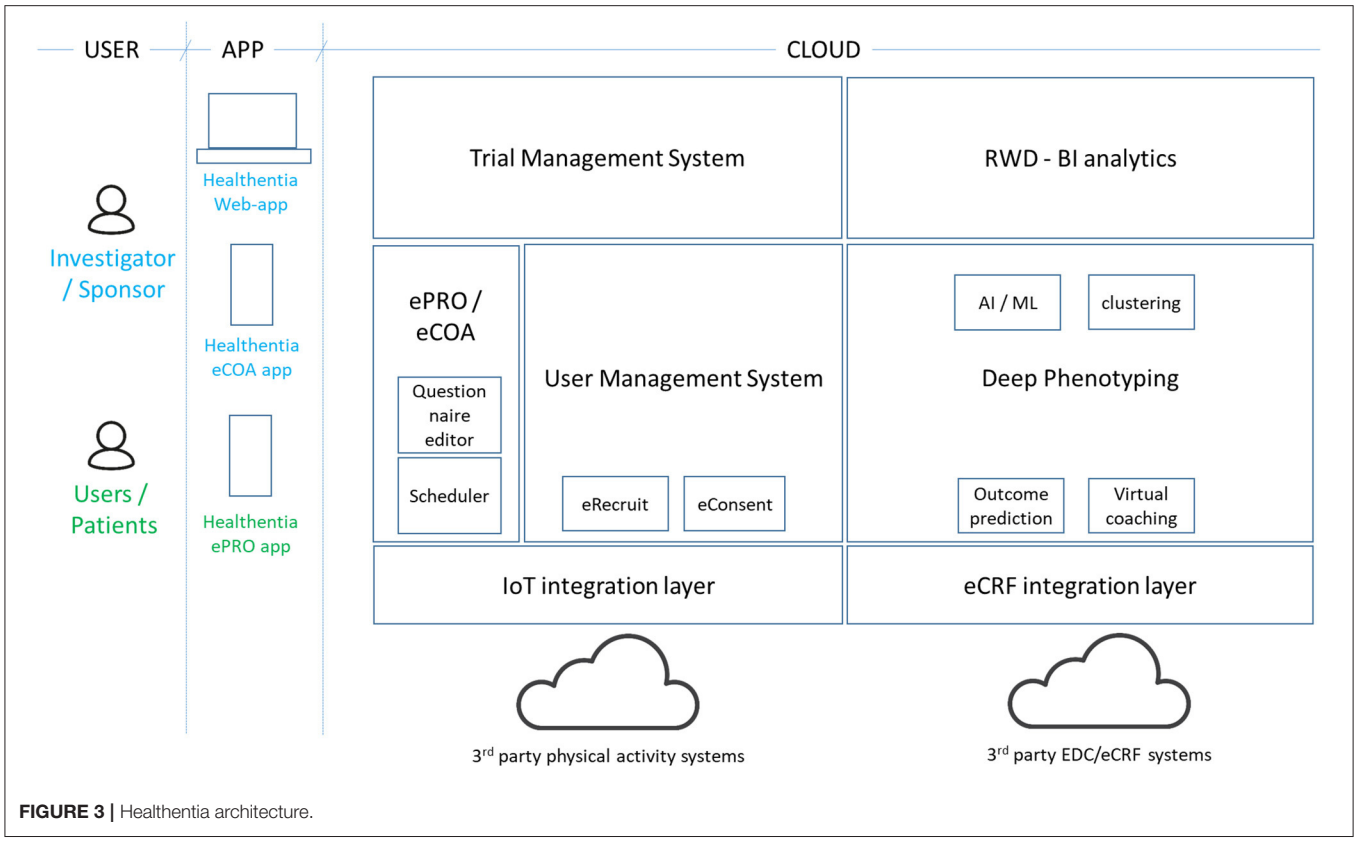


FIGURE 3 | Healthentia architecture.

Healthentia Platform

Healthentia (4) is an eClinical solution that facilitates clinical trial optimization by accelerating the trial processes, reducing the failure rate, and validating drug/intervention efficacy and effectiveness with RWD insights. In this way, pharmaceutical companies can achieve cost savings, accelerate the drug approval process, and obtain useful insights to develop drugs and interventions of higher efficacy. Its architecture is shown in **Figure 3**.

The Healthentia solution extends the use of a traditional ePRO/eCOA application by adding behavioral and health-related data collected from Internet of Things (IoT) devices. Utilizing ML on this data, it is possible to discover behavioral biomarkers and cluster patients into behavioral phenotypes, which allows the activation of smart services to predict clinical outcomes, generate prevention alarms, and link phenotypes with drug/intervention efficacy. In addition, based on reported outcomes, the AI module generates automatic alerts in case of adverse events. These automatic and prevention alarms support decision making by the investigator during the clinical trial, for the benefit of the health of the individual patient. For the AI module to do so online, the biomarker needs to be trained first as discussed in section Manual RWD Selection Stage.

On top of *in vivo* clinical studies, Healthentia allows the running of *in silico* trials that use the deep phenotyping outcomes together with legacy data, to create synthetic control arms and support pharmaceutical companies to design optimized studies.

Healthentia is available for clinical studies, under a strict regulatory framework, and a SaaS environment, which is open to the wider community. The SaaS version includes further features, such as eRecruitment, eConsent, and Virtual Coaching. Healthentia is already in use in APACHE, and we have received ethical clearance for its use in more studies with a top pharmaceutical company and a hospital. Results from its AI module (for biomarker training) have been published (36), albeit in a completely different domain.

EXPECTED BENEFITS, EARLY FINDINGS, AND NEXT STEPS

The APACHE study has been running for a few weeks now, albeit with lower enrolment rates than expected. The first RWD are being collected, as shown in **Figure 4**, where the real-time monitoring functionalities of Healthentia offer investigators views of the RWD, like the depicted physical activity and MINI-MAC cancer scale.

In the APACHE study, we have introduced new patient-centered variables for risk stratification (i.e., toxicity onset, malnutrition, or mental coping issues), allowing the prospective setup of rapid and fully personalized therapeutic approaches. The integration of such variables into clinical nomograms and multidimensional predictive models is contributing to realizing decisional support systems. The study contributes to the active monitoring of toxicity and therapy-related side effects, aiming

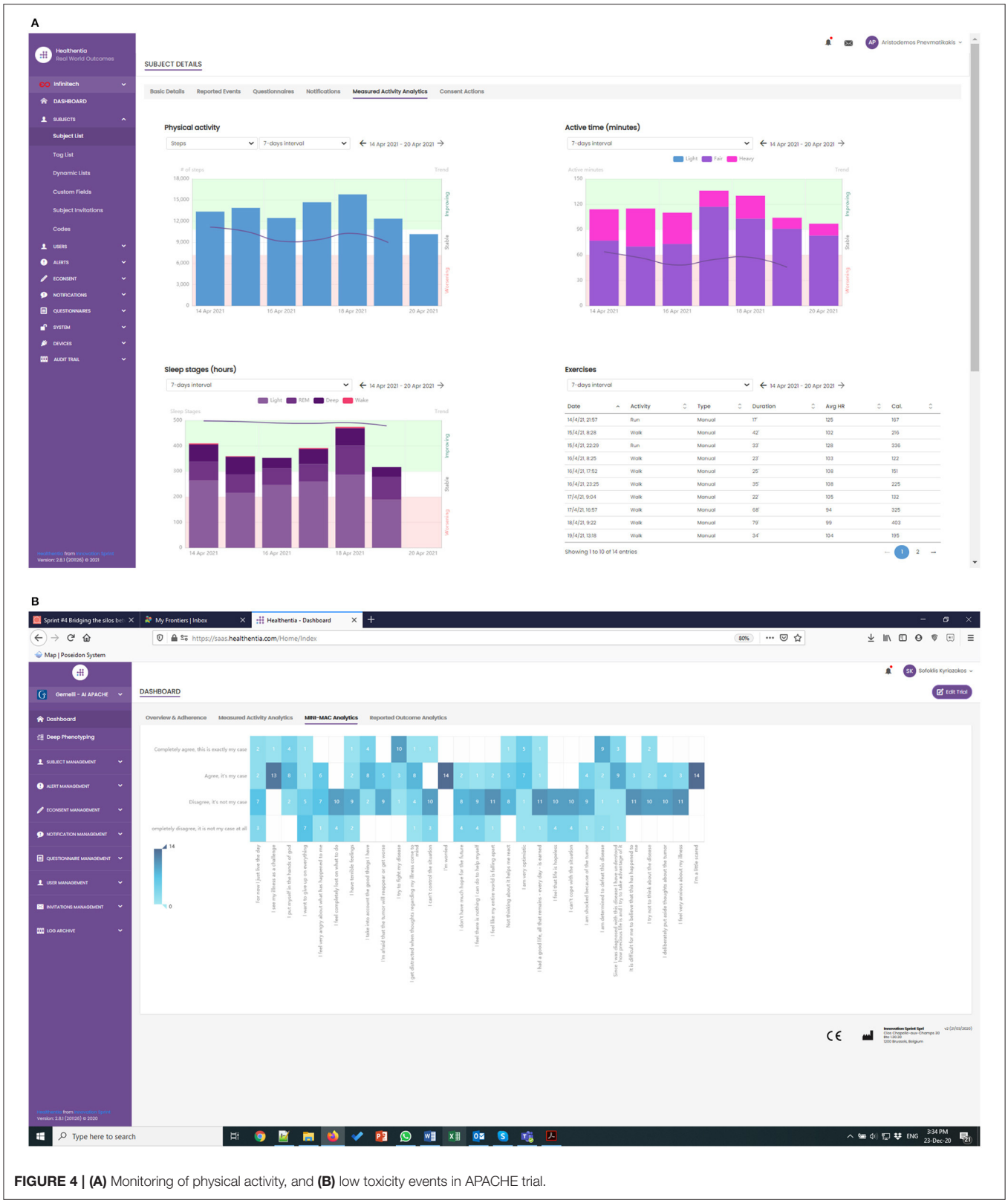


FIGURE 4 | (A) Monitoring of physical activity, and **(B)** low toxicity events in APACHE trial.

at their reduction, and in the optimization of monitoring and follow-up strategies of the patients. Finally, the use of Healthentia enhances self-awareness of the patients about global clinical status and participated clinical decision making.

Clinicians expect different benefits from the use of such advanced monitoring techniques. First of all, the reduction of toxicities may hamper QoL of the patient and their compliance to the oncological multimodal treatments, thanks to the identification of alert signs and early symptoms that may be overlooked during the visits or considered negligible by the patients themselves and not properly reported to the attending physician. This may be translated into an active personalization of the ongoing treatments (i.e., radiotherapy replanning secondary to bowel toxicity) with significant expected advantages in terms of treatment quality and overall clinical outcomes. Furthermore, the use of this monitoring approach may represent a key resource for coping strategies enhancement, especially in hospitals where a psychoncology service is not available.

As the volume of collected RWD increases, our next steps in the analysis side have to do with modeling the different scores and discovering a biomarker for the predictions of the QoL ones. Due to conflicting studies that do not allow to run multiple trials on the same cohort of patients, the enrollment has been slow, and therefore an amendment has been proposed to enlarge the cohort and grant the success of the trial. At the time being, patients affected by other pelvic cancer undergoing CRT are allowed in the study (e.g., endometrial, vaginal, vulvar, and rectal cancer). This expansion of the inclusion criteria is justified by the similarities of cervical to the other pelvic cancers in terms of treatment (there is a radiotherapy does overlap) and most importantly the common types of possible toxicity (linked to the irradiation of the same pelvic organs, i.e., bowel, bladder, rectum) that allow the reuse of the same questionnaires and eliminate differences in data analysis.

The limitation of the APACHE study lies with the automatic measurements. These need the compliance of the study participant, whose familiarity with activity trackers significantly declines with their age. The participant needs to understand the importance of the physiological data being collected and make sure their activity tracker is worn and is charged. Working with the unavoidable gaps in the collected physiological data is something we are investigating.

Scoring Considerations

In section Study Description we presented the study design, which is complemented by the scoring mechanisms that we have been prepared prior to the launch of the study. The physiological scores combine diverse parameters, and their combination weights will be investigated to obtain meaningful results. All the other scores combine parameters that carry similar weights and hence there are three combination options:

A linear combination does not discriminate between number of events or their severity. As an example, consider four parameters with values 0–4, 4 indicating maximum severity. An average score of 1 is obtained with four 1's or three 0's and one 4. There are cases where the latter is more alarming than the former. Some of the questionnaires used in APACHE do have formal scoring suggestions (37–41), following the linear case.

Non-linear combination, on the other hand, allows the investigator to put more weight on either the number of events or their severity. Consider a set of possible events x_n, each with a value of zero if there is no symptom and integer values larger than zero indicating increased severity of the symptom. The events are combined into a single score s using:

$$s = \frac{1}{N}\left(\sum_{n=1}^{N} x_n^a\right)^{1/a}$$

Selecting $a > 1$ the investigator has a score that puts emphasis on event severity vs. event count. On the other hand, selecting $a < 1$ puts the emphasis on the event count. Selecting $a = 1$ leads to the linear case.

As already mentioned, the formal scoring of most of the questionnaires is linear. This is followed to facilitate clinical research. But to facilitate ML, we also experiment with non-linear scoring in APACHE. We are selecting the value of the exponent for each combination leading to the different scores, based on what the investigators need to emphasize with each one of them.

Iterative Design Stage

The iterative design phase has recently started with an initial algorithm selection. Since there are 50 patients in the trial, there are 50 feature vector instances to train and evaluate the biomarkers. For this reason, algorithms like neural networks are not expected to be used. The biomarkers will most probably be based on a decision-tree classifier or random forests. Linear methods (with careful feature engineering) will be used as a baseline, together with their multiclass variants, since the decision boundaries are not expected to be linear.

The biomarkers will be trained using the leave-one-out method, each time keeping one patient for testing, training with 45 and using the remaining four for hyper-parameter tuning during validation. The performance will be reported as the number of correctly identified patients out of the 50 leave-one-out experiments.

Early Findings and Next Steps

At the definition stage of the APACHE biomarkers discovery, the clinically significant outcomes that need to be predicted by the biomarker are selected. These are dictated by the goal of APACHE, that is, the QoL in terms of low toxicity adverse effects, as is quantified by the three scores of the lifestyle aspect. As APACHE trial starts producing RWD, we will be training our biomarker to predict significant variations of these scores. At some milestone of the trial (currently planned for its end on the 52nd week) the predictors of the biomarker will be able to determine if a significant improvement of the 3 scores associated with the QoL of the patients is to be expected.

Having trained the biomarker, we will be exploiting the explainable AI techniques described in section Composite Lifestyle Biomarker Discovery to determine the aspects in the life of the patient to coach in favor or against. This way, as discussed in section Using the Biomarkers for Digital Therapeutics, we will be driving DTx in this therapeutic area.

Although the RWD collection and scoring presented in this study is customized to the needs of the APACHE trial for cervical cancer, the methodology for capturing and combining data, and

most importantly, that for discovering biomarkers, is applicable to other conditions as well. Preliminary results are available for the application of this methodology on RWD for obesity, where the discovered biomarker predicts significant short-term weight variations and general well-being, where the biomarker classifies the health outlook of general population, aiming at using it for risk assessment and the analysis of its decisions in virtual coaching. We have published these early results in Pnevmatikakis et al. (36), and our next steps in biomarker discovery research involve applying the discovery methodology in the APACHE data to predict the low-toxicity events.

CONCLUSIONS

The APACHE study addresses a very important milestone, that is represented by the clinical validation of AI technologies when creating models based on PROMs capable of predicting any outcome of clinical value. In this modeling, endeavor the clinical validation still represents a bottleneck. In particular, the complexity of promoting RWE from basic clinical decision support (needs validation from an accountable "real doctor") to a fully validated (rather "qualified") digital biomarker RWD-from-lifestyle-raw-material based is quite significant. The challenge is represented by the robust regulatory framework set to qualify a classical biomarker, herein adopted to evaluate a digital one vs. strictly technological standards, requirements, and credentials. If we consider the privacy endeavor (related to innovative data capture and handling solutions) and the cyber-sec one, these alone represent entirely new dimensions entering the ethical/regulatory dialog.

In our perspective, extracting evidence with predictive values from lifestyle in a very homogeneous cohort (and a technological endeavor ethically and regulatory robust) of subjects undergoing state-of-the-art treatment magnifies its value by offsetting this RWE toward a very stable, and to a certain extent expected, clinical outcome progression (observational nature of the approach). In other words, a study like this creates the idea sandbox to evaluate the training of an ML algorithm in a low noise setting.

Clinically, extracting such RWE has significant implications. Lifestyle-driven and outcome-connected digital biomarkers with the predictive value could enrich the diagnostic tools with agile (and relatively inexpensive) indicators (easy to collect in a continuous fashion), for example, of the onset of significant toxicity from an oncological treatment.

Training cycle by training cycle, moreover, these digital biomarkers could pave the way to smart coaching that, in turn, could be promoted toward validated digital content as an active ingredient in a DTx perspective.

DATA AVAILABILITY STATEMENT

The clinical study APACHE and its raw data are property of Gemelli and cannot be shared.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Gemelli. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SK, AP, and AC were the main editors. KK has contributed to the customization of Healthentia to support the APACHE trial. SK, AP, and AC have contributed to the ML of Healthentia. LB, VV, and GS have contributed to the study design and faciliating the APACHE trial. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The process of digital composite biomarker discovery is based on a methodology developed by Innovation Sprint Sprl, as part of the Digital Biotech activity. In this activity, Innoviris.brussels has contributed through the project Healthentia: Deep Learning of Patients (2020-RDIDS-20) by cofunding with Innovation Sprint the customization and validation of Healthentia for the APACHE study, and the purchasing of wearable devices and cloud hosting. Gemelli hospital is running the APACHE study using Healthentia and is financing all elements of the local running (e.g., investigator, insurance). The authors wish to acknowledge the valuable contribution of the reviewers in improving the original manuscript.

REFERENCES

1. Nivel DL, Jared McCormick T, Straus SE, Hemmelgarn BR, Jeffs L, Barnes TRM, et al. Reproducibility of clinical research in critical care: a scoping review. *BMC Med.* (2018) 16:26. doi: 10.1186/s12916-018-1018-6
2. Henegan C, Goldacre B, Mahtani KR. Why clinical trials outcomes fail to translate into benefits for patients. *Trials.* (2017) 18:22 doi: 10.1186/s13063-017-1870-2
3. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun.* (2018) 11:156–64. doi: 10.1016/j.conctc.2018.08.001
4. Innovation Sprint. *Healthentia: Driving Real World Evidence in Research & Patient Care.* (2021). Available online at: https://innovationsprint.eu/healthentia (accessed May 1, 2021).
5. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit Med.* (2019) 2:14. doi: 10.1038/s41746-019-0119-8
6. Kovalchick C, Sirkar R, Regele OB, Kourtis LC, Schiller M, Wolpert H, et al. Can composite digital monitoring biomarkers come of age? A framework for utilization. *J Clin Transl Sci.* (2017) 1:373–80. doi: 10.1017/cts.2018.4
7. Garrow JS, Webster J. Quetelet's index (W/H2) as a measure of fatness. *Int J Obes.* (1985) 9:147–53.

8. Theodoridis S, Koutroumbas K. *Pattern Recognition, Fourth Edition (4th. ed.)*. Orlando, FL: Academic Press, Inc. (2008).
9. US Food and Drug Administration. *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff*. (2017). Available online at: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf (accessed August 2021).
10. Organization for Economic Co-Operation and development (OECD)/World health Organization (WHO). *Access to New Medicines in Europe: Technical Review of Policy Initiatives and Opportunities for Collaboration and Research*. (2015). Organization for Economic Co-Operation and development (OECD)/World health Organization (WHO).
11. Crown WH. Real world evidence, casual inference and machine learning. *Value Health.* (2019) 22:587–92. doi: 10.1016/j.jval.2019.03.001
12. Rajkamar A, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* (2018) 1:18. doi: 10.1038/s41746-018-0029-1
13. Perry W, Hossain R, Taylor RA. Assessment of the feasibility of automated, real time clinical decision support in the emergency department using HER data. *BMC Emerg Med.* (2018) 18:19. doi: 10.1186/s12873-018-0170-9
14. Rathnam C, Lee S, Jiang X. An algorithm for direct causal learning of influences on patient outcome. *Artif Intell Med.* (2017) 75:1–15. doi: 10.1016/j.artmed.2016.10.003
15. Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, Druzdzel MJ, et al. Bayesian networks for risk prediction using real world data: a tool for precision medicine. *Value Health*. (2019) 22:439–45. doi: 10.1016/j.jval.2019.01.006
16. FDA. *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Products Development to Support Labelling Claims*. Silver Spring, MD (2009).
17. Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, et al. Recommendations on health-realted quality of life research to support labeling and promotional claims in the United States. *Qual Life Res.* (2000) 9:887–900. doi: 10.1023/a:1008996223999
18. Margaret G. Lifestyle determinants of health: Isn't it all about genes and environment? *Nurs Outlook*. (2017) 65:505–5. doi: 10.1016/j.outlook.2017.04.011
19. Joseph-Shehu EM, Ncama BP, Irinoye OO. Health-promoting lifestyle behaviour: a determinant for noncommunicable diseases risk factors among employees in a Nigerian University. *Glob J Health Sci.* (2019) 11, 15–26. doi: 10.5539/gjhs.v11n12p15
20. OECD Better Life Index. Available online at: http://www.oecdbetterlifeindex.org (accessed August 2021).
21. Guthrie NL, Carpenter J, Edwards KL, Appelbaum KJ, Dey S, Eisenberg DM, et al. Emergence of digital biomarkers to predict and modify treatment efficacy: machine learning study. *BMJ Open*. (2019) 9:e030710. doi: 10.1136/bmjopen-2019-030710
22. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY:Springer (2006).
23. Pnevmatikakis L. Polymenakos, 'Subclass Linear Discriminant Analysis for Video-Based Face Recognition', *J Visual Commun Image Represent.* (2009) 20:543–51. doi: 10.1016/j.jvcir.2009.08.001
24. Moghaddam B. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. (2002) 24:780–8. doi: 10.1109/TPAMI.2002.1008384
25. Zhu M, Martínez AM. Subclass discriminant analysis. *IEEE Trans Pattern Anal Mach Intell*. (2006) 28:1274–86. doi: 10.1109/TPAMI.2006.172
26. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. *Neural Comput*. (2000) 12:2385–404. doi: 10.1162/089976600300014980
27. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
28. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Networks*. (2015) 61:85–117. doi: 10.1016/j.neunet.2014.09.003
29. Pearson K. Notes on regression and inheritance in the case of two parents. *Proc R Soc London*. (1895) 58:240–2. doi: 10.1098/rspl.1895.0041
30. Galton F. Typical laws of heredity. *Nature*. (1877) 15:532–3. doi: 10.1038/015532a0
31. Lundberg SM, Lee SI. Consistent feature attribution for tree ensembles. In: *Proceedings of the 34th International Conference on Machine Learning, JMLR: W&CP*. (2017). pp. 15–21.
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions.In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. editors. *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc (2017). pp. 4766–75
33. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9
34. Palanica A, Docktor MJ, Lieberman M, Fossat Y. The need for artificial intelligence in digital therapeutics. *Digit Biomark*. (2020) 4:21–5. doi: 10.1159/000506861
35. Wang T, Azad T, Rajan R. *The Emerging Influence of Digital Biomarkers on Healthcare, RockHealth report*. Available online at: https://rockhealth.com/reports/the-emerging-influence-of-digital-biomarkers-on-healthcare/ (accessed August 2021).
36. Pnevmatikakis A, Kanavos S, Matikas G, Kostopoulou K, Cesario A,Kyriazakos S. Risk assessment for personalized health insurance based on real world data. *Risks*. (2021) 9. doi: 10.3390/risks9030046. Available online at: https://www.mdpi.com/2227-9091/9/3
37. Watson M, Greer S, Young J, Inayat Q, Burgess C, Robertson B. Development of a questionnaire measure of adjustment to cancer: the MAC scale. *Psychol Med.* (1988) 18:203–9. doi: 10.1017/S0033291700002026
38. Ferguson M, Capra S, Bauer J, Banks M. Development of a valid and reliable malnutrition screening tool for adult acute hospital patients. *Nutrition.* (1999) 15:458–64. doi: 10.1016/S0899-9007(99)00084-2
39. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. *The EORTC QLQ-C30 Scoring Manual (3rd Edition). Published by: European Organisation for Research and Treatment of Cancer*. (2001). Brussels. Available online at: https://qol.eortc.org/manuals/ (accessed August 2021).
40. Graf C. The Lawton instrumental activities of daily living scale. *Am J Nurs.* (2008) 108:52–62; quiz 62-3. doi: 10.1097/01.NAJ.0000314810.46029.74
41. Zhou X, Garbinsky D, Gnanasakthy A. *Methods for Reporting the Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) Data in Cancer Clinical Trials*. Durham, NC: Duke Industry Statistics Symposium (DISS) (2019). doi: 10.1016/j.jval.2018.04.1528

Conflict of Interest: SK, AP, AC, and KK were employed by company Innovation Sprint Sprl.

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 Digital Health

TYPE Original Research
PUBLISHED 09 November 2022
DOI 10.3389/fdgth.2022.1029191

OPEN ACCESS

EDITED BY
Daniel Donoho,
Children's National Hospital, United States

REVIEWED BY
Hani J. Marcus,
University College London, United Kingdom
Guillaume Kugener,
University of Southern California, United States
Timing Liu,
University of Cambridge, United Kingdom

*CORRESPONDENCE
Parisa Rashidi
parisa.rashidi@bme.ufl.edu

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 26 August 2022
ACCEPTED 14 October 2022
PUBLISHED 09 November 2022

CITATION
Shickel B, Silva B, Ozrazgat-Baslanti T, Ren Y, Khezeli K, Guan Z, Tighe PJ, Bihorac A and Rashidi P (2022) Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks. Front. Digit. Health 4:1029191. doi: 10.3389/fdgth.2022.1029191

Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks

Benjamin Shickel[1,4], Brandon Silva[2,4], Tezcan Ozrazgat-Baslanti[1,4], Yuanfang Ren[1,4], Kia Khezeli[2,4], Ziyuan Guan[1,4], Patrick J. Tighe[3,4], Azra Bihorac[1,4] and Parisa Rashidi[2,4]*

[1]Department of Medicine, University of Florida, Gainesville, FL, United States, [2]Department of Biomedical Engineering, University of Florida, Gainesville, FL, United States, [3]Department of Anesthesiology, University of Florida, Gainesville, FL, United States, [4]Intelligent Critical Care Center (IC3), University of Florida, Gainesville, FL, United States

Transformer model architectures have revolutionized the natural language processing (NLP) domain and continue to produce state-of-the-art results in text-based applications. Prior to the emergence of transformers, traditional NLP models such as recurrent and convolutional neural networks demonstrated promising utility for patient-level predictions and health forecasting from longitudinal datasets. However, to our knowledge only few studies have explored transformers for predicting clinical outcomes from electronic health record (EHR) data, and in our estimation, none have adequately derived a health-specific tokenization scheme to fully capture the heterogeneity of EHR systems. In this study, we propose a dynamic method for tokenizing both discrete and continuous patient data, and present a transformer-based classifier utilizing a joint embedding space for integrating disparate temporal patient measurements. We demonstrate the feasibility of our clinical AI framework through multi-task ICU patient acuity estimation, where we simultaneously predict six mortality and readmission outcomes. Our longitudinal EHR tokenization and transformer modeling approaches resulted in more accurate predictions compared with baseline machine learning models, which suggest opportunities for future multimodal data integrations and algorithmic support tools using clinical transformer networks.

KEYWORDS

transformer, deep learning, electronic health records, critical care, patient acuity, clinical decision support

1. Introduction

Through the course of a typical intensive care unit (ICU) admission, a variety of patient-level data is collected and recorded into electronic health records (EHR) systems. Patient data is diverse, including measurements such as vital signs, laboratory tests, medications, and clinician-judged assessment scores. While primarily used for ad-hoc clinical decision-making and administrative tasks such as billing, patient-centric data can also be used to build automated machine learning systems for assessing overall patient health and predicting recovering or worsening patient trajectories.

Patient mortality risk is often used as a proxy for overall ICU patient acuity, both in traditional illness severity scores like SOFA (1, 2) and more recent machine learning approaches such as DeepSOFA (3). Whether manually calculated or algorithmically computed, nearly all of these systems rely on measurements from a set of handpicked clinical descriptors thought to be most indicative of overall patient health. Given the breadth of data available in modern EHR systems, there is untapped potential for enhanced patient modeling contained in the large amount of unused patient data.

Several recent studies have demonstrated the predictive accuracy and patient modeling capacity of deep learning implementations in healthcare, using models such as recurrent neural networks (RNN) (3–8) and convolutional neural networks (CNN) (9, 10).

Recently, Transformer models (11) have garnered increased attention in the deep learning community due to their state-of-the-art results on a variety of natural language processing (NLP) tasks, particularly when using schemes such as Bidirectional Encoder Representations from Transformers (BERT) (12). There are also more recent advances in analyzing frequency of data in Frequency Enhanced Decomposed Transformer Zhou et al. (13) that exploits the sparseness of time series data.

From a temporal perspective, one advantage the Transformer offers is its parallel processing characteristics. Rather than processing data points sequentially, the Transformer views all available data at once, modeling attention-based relationships between all input time steps. In contrast, models such as RNNs require distinct temporal separation within input sequences, and usually demand a regular sample interval between adjacent time steps. As clinical EHR data is recorded at highly irregular frequency and is often missing measurements, a large amount of data preprocessing is typically required in the form of temporal resampling to a fixed frequency, and an imputation scheme to replace missing values. Furthermore, given that several EHR measurements are often recorded at the same timestamp, typical machine learning workflows aggregate temporally adjacent measurements into mean values contained in resampled time step windows, or perform random shuffling procedures before training models. Given its parallel and fundamentally temporally agnostic attributes, the Transformer is capable of distinctly processing all available measurements, even those occurring at the same timestamp. Additionally, the Transformer is able to process whichever data happens to be available, reducing the need for potentially bias-prone techniques to account for data missingness.

In this study, we showcase the feasibility of a highly flexible Transformer-based patient acuity prediction framework in the critical care setting. Our contributions can be summarized by the following:

- Our flexible system design incorporates a diverse set of EHR input data that does not require *a priori* identification of clinically relevant input variables, and can work with any data contained in EHR platforms.
- In contrast to recent Transformer approaches that either use discrete medical concepts (14–16) or continuous measurements from a handpicked set of features (17), we introduce a data embedding scheme that jointly captures both concept and corresponding measurement values of a wide variety of disjoint clinical descriptors.
- In our novel embedding module, we introduce a mechanism for combining both absolute and relative temporality as an improvement over traditional positional encoding.
- We present an input data scheme with minimal preprocessing, obfuscating the need for potentially biased temporal resampling or missing value imputation common in many other sequential machine learning approaches.
- We expand BERT's [CLS] token for classification into several distinct tokens for predicting multiple-horizon patient mortality and ICU readmission in a novel multi-task learning environment.
- Rather than typical concatenation with sequential representation, we incorporate static patient information in a novel way using a global self-attention token so that every sequential time step is compared with the static pre-ICU representation.
- We show that the Longformer (18) can be applied to long EHR patient data sequences to minimize required computation while retaining superior performance.

2. Methods

2.1. Cohort

The University of Florida Integrated Data Repository was used as an honest broker to build a single-center longitudinal dataset from a cohort of adult patients admitted to intensive care units at University of Florida Health between January 1st, 2012 and September 22nd, 2019. Our project was approved by the Institutional Review Board of the University of Florida and the University of Florida Privacy Office (IRB201901123). Full cohort statistics is described in **Table 1**.

We excluded ICU stays lasting less than 1 h (to reduce EHR data artifacts and provide predictive models with adequate patient data) or more than 10 days, to limit outliers based on tokenized sequence length and following several existing studies using ICU encounters for predictive modeling (19). Excluding patients based on length of stay resulted in roughly 95% of the original ICU cohort. Our final cohort consisted of 73,190 distinct ICU stays from 69,295 hospital admissions and 52,196 unique patients. The median length of stay in the ICU was 2.7 days.

We divided our total cohort of ICU stays into a development cohort of 60,516 ICU stays (80%) for training our models, and a validation cohort of 12,674 ICU stays

TABLE 1 Summary statistics for experimental ICU cohorts.

	Development cohort (n = 60, 516)	Validation cohort (n = 12, 674)
Patients, *n*	41,881	10,315
Hospital encounters, n	57,168	12,127
Age, years, median (25th, 75th)	61.0 (49.0, 71.0)	62.0 (49.0, 73.0)
Female, *n* (%)	27,380 (45.2)	5,616 (44.3)
Body mass index, median (25th, 75th)	26.9 (23.0, 32.0)	27.3 (23.3, 32.2)
Hospital length of stay, days, median (25th, 75th)	6.7 (3.6, 12.1)	6.4 (3.3, 11.5)
ICU length of stay, days, median (25th, 75th)	2.8 (1.5, 5.1)	2.9 (1.6, 5.5)
Time to hospital discharge, days, median (25th, 75th)	1.9 (0.0, 4.8)	1.1 (0.0, 4.1)
Hispanic, *n* (%)	2,130 (3.5)	539 (4.3)
Non-English speaking, *n* (%)	1,092 (1.8)	233 (1.8)
Marital status, *n* (%)		
Married	26,084 (43.1)	5,457 (43.1)
Single	21,844 (36.1)	4,931 (38.9)
Divorced	11,905 (19.7)	2,142 (16.9)
Smoking status, *n* (%)		
Never	20,180 (33.3)	4,653 (36.7)
Former	19,378 (32.0)	4,167 (32.9)
Current	12,094 (20.0)	2,326 (18.4)
Insurance status, *n* (%)		
Medicare	31,447 (52.0)	6,543 (51.6)
Private	13,115 (21.7)	2,912 (23.0)
Medicaid	10,208 (16.9)	1,999 (15.8)
Uninsured	5,746 (9.5)	1,220 (9.6)
Comorbidities, *n* (%)		
Charlson comorbidity index, median (25th, 75th)	2.0 (0.0, 4.0)	2.0 (0.0, 4.0)
Myocardial infarction	7,537 (12.5)	1,985 (15.7)
Congestive heart failure	14,897 (24.6)	3,380 (26.7)
Peripheral vascular disease	10,005 (16.5)	2,185 (17.2)
Cerebrovascular disease	8,981 (14.8)	1,720 (13.6)
Chronic pulmonary disease	17,938 (29.6)	3,473 (27.4)
Metastatic carcinoma	3,377 (5.6)	812 (6.4)
Cancer	8202 (13.6)	1,808 (14.3)
Mild liver disease	4,745 (7.8)	960 (7.6)
Moderate/severe liver disease	1,856 (3.1)	374 (3.0)
Diabetes without complications	14,137 (23.4)	2,395 (18.9)
Diabetes with complications	5,052 (8.3)	1,736 (13.7)
AIDS	442 (0.7)	53 (0.4)
Dementia	1,692 (2.8)	559 (4.4)
Paraplegia/hemiplegia	3,465 (5.7)	769 (6.1)
Peptic ulcer disease	1,110 (1.8)	187 (1.5)
Renal disease	11,878 (19.6)	2,493 (19.7)
Rheumatologic disease	1,794 (3.0)	342 (2.7)
Neighborhood characteristics, median (25th, 75th)		
Total population, $n \times 10^3$	17.0 (10.6, 26.4)	17.6 (10.6, 26.7)
Distance to hospital, km	39.3 (17.9, 69.1)	42.4 (20.2, 76.5)
Median income, dollars $\times 10^3$	40.1 (33.8, 46.7)	40.1 (35.1, 47.4)
Poverty rate, %	19.6 (14.0, 27.7)	19.3 (13.7, 26.7)
Rural area, n	22543 (37.3)	4691 (37.0)
Clinical outcomes, *n* (%)		
ICU readmission before hospital discharge	3,583 (5.9)	613 (4.8)
Inpatient mortality	5,813 (9.6)	1,131 (8.9)
7-day mortality	5,237 (8.7)	1,022 (8.1)
30-day mortality	7,056 (11.7)	1,380 (10.9)
90-day mortality	9,197 (15.2)	1,785 (14.1)
1-year mortality	12,991 (21.5)	2,288 (18.1)

(20%) for evaluating their predictive performance. 10% of the development set was used for within-training validation and early stopping. The cohort was split chronologically, where the earliest 80% of ICU stays was used for training, and the most recent 20% used for evaluation. To ensure the same patient did not appear in both development and validation sets, all ICU stays of patients with multiple admissions spanning the cohort threshold were grouped into the development cohort.

2.2. Data

We extracted patient data from several EHR data sources: sociodemographics and information available upon hospital admission, summarized patient history, vital signs, laboratory tests, medication administrations, and numerical assessments from a variety of bedside scoring systems. We did not target or manually select any specific ICU variables, instead using all such data contained in our EHR system. A full list of variables used in our experiments is shown in Table 2.

Static data: For each ICU stay, we extracted a set of non-sequential clinical descriptors pertaining to patient characteristics, admission information, and a summarized patient history from the previous year. Patient-level features included several demographic indicators, comorbidities, admission type, and neighborhood characteristics derived from the patient's zip code. Patient history consisted of a variety of medications and laboratory test results up to one year prior to hospital admission (Table 2). Historical patient

TABLE 2 Summary of variables used in Transformer experiments.

Variable	Type
Patient demographics	
Age	Static
Sex	Static
Ethnicity	Static
Race	Static
Language	Static
Marital status	Static
Smoking status	Static
Insurance provider	Static
Patient residential information	
Total population	Static
Distance from hospital	Static
Rural/Urban	Static
Median income	Static
Proportion black	Static
Proportion hispanic	Static
Percent below poverty line	Static
Patient admission information	
Height	Static
Weight	Static
Body mass index	Static
17 comorbidities present at Admission	Static
Charlson comorbidity index	Static
Presence of chronic kidney disease	Static
Admission type	Static
Patient history: medications[a]	
ACE inhibitors	Static
Aminoglycosides	Static
Antiemetics	Static
Aspirin	Static
Beta blockers	Static
Bicarbonates	Static
Corticosteroids	Static
Diuretics	Static
NSAIDS	Static
Vasopressors/Inotropes	Static
Statins	Static
Vancomycin	Static
Nephrotoxic drugs	Static
Patient history: laboratory test results[b]	
Serum hemoglobin	Static
Urine hemoglobin	Static
Serum glucose	Static
Urine glucose	Static
Urine red blood cells	Static
Urine protein	Static
Serum urea nitrogen	Static
Serum creatinine	Static
Serum calcium	Static
Serum sodium	Static
Serum potassium	Static
Serum chloride	Static
Serum carbon dioxide	Static
White blood cells	Static
Mean corpuscular volume	Static
Mean corpuscular hemoglobin	Static
Hemoglobin concentration	Static
Red blood cell distribution	Static
Platelets	Static
Mean platelet volume	Static
Serum anion gap	Static
Blood pH	Static
Serum oxygen	Static
Bicarbonate	Static
Base deficit	Static
Oxygen saturation	Static
Band count	Static
Bilirubin	Static
C-reactive protein	Static
Erythrocyte sedimentation rate	Static
Lactate	Static
Troponin T/I	Static
Albumin	Static
Alaninen	Static
Asparaten	Static
ICU vital signs	
Systolic blood pressure[c]	Temporal
Diastolic blood pressure[c]	Temporal
Mean arterial pressure[c]	Temporal
Heart rate	Temporal
Respiratory rate	Temporal
Oxygen flow rate	Temporal
Fraction of inspired oxygen (FIO2)	Temporal
Oxygen saturation (SPO2)	Temporal
End-tidal carbon dioxide (ETCO2)	Temporal
Minimum alveolar concentration (MAC)	Temporal
Positive end-expiratory pressure (PEEP)	Temporal
Peak inspiratory pressure (PIP)	Temporal
Tidal volume	Temporal
Temperature	Temporal
ICU Assessment Scores[d]	
ASA physical status classification	Temporal
Braden scale	Temporal

TABLE 2 Continued

Variable	Type
Confusion assessment method (CAM)	Temporal
Modified early warning score (MEWS)	Temporal
Morse fall scale (MFS)	Temporal
Pain score	Temporal
Richmond agitation-sedation scale (RASS)	Temporal
Sequential organ failure assessment (SOFA)	Temporal
ICU laboratory tests[e]	
106 distinct lab tests present in EHR system	Temporal
ICU medications[e]	
345 distinct medications present in EHR system	Temporal

[a]Extracted features included total counts of administered medications up to one year prior to hospital admission.
[b]Extracted features included total counts of recorded laboratory test results and minimum, maximum, mean, and standard deviation of measurement values up to one year prior to hospital admission. Both serum and urine-based tests extracted separately when available.
[c]Invasive and non-invasive readings for systolic blood pressure, diastolic blood pressure, and mean arterial pressure were treated as distinct event tokens.
[d]For assessment scores with multiple sub-components, each component was treated as a distinct timestamped measurement, resulting in 30 such assessment measurements.
[e]We retained distinct laboratory tests and medications that were administered in at least 1% of the training cohort of ICU stays.

measurement features were derived from a set of statistical summaries for each descriptor (minimum, maximum, mean, standard deviation).

Temporal data: For each ICU stay, we extracted all available vital signs, laboratory tests, medication administrations, and bedside assessment scores recorded in our EHR system while the patient was in the ICU (Table 2). We refer to each extracted measurement as a clinical event. Each event was represented as a vector containing the name of the measurement (e.g. "noninvasive systolic blood pressure"), the elapsed time from ICU admission, the current measured value, and eight cumulative value-derived features corresponding to prior measurements of the same variable earlier in the ICU stay (mean, median, count, minimum, maximum, standard deviation, first value, elapsed time since most recent measurement). For bedside assessment scores with multiple sub-components, we treated each sub-component as a distinct measurement. Invasive and noninvasive measurements were treated as distinct tokens. We excluded ICU stays with sequence lengths longer than 12,000 tokens, and the resulting mean sequence length in our cohorts was 1,996.

Data processing: Categorical features present in the pre-ICU static data were converted to one-hot vectors and concatenated with the remaining numerical features. Missing static features were imputed with training cohort medians, but no such imputation was required for the tokenized temporal ICU data. Binary indicator masks were computed and concatenated with static features to capture patterns of missingness.

Static features were standardized to zero mean and unit variance based on values from the training set. For each variable name in the temporal ICU data, corresponding continuous measurement value features were individually standardized in the same manner. ICU measurement timestamps were converted to number of elapsed hours from ICU admission, and were similarly standardized based on training cohort values.

ICU measurement names were converted to unique integer identifiers in a similar manner to standard tokenization mapping procedures in NLP applications. Each temporal clinical event was also associated with an integer position index. While similar to the positional formulations in NLP applications, we introduce one key distinction that is more suitable for Transformers based on EHR data: we do not enforce the restriction that positional indices are unique, and if two clinical events occurred at the same EHR timestamp, they are associated with the same sequential position index.

Each temporal measurement token consisted of integer positional identifier, integer variable identifier, continuous elapsed time from ICU admission, and eight continuous features extracted from current and prior measurement values.

Following data extraction and processing, each ICU stay was associated with two sets of data: (1) a single vector $x_s \in \mathbb{R}^{718\times 1}$ of 718 static pre-ICU features, and (2) a matrix of T temporal ICU measurements $x_t \in \mathbb{R}^{T\times 12}$ including token position and identifier. Across our entire population, the temporal ICU measurements included 19 unique vital signs, 106 unique laboratory tests, 345 unique medication administrations, and 29 bedside assessment score components; however, each ICU stay only included a subset of such total variables, and its corresponding temporal sequence only included what was measured during the corresponding ICU stay. One of the benefits of our proposed EHR embedding framework is the lack of resampling, propagation, imputation, or other such temporal preprocessing typically performed in related sequential modeling tasks.

2.3. Clinical outcomes

For each ICU stay, we sought to predict six clinical outcomes related to patient illness severity: ICU readmission within the same hospital encounter, inpatient mortality, 7-day mortality, 30-day mortality, 90-day mortality, and 1-year mortality. Our model is formulated as a multi-task design, and simultaneously estimates risk for all six clinical prediction targets.

2.4. Model architecture

The primary driver behind our ICU patient acuity estimation model is the transformer encoder (11). Our modified model utilizes the global and sliding window mechanism introduced by the Longformer (18) along with special classification tokens from BERT (12). **Figure 1** shows a high-level overview of our Transformer architecture. Our longitudinal tokenization pipeline and Transformer modeling architecture code will be available upon request for interested researchers.

Novel embedding: In typical Transformer implementations, one-dimensional input sequences consist of integer-identified tokens (such as textual tokens or discrete clinical concepts) that are embedded using a lookup table, after which a positional encoding vector is added to inject local temporality. For existing applications of Transformers with EHR data, the values of a given measurement are not factored into its representation.

Our embedding scheme introduces three novelties that offer improvements for clinical prediction tasks. First, positional indices are derived from EHR record times and are not unique (see Section 2.2), allowing for multiple tokens to share the same positional index and resulting positional encoding. Rather than enforce an arbitrary sequence order or implement a random shuffling procedure for simultaneous tokenized events, this modification is more flexible with respect to clinical workflows.

Second, in addition to novel framing of relative and local temporal relationships through positional encoding modifications, each clinical event token also explicitly includes absolute temporality in the form of a feature indicating the elapsed hours from ICU admission. We hypothesized that the injection of both relative and absolute temporality would allow the Transformer to better model patient trajectories.

Finally, each clinical event in our tokenized input sequences consists of several continuous measurement values in addition to the discrete token identifiers (see Section 2.2). To our

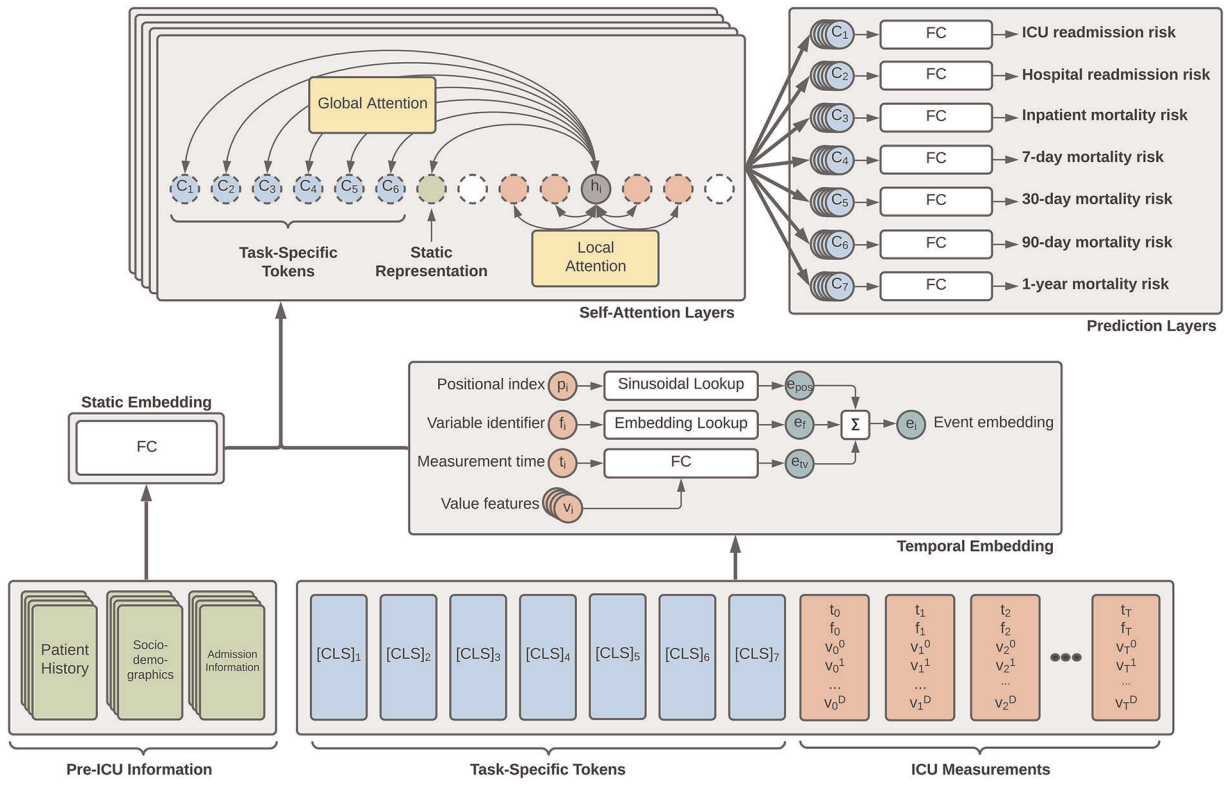


FIGURE 1
Overview of our proposed generalized EHR Longformer network for simultaneously predicting multiple patient outcomes in the ICU. Pre-ICU information includes summarized history of patient medications and laboratory tests, sociodemographic indicators, and features relating to hospital admission. Temporal ICU measurements take the flexible form of tuples: (p, non-unique positional index of clinical event based on timestamp; t, elapsed time from ICU admission, f, unique measurement identifier integer; $\vec{v}$, set of continuous features derived from measured values). Task-specific [CLS] tokens are assigned t = time of prediction and $\vec{v} = 0$. Tokens are individually embedded and passed through a stack of Longformer layers with sliding self-attention windows. Global attention is applied to static feature representation and prediction tokens. The concatenation of each layer's [CLS] representations are used for a given task to predict the desired mortality risk. Not shown: Transformer feedforward network and nonlinear activations. FC: fully-connected layers.

knowledge, no other work integrates both discrete and continuous data in this manner, with the majority of recent research opting for discrete medical codes only (Section 4.2). We augment discrete variable tokens with continuous measurement values into our embedding to better capture recovery or worsening trends as a patient progresses through an ICU stay.

Our embedding module consists of (1) a traditional lookup table used for measurement name identifier, (2) a sinusoidal positional embedding table, and (3) a single fully-connected layer for embedding absolute time and value-derived features. The final sequence embedding is the summation of three embedded vectors: (1) the embedding of absolute time with corresponding cumulative values, (2) the measurement token identifier embedding, and (3) a traditional sinusoidal positional encoding. In our implementation, the sinusoidal positional encoding is based on the position of unique measurement times in the input sequence: for an example sequence of measurement hours [0.1, 0.2, 0.2, 0.3, 0.3], the positional indices are computed as [0, 1, 1, 2, 2].

Novel multi-task global tokens: In the original BERT implementation, a single special [CLS] token is prepended to input sequences that is meant to capture a global representation of the entire sequence. We extend this notion by prepending each sequence with 6 such special tokens: one for each of our clinical outcomes. As each token in our data scheme consists of a (time, name, values) 12-tuple, we set time of each [CLS] token equal to the total ICU length of stay and all values equal to zero. The special token identifiers are embedded in a similar fashion to other ICU measurement tokens. In our experiments, we include an additional prediction target for long-term hospital readmission that is used for regularization, but not included in our patient acuity estimation. In the Longformer implementation in our encoder, we set each of the multi-task tokens to compute global attention, so that self-attentions are computed among all sequence elements for each clinical outcome token.

Novel inclusion of static patient data: In many sequential models for clinical prediction, a final encounter representation is obtained by concatenating the pre-sequence static patient representation with the sequential representation. In our work, we prepend each ICU sequence with the representation obtained from passing the static patient information vector through a fully-connected network. We assign this static token as global, so that every time step computes attention with the static data. We hypothesized that this more fine-grained injection of patient information at every time step would improve the capacity of our model to learn important and more personalized patient trajectory patterns.

Model details: Our final model consisted of an embedding layer, followed by 8 Longformer layers, and a separate linear prediction layer for each of our 6 clinical outcomes. For making a task-specific prediction, the task-specific linear layer uses the concatenation of representations corresponding to its special [CLS] token at each of the 8 layers. In our initial Longformer implementation, we used a hidden size of 128, a feedforward size of 512, 8 attention heads, a sliding window of size 128, dropout of 0.1, and a batch size of 21. Hyperparameters were chosen with respect to hardware constraints; hyperparameter optimization will be a focus of future work.

Experiment details: Models were trained using a development set of 60,516 ICU stays corresponding to 80% of our total ICU cohort. 10% of this development set was used for early stopping based on the mean AUROC among all six clinical outcomes and a patience of four epochs. All experiments were conducted on a local Linux server equipped with two i7-7820X 3.6 GHz CPUs, 3 NVIDIA GeForce RTX 2080Ti GPUs, 512GB SSD storage, and 128GB RAM. Models were designed and run using the PyTorch and Hugging Face Python libraries.

In this feasibility study, we compared performance against six other ICU prediction models:

- Longformer using tokenized data sequences with only discrete code identifiers. In this variant of our proposed framework, we do not include the continuous measurement values in the representation of each event token.
- Recurrent neural network (RNN) with gated recurrent units (GRU) using continuous multivariate time series inputs. In this experiment, the flexibility of our tokenization scheme is removed, and more traditional "tabularized" input data sequences were constructed where each variable is assigned a distinct column. Sequences were constructed with continuous current values and resampled to 1-hour frequency to align with common practice found in literature. Multi-task predictions were drawn from the final hidden state of the GRU encoder. Static patient information was concatenated with the sequence representation and fed through fully-connected layers before classification.
- GRU with attention mechanism. This variant is identical to the above, but with the addition of a simple attention mechanism over the hidden states of the GRU. States are weighted by alignment scores and summed to yield a final attention-based sequential representation.
- Tokenized GRU with attention. In this final experimental setting, we used the same novel EHR embedding and tokenization approach as with our Transformer model architecture (see Section 2.2), but instead use a GRU with attention mechanism in place of the Transformer model.
- CatBoost (20) gradient boosting algorithm. The algorithm employs gradient boosting on decision trees for both regression and classification tasks. Gradient boosting algorithms have shown benefits over random forests and

require comparatively less hyperparameter tuning for optimal performance. For this experiment, the embedding layers are removed and the CatBoost model is trained on samples containing both the pre-ICU information and concatenated ICU measurements.

- XGBoost (21) gradient boosting algorithm. This experiment and associated data processing is identical to CatBoost, except an XGBoost model is used for prediction.

3. Results

At present time, the primary aim of our novel mortality prediction model is not to show state-of-the-art improvements in model accuracy; rather, we present this work as a feasibility study for future research. We believe our novel modifications of existing Transformer architectures for use in clinical EHR applications will result in highly flexible and more personalized patient representations and predictions across a variety of clinical tasks.

In this first iteration of our experiments, we did not perform any hyperparameter optimization, instead choosing sensible settings that both highlight the novel aspects of the architecture and work with our hardware constraints. In passing, we note that often parameter tuning is an essential component of enhancing performance, and future iterations of this work will focus on optimizing crucial parameters such as learning rate, dropout, number of self-attention heads, number of self-attention layers, hidden dimension, and size of the sliding self-attention window.

Our results are shown in **Table 3**. Our Transformer architecture with novel EHR embedding and tokenization scheme yielded slightly superior mean AUROC (0.929) across all six clinical prediction tasks, with individual task AUROC ranging from 0.843 (ICU readmission) to 0.983 (7-day mortality). The Transformer using tokenized embeddings that omit continuous measurement values resulted in the lowest mean AUROC (0.773) and worst performance across most of the clinical outcomes, ranging from 0.512 (ICU readmission) to 0.900 (7-day mortality). It outperformed the XGBoost model for inpatient and 7-day mortality.

In terms of GRU baseline models, the traditional model and data processing scheme resulted in the lowest baseline accuracy, with mean AUROC of 0.900 and task AUROC ranging from 0.750 (ICU readmission) to 0.972 (7-day mortality). The augmentation of this model and data scheme with traditional attention mechanism improved the performance to a mean AUROC of 0.909.

The best GRU baseline model used our novel EHR embedding, tokenization, and representation pipeline. This model yielded a mean AUROC of 0.927 with individual task AUROC ranging from 0.831 to 0.982. It performed best for predicting 30-day mortality and 90-day mortality, although the relative difference compared with the transformer is minimal. For the gradient boosting algorithms, CatBoost outperformed XGBoost across all outcomes (mean AUROC: 0.863 vs. 0.836) except for predicting ICU readmission (AUROC: 0.759 vs. 0.762). The CatBoost model performed similarly to the baseline GRU model for all other outcomes. The tree-based models were predominantly outperformed by GRU models with attention.

Across all models and data representation schema, ICU readmission proved the most difficult task. Among the multiple prediction horizons for patient mortality, models were best able to predict 7-day mortality, followed by inpatient mortality, 30-day mortality, 90-day mortality, and 1-year mortality.

4. Discussion

4.1. Principal findings

This work presents a novel ICU acuity estimation model inspired by recent breakthroughs in Transformer architectures. Our proposed model framework incorporates several novel modifications to the existing Transformer architecture that make it more suitable for processing EHR

TABLE 3 Multi-task prediction results expressed as area under the receiver operating characteristic curve (AUROC).

Model	Data	Mean	Readmission	Mortality				
			ICU	Inpatient	7-Day	30-Day	90-Day	1-Year
Transformer	Tokenized events (discrete only)	0.773	0.512	0.889	0.900	0.831	0.777	0.727
Transformer	Tokenized events + continuous measurement values	**0.929**	**0.843**	**0.978**	**0.983**	0.953	0.923	**0.892**
GRU	Resampled multivariate time series	0.900	0.750	0.960	0.972	0.938	0.907	0.872
GRU with attention	Resampled multivariate time series	0.909	0.770	0.965	0.975	0.946	0.914	0.882
GRU with attention	Tokenized events + continuous measurement values	0.927	0.831	0.977	0.982	**0.954**	**0.925**	0.891
CatBoost	Tokenized events + continuous measurement values	0.863	0.759	0.901	0.915	0.890	0.868	0.847
XGBoost	Tokenized events + continuous measurement values	0.836	0.762	0.867	0.878	0.859	0.833	0.817

data of varying modalities. Through initial feasibility experiments, our model was on par with, or outperformed, common variants of RNN baselines, and we feel our approach holds promise for incorporating additional EHR-related outcome prediction tasks and additional sources of EHR input data.

One of the advantages of our work is that input elements are treated as distinct. For example, if heart rate, respiratory rate, and SPO2 were recorded at the same timestamp in an EHR system, our framework operates on these individual elements, rather than combining them into a single aggregated time step as in similar RNN or CNN-based work. From an interpretability standpoint, combined with the inherent self-attention mechanisms of the Transformer, isolation of inputs allows for improved clarity with respect to important or contributing clinical factors. While one area of recent sequential interpretability research involves multivariate attribution for aggregated time steps (5, 22), Transformer-based approaches such as ours obfuscate the need for multivariate attribution, as attentional alignment scores are assigned to individual measurements. This property highlights the potential for EHR Transformers to shed increased transparency and understanding for clinical prediction tasks built upon complex human physiology.

Furthermore, while many sequential applications of deep learning to EHR (including recent implementations of Transformer techniques) make use only of discrete clinical concepts, our proposed framework extends the representational capacity by integrating continuous measurement values alongside these discrete codes and events. The inclusion of continuous measurement values represents an important step forward, as the measured result of a clinical test or assessment can provide crucial information alongside a simple presence indicator that can help complex models develop a better understanding of patient state and overall health trajectory.

Given the flexible nature of our Transformer framework, each patient input sequence only contains the measurements that were made during the ICU encounter. The advantages for EHR applications are twofold. First, in traditional RNN or CNN-based work, the distance between time steps is assumed to be fixed, and this is typically achieved by resampling input sequences to a fixed frequency by aggregating measurements within resampled windows, and propagating or imputing values into windows without present values. Such a scheme has the potential for introducing bias, and when using our novel EHR embedding paradigm and Transformer-based modeling approach, the problem of missing values is made redundant given the explicit integration of both absolute and relative temporality for each irregularly measured clinical event. Additionally, in typical deep sequential applications using EHR data, the number of input features at each time step must be constant. This is achieved by an *a priori* identification and extraction of a subset of clinical descriptors thought to be relevant indicators for a given prediction task. As we have shown, when using a Transformer-based approach with our flexible tokenization scheme, any and all EHR measurements can be easily incorporated into the prediction framework, even when some types do not exist for a given patient or ICU encounter, and do not necessitate bias-prone imputation techniques.

While the Transformer offers several benefits over existing sequential deep learning models such as the RNN, it is not without drawbacks. Because the self-attention mechanism is highly parallelizable and does not require step-wise iterative processing of a sequence (unlike the RNN), there is a tradeoff between faster computation and a much larger memory footprint (complexity $\mathcal{O}(n^2)$ without scope modifications). As such, Transformers may be infeasible to implement in training environments with limited computational resources.

In our approach, we introduced a novel method for incorporating static, pre-sequential patient information and patient history into the overall prediction model. Typically, such static information is concatenated with a final sequential representation before making a prediction. We instead include static information as a distinct token in the input sequence, and assign global attention using the Longformer self-attention patterns. In effect, static patient-level information is injected into the self-attention representation of every ICU measurement, allowing more fine-grained and personalized incorporation of changes in overall patient health trajectories.

Another novel contribution we feel can be applied to even non-EHR tasks is the expansion of the special BERT classification token into a separate token per classification target in a multi-task prediction setting. Given the global self-attention patterns between all task tokens and every sequential input element, such a scheme allows the model to develop task-specific data representations that can additionally learn from each other.

As with other retrospective machine learning models for predicting patient outcomes from longitudinal data, our transformer framework offers the potential for augmenting clinical decision-making with dynamic data-driven risk estimations that can be used to help forecast patient trajectory and guide treatment and care strategies. Intended not to mandate particular course of action, tools such as ours can complement existing standards of care and provide clinicians with additional support.

4.2. Related work

4.2.1. Transformer models

First introduced by Vaswani et al. (11) for machine translation tasks, the Transformer is a deep learning architecture built upon layers of self-attention mechanisms.

The Transformer views attention as a function of keys K, queries Q, and values V. In the work of Vaswani et al. (11), all three elements came from the same input sequence, and is why their style of attention is referred to as self-attention. In a similar manner to previously described works, compatibility between a key and query is used to weight the value, and in the case of self-attention, each element of an input sequence is represented as a contextual sum of the alignment between itself and every other element. Similar to the memory networks of Sukhbaatar and Szlam (23), the Transformer also involves the addition of a positional encoding vector to preserve relative order information between input tokens.

An end-to-end Transformer architecture typically includes both an encoder and decoder component. While critical for many NLP tasks such as machine translation, our architecture utilizes only the Transformer encoder, which encodes input sequences into hidden representations that are subsequently used for predicting patient mortality.

A comprehensive overview of the Transformer and BERT is beyond the scope of this section; we refer interested readers to Vaswani et al. (11) and Devlin et al. (12), respectively.

Briefly, the first stage of a Transformer encoder typically includes an embedding component, where each input sequence element is converted to a hidden representation that is fed into the remainder of the model. In its original NLP-centered design where inputs are sequences of textual tokens, a traditional embedding lookup table is employed to convert such tokens into continuous representations. Unlike similar sequential models like RNNs or CNNs, the Transformer is fundamentally temporally agnostic and processes all tokens simultaneously rather than sequentially. As such, the Transformer embedding module must inject some notion of temporality into its element embeddings. In typical Transformer implementations, this takes the form of a positional encoding vector, where the position of each element is embedded by sinusoidal lookup tables, which is subsequently added to the token embeddings. The primary aim of such positional embeddings is to allow the model to understand local temporality between nearby sequence elements.

At each layer of a Transformer encoder, a representation of every input sequence element is formed by summing self-attention compatibility scores between the element and every other element in the sequence. Typical with other deep learning architectures, as more layers are added to the encoder, the representations become more abstract.

The recent NLP method BERT (12) is based on Transformers, and at present time represent state of the art in a variety of natural language processing tasks. In addition to its novel pretraining scheme, BERT also prepends input sequences with a special [CLS] token before a sequence is passed through the model. The goal of this special token is to capture the combined representation of the entire sequence, and for classification tasks is used for making predictions.

Transformers are also being used in computer vision as well, with great success. For example, videos especially benefit from Transformers which can learn the temporal and spatial features of vision data. They have shown to before the same or better for vision tasks, while also reducing vision-specific induction bias Han et al. (24). For video data, they can be used for trajectory tracking of objects like balls Patrick et al. (25) using attention on objects in images, as well as approximate self attention to reduce quadratic dependency.

While the Transformer is in one sense more efficient than its sequential counterparts due to its ability to parallelize computations at each layer, one of the main drawbacks is its required memory consumption. Since each input element of a sequence of length n must be compared with every other input element in the sequence, typical Transformer implementations require memory on the order of $\mathcal{O}(n^2)$. While acceptable for relatively short sequences, the memory consumption quickly becomes problematic for very long sequences. Decreasing the memory requirement of Transformers is an area of ongoing research.

One potential solution was proposed by Beltagy et al. (18) in their Longformer architecture. Rather than computing full n^2 self-attentions, they propose a sliding self-attention window of specified width, where each input sequence element is compared only with neighboring sequence elements within the window. They extend this to include user-specified global attention patterns (such as on the special [CLS] tokens for classification) that are always compared with every element in the sequence. Through several NLP experiments, they demonstrate the promising ability of the Longformer to approximate results from a full Transformer model.

4.2.2. Transformers in healthcare

Given the similarity between textual sequences and temporal patient data contained in longitudinal EHR records, several works have begun exploring the efficacy of Transformers and modifications of BERT for clinical applications using electronic health records. In terms of patient data modalities, existing implementations of Transformers in a clinical setting tend to fall under three primary categories:

Perhaps the most aligned with the original BERT implementation, several studies adapt and modify BERT for constructing language models from unstructured text contained in clinical notes. The ClinicalBERT framework of Huang et al. (26) used a BERT model for learning continuous representations of clinical notes for predicting 30-day hospital readmission. Zhang et al. (27) pretrained a BERT model on clinical notes to characterize inherent bias and fairness in clinical language models.

Song et al. (17)'s SAnD architecture developed Transformer models for several clinical prediction tasks using continuous multivariate clinical time series.

The majority of existing EHR Transformer research has focused on temporal sequences of discrete EHR billing codes. Li et al. (16)'s BEHRT framework modified the BERT paradigm for predicting future disease from diagnosis codes. Med-BERT (15) demonstrated the performance advantages of a contextualized clinical pretraining scheme in conjunction with a BERT modification. RAPT (28) used a modified Transformer pretraining scheme to overcome several challenges with sparse EHR data. SETOR (29) utilized neural ordinary differential equations with medical ontologies to construct a Transformer model for predicting future diagnoses. RareBERT (30) extends Med-BERT for diagnosis of rare diseases. Meng et al. (31) used Transformers for predicting depression from EHR. Hi-BEHRT (16) extends BEHRT using a hierarchical design to expand the receptive field to capture longer patient sequences. Choi et al. (32) and Shang et al. (33)'s G-BERT architecture capitalize on the inherent ontological EHR structure.

In contrast to the isolated data modalities implemented in existing EHR Transformers, the novel embedding scheme utilized in our models combines both discrete and continuous patient data to generate a comprehensive representation of distinct clinical events and measurements.

4.5. Limitations

This feasibility study has several limitations and is intended as a methodological guiding framework for future multimodal and multi-task EHR Transformer research. Our retrospective dataset is limited to patients from a single-center cohort. Future work will evaluate performance in external validation cohorts such as MIMIC-IV (34). We also present results with parameters that maximize our limited hardware capacity; future work will focus on several hyperparameter tuning and model selection procedures. The baseline models we present for comparison are drawn from simplified implementations found in clinical deep learning research, and more recent approaches may offer enhanced predictive performance. From the results in **Table 3**, one might conclude that our EHR embedding procedure had a larger impact than use of the Transformer architecture, given the competitive AUROC of the attentional GRU baseline when implementing our tokenization pipeline for estimating risk of patient mortality. Future work will focus on disentangling the relative impacts of both model and data representation designs.

4.6. Conclusions and next steps

We feel there is still great potential for exploring additional benefits of our approach with diverse EHR data for a variety of clinical modeling and prediction tasks, especially in the realm of clinical interpretability. Given our promising pilot study results, future versions of this work will perform hyperparameter optimization with a focus on maximizing predictive accuracy. Additionally, since transformers are fundamentally composed of attention mechanisms, they can be analyzed with respect to particular outcomes, time points, or variables of interest to highlight important contributing factors to overall risk estimation. Future research will emphasize analyzing self-attention distributions between input variables and clinical outcomes to further the clinical explainability and enhance the clinical trust of Transformers in healthcare. We believe there is great potential for multimodal patient monitoring using flexible EHR frameworks such as ours. Future research will also focus on augmenting our multi-modal datasets with additional clinical data modalities such as clinical text and images, and pre-training our Transformer architectures with self-supervised prediction schemes across a variety of input data and clinical outcomes.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: UFHealth cohort data are available from the University of Florida Institutional Data Access/Ethics Committee for researchers who meet the criteria for access to confidential data and may require additional IRB approval. Requests to access these datasets should be directed to https://idr.ufhealth.org.

Author's contributions

Dr. Shickel conceived the model design, performed data extraction and processing, developed the data embedding pipeline and Transformer prediction framework, and performed Transformer experiments. Mr. Silva validated model performance, tuned hyperparameters, and trained baseline models for comparison with deep learning approaches. Mr. Khezeli, Dr. Tighe, and Dr. Bihorac reviewed study and manuscript for scientific accuracy. Dr. Rashidi conceptualized the study design and provided support and guidance. Dr. Shickel, Mr. Silva, Dr. Bihorac, and Dr. Rashidi had full access to the data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. Administrative, technical, material support, and study supervision, was provided by Dr. Bihorac and Dr. Rashidi. All authors contributed to the acquisition, analysis, and interpretation of data. All authors contributed to critical revision of the manuscript for important intellectual content. All authors contributed to the article and approved the submitted version.

Funding

BS was supported by R01GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS) and OT2OD032701 from the NIH Office of the Director (NIH/OD). TO-B was supported by K01DK120784, R01DK123078, and R01DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), R01GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01EB029699 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), OT2OD032701 from the NIH Office of the Director (NIH/OD), and UF Research AWD09459, and the Gatorade Trust, University of Florida. AB was supported by R01GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01EB029699 and R21EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), R01DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), and OT2OD032701 from the NIH Office of the Director (NIH/OD). PR was supported by National Science Foundation CAREER award 1750192, R01EB029699 and R21EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01GM110240 from the National Institute of General Medical Science (NIH/NIGMS), R01NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), R01DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), and OT2OD032701 from the NIH Office of the Director (NIH/OD). Additionally, the Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under University of Florida Clinical and Translational Science Awards UL1TR000064 and UL1TR001427.

Acknowledgments

We acknowledge the University of Florida Integrated Data Repository (IDR) and the UF Health Office of the Chief Data Officer for providing the analytic data set for this project. We thank the NVIDIA Corporation for their support through the Academic Hardware Grant Program.

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. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Med* (1996) 22:707–10. doi: 10.1007/BF01709751

2. Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, et al., Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. *Crit Care Med* (1998) 26:1793–800. doi: 10.1097/00003246-199811000-00016

3. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci Rep* (2019) 9:1879. doi: 10.1038/s41598-019-38491-0

4. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. *Proceedings of Machine Learning for Healthcare 2016 JMLR W&C Track 56*. Boston, MA: Proceedings of Machine Learning Research (2015). p. 1–12. doi: 10.1002/aur.1474.Replication

5. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc. (2016). p. 3512–3520.

6. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* (2016) 292:344–50. doi: 10.1093/jamia/ocw112

7. Sha Y, Wang MD. Interpretable predictions of clinical outcomes with an attention- based recurrent neural network. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,, Health Informatics*. New York, NY: Association for Computing Machinery (2017). p. 233–240

8. Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. In: *4th International Conference on Learning Representations*. San Juan, Puerto Rico (2016).

9. Lin L, Xu B, Wu W, Richardson T, Bernal EA. Medical time series classification with hierarchical attention-based temporal convolutional networks: a case study of myotonic dystrophy diagnosis. In: *CVPR Workshops*. New York, NY: Institute for Electrical and Electronics Engineers (2019). p. 83–86.

10. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: a convolutional net for medical records (2016). p. 1–9.

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30:5998–6008. doi: 10.1017/S0952523813000308

12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding [Preprint] (2018). Available at: arXiv:1811.03600v2.

13. Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting. *CoRR* (2022). Available at: arXiv:abs/2201.12740.

14. Li Y, Rao S, Roberto J, Solares A, Hassaine A, Ramakrishnan R, et al., BEHRT: transformer for electronic health records. *Sci Rep* (2020) 10:1–12. doi: 10.1038/s41598-020-62922-y

15. Rasmy L. Med-BERT: pretrained contextualized embeddings on large- scale structured electronic health records for disease prediction. *NPJ Digit Med* (2021) 4:1–13. doi: 10.1038/s41746-021-00455-y

16. Li Y, Mamouei M, Salimi-khorshidi G, Rao S, Hassaine A, Canoy D, et al. Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *arXiv* (2021).

17. Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend, diagnose: clinical time series analysis using attention models. In: *Thirty-second AAAI Conference on Artificial Intelligence*. Red Hook, NY: Curran Associates, Inc. (2018).

18. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. *arXiv* (2020).

19. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability, fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep* (2022) 12:1–28. doi: 10.1038/s41598-022-11012-2

20. Dorogush AV, Gulin A, Gusev G, Kazeev N, Prokhorenkova LO, Vorobev A. Fighting biases with dynamic boosting. *CoRR* (2017). Available at: abs/1706.09516.

21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *CoRR* (2016). Available at: abs/1603.02754.

22. Qin Y, Song D, Cheng H, Cheng W, Jiang G, Cottrell GW. A dual-stage attention-based recurrent neural network for time series prediction. *International Joint Conference on Artificial Intelligence (IJCAI)*. Red Hook, NY: Curran Associates, Inc. (2017). p. 2627–2633.

23. Sukhbaatar S, Szlam A. End-to-end memory networks. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc. (2015). p. 2440–2448.

24. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell* (2022): 1–1. doi: 10.1109/TPAMI.2022.3152247. https://ieeexplore.ieee.org/document/9716741

25. Patrick M, Campbell D, Asano Y, Misra I, Metze F, Feichtenhofer C, et al. Keeping your eye on the ball: trajectory attention in video transformers. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc. (2021). p. 12493–12506.

26. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv* (2019).

27. Zhang H, Lu AX, Mcdermott M. HurtfulWords: quantifying biases in clinical contextual word embeddings. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. New York, NY: Association for Computing Machinery (2020). p. 110–120.

28. Ren H, Wang J, Zhao WX. RAPT: pre-training of time-aware transformer for learning robust healthcare representation. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York, NY: Association for Computing Machinery (2021). p. 3503–3511.

29. Peng X, Long G, Shen T, Wang S, Jiang J. Sequential diagnosis prediction with transformer and ontological representation. *arXiv* (2021).

30. Prakash P, Chilukuri S, Ranade N, Viswanathan S. RareBERT: transformer architecture for rare disease patient identification using administrative claims. In: *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) RareBERT*. Palo Alto, CA: AAAI Press (2021). p. 453–460.

31. Meng Y, Speier W, Ong MK, Arnold CW. Transformers using multimodal electronic health record data to predict depression. *IEEE J Biomed Health Inform* (2021) 25:3121–9. doi: 10.1109/JBHI.2021.3063721

32. Choi E, Xu Z, Li Y, Dusenberry MW, 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*. Palo Alto, CA: AAAI Press (2020). p. 606–613.

33. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. *arXiv* (2019).

34. Johnson AEW, Stone DJ, Celi LA, Pollard TJ. The mimic code repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc* (2018) 25:32–9. doi: 10.1093/jamia/ocx084

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 21 November 2022
DOI 10.3389/fdgth.2022.997219

OPEN ACCESS

EDITED BY
Max Little,
University of Birmingham, United Kingdom

REVIEWED BY
Tyler John Loftus,
University of Florida, United States,
Inmaculada Mora-Jiménez,
Rey Juan Carlos University, Spain

*CORRESPONDENCE
William Bolton
william.bolton@imperial.ac.uk

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 18 July 2022
ACCEPTED 27 October 2022
PUBLISHED 21 November 2022

CITATION
Bolton WJ, Rawson TM, Hernandez B, Wilson R, Antcliffe D, Georgiou P and Holmes AH (2022) Machine learning and synthetic outcome estimation for individualised antimicrobial cessation.
Front. Digit. Health 4:997219.
doi: 10.3389/fdgth.2022.997219

Machine learning and synthetic outcome estimation for individualised antimicrobial cessation

William J. Bolton[1,2,3]*, Timothy M. Rawson[1,4], Bernard Hernandez[1,5], Richard Wilson[1,4], David Antcliffe[6,7], Pantelis Georgiou[1,5] and Alison H. Holmes[1,4,8]

[1]Centre for Antimicrobial Optimisation, Imperial College London, London, United Kingdom, [2]AI4Health Centre for Doctoral Training, Imperial College London, London, United Kingdom, [3]Department of Computing, Imperial College London, London, United Kingdom, [4]National Institute for Health Research, Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, London, United Kingdom, [5]Centre for Bio-inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom, [6]Department of Critical Care, Imperial College Healthcare NHS Trust, London, United Kingdom, [7]Faculty of Medicine, Imperial College London, London, United Kingdom, [8]Department of Infectious Diseases, Imperial College London, London, United Kingdom

The decision on when it is appropriate to stop antimicrobial treatment in an individual patient is complex and under-researched. Ceasing too early can drive treatment failure, while excessive treatment risks adverse events. Under- and over-treatment can promote the development of antimicrobial resistance (AMR). We extracted routinely collected electronic health record data from the MIMIC-IV database for 18,988 patients (22,845 unique stays) who received intravenous antibiotic treatment during an intensive care unit (ICU) admission. A model was developed that utilises a recurrent neural network autoencoder and a synthetic control-based approach to estimate patients' ICU length of stay (LOS) and mortality outcomes for any given day, under the alternative scenarios of if they were to stop vs. continue antibiotic treatment. Control days where our model should reproduce labels demonstrated minimal difference for both stopping and continuing scenarios indicating estimations are reliable (LOS results of 0.24 and 0.42 days mean delta, 1.93 and 3.76 root mean squared error, respectively). Meanwhile, impact days where we assess the potential effect of the unobserved scenario showed that stopping antibiotic therapy earlier had a statistically significant shorter LOS (mean reduction 2.71 days, p-value <0.01). No impact on mortality was observed. In summary, we have developed a model to reliably estimate patient outcomes under the contrasting scenarios of stopping or continuing antibiotic treatment. Retrospective results are in line with previous clinical studies that demonstrate shorter antibiotic treatment durations are often non-inferior. With additional development into a clinical decision support system, this could be used to support individualised antimicrobial cessation decision-making, reduce the excessive use of antibiotics, and address the problem of AMR.

KEYWORDS

antimicrobial resistance, artificial intelligence, clinical decision support systems, decision-making, individualised antimicrobial prescribing, precision prescribing, antibiotic cessation, outcome estimation

Introduction

Bacterial antimicrobial resistance (AMR) is a global threat (1, 2), which resulted in an estimated 1.27 million deaths in 2019 (3). One key strategy to tackle AMR is to optimise antimicrobial use and prolong current antimicrobials' therapeutic life. Clinical decision support systems (CDSSs) are software designed to provide information to healthcare professionals, patients, or other individuals in order to make informed clinical decisions. With the advent of artificial intelligence (AI) and the ever increasing prevalence of electronic health records (EHRs), numerous CDSSs utilising machine learning (ML) trained on historical patient data have been developed to assist with managing infections (4). Recent research has focused on the diagnoses of bacterial infections (5–7), resistance prediction (8), and antimicrobial therapy selection (9, 10).

One challenge when treating a patient who has a bacterial infection is determining when it is appropriate to stop antibiotic treatment (11). The decision to cease antibiotics too early can result in the patient's condition worsening, while unnecessary exposure increases the risk of toxicity (12) and drives the evolution of AMR (13). Even over-treating for a short duration can have a significant impact on a population level and enhances the development of resistance (14). Furthermore, excessive treatment is responsible for most avoidable antibiotic adverse events including gastrointestinal distress and allergic reactions (15, 16). Numerous studies have shown that on a population level, shorter treatment durations are often non-inferior to longer ones (17–21). The challenge is that the resulting recommendations do not take into account the individual patient's characteristics or specific scenarios. It is difficult for clinicians to have confidence in individualised treatment decisions for their patient, when there is a poor understanding of the factors that facilitate or inhibit an individual from receiving a short duration of antibiotic therapy. Therefore, durations are often unnecessarily extended (22) and decided by habit or arbitrarily based on population evidence. Antibiotic cessation should be a collective, data-driven decision, given choices are made in a more favourable environment once time has passed from presentation and significant amounts of information have been gathered. Despite this, systems to help support individualised antibiotic duration and cessation decision-making are often neglected and under-researched with little innovation in this area (23, 24).

Given the current standard of care uses clinical factors to determine if a patient should stop antibiotics or not, we hypothesise that an AI-based CDSS using routinely collected EHR data may be able to support individualised antibiotic cessation decision making and overcome prescriber concerns of poor patient outcomes that is likely a major driver of over treatment (25, 26). We approach this problem by estimating clinical outcomes under alternative scenarios with the aim of showing non-inferiority or a direct benefit of antibiotic cessation. More specifically, a machine learning and synthetic control-based approach was developed to estimate patients' LOS and mortality outcomes for any given day, if they were to stop vs. continue antibiotic treatment. **Figure 1** shows a graphical abstract of the approach and methodology employed in this retrospective research study.

Methods

Dataset

MIMIC-IV is a large de-identified real-world clinical dataset that is publicly available for clinical research (27, 28). It contains EHR information for over 40,000 patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, Unites States, between 2008 and 2019. The patient population was filtered to those who received intravenous antibiotic treatment for a duration between 1 and 21 days during an ICU stay. Input features were extracted, analysed, and selected based on prevalence, correlation, as well as infectious disease doctors and critical care consultants advice. Length of stay (LOS) (continuous value) and mortality (binary) labels were extracted for each patient stay; however, it should be noted that these are not temporally dynamic. An overview of statistics for each dataset is shown in **Table 1**

Some features were calculated based on other variables. Cumulative overall antibiotic treatment length was determined for each day of each ICU stay that considered consecutive treatment days irrespective of the antibiotic given. In addition, whether the patient had received re-treatment for antibiotics or not and their age at the time of ICU admission were also computed. Standard pre-processing was applied to features including outliers being removed and values normalised, as well as missing values forward filled or highlighted. Features were aggregated by day for each unique stay to create a regular temporal dataset. In general, there was a high degree of missingness, and so patients with greater than 50% of values missing each day were removed. The resulting dataset contained 43 input features (**supplementary Table S1**) including lab test results, clinical parameters, ventilation settings, and demographics.

Model architecture

The objective of our model is to estimate the patients' LOS and mortality outcomes for any given day, if they were to stop

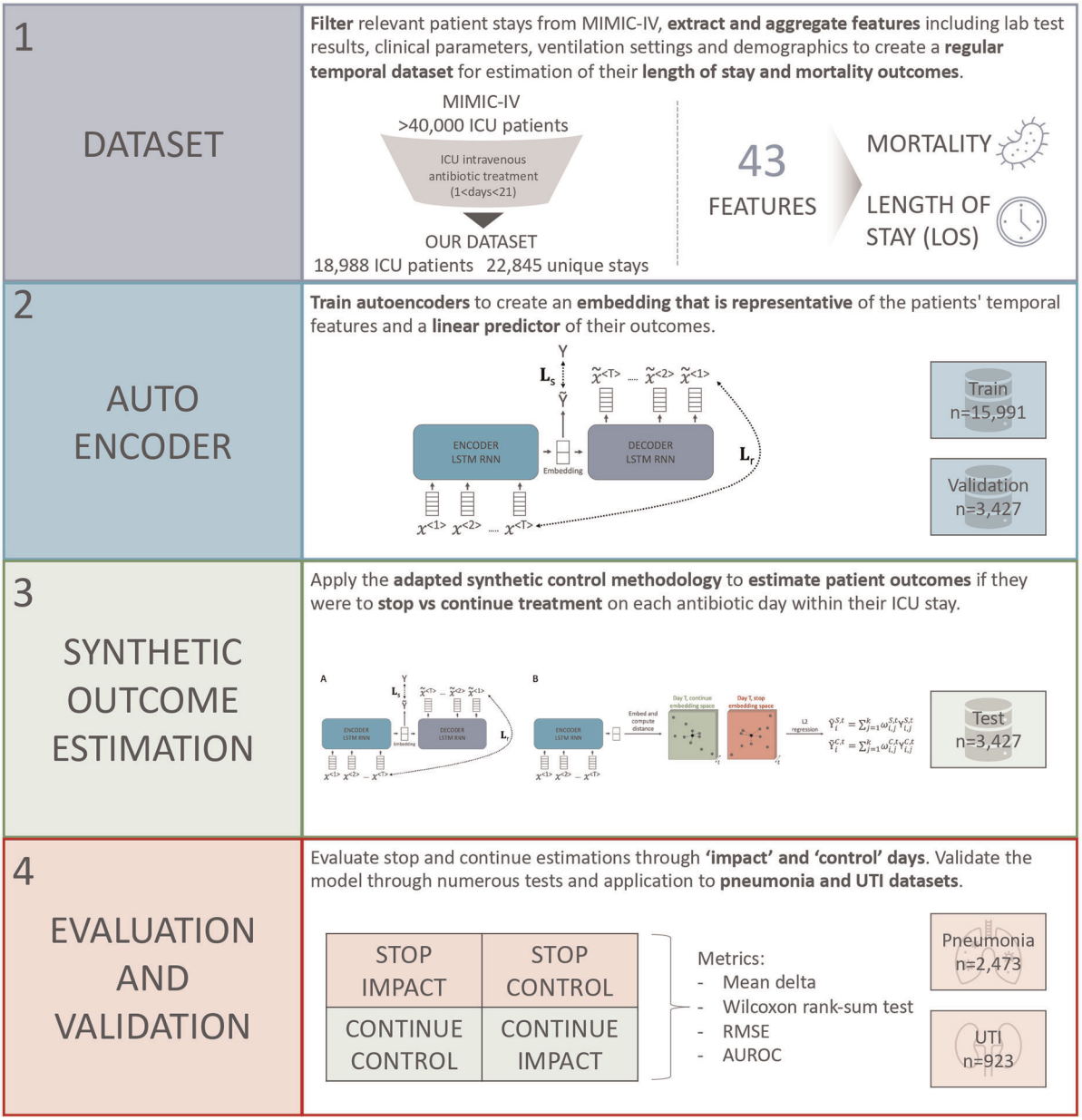


FIGURE 1
Overview of the steps taken in this research study to develop a model for antimicrobial cessation synthetic outcome estimation.

vs. continue antibiotic treatment. It uses a bi-directional long short-term memory (LSTM) autoencoder, which takes in a sequence of patient input features ($x^{\langle 1\rangle}, x^{\langle 2\rangle} \ldots x^{\langle T\rangle}$), creates an embedding representation, and outputs a sequence of reconstructed features ($\tilde{x}^{\langle T\rangle} \ldots \tilde{x}^{\langle 2\rangle}, \tilde{x}^{\langle 1\rangle}$). This autoencoder is trained through two loss functions (29), which are summed together to create a combined loss for backpropagation. First, the reconstruction loss L_r is calculated by the root mean squared error (RMSE) between outputs that are trying to reproduce the inputs and the real input data. Second, a supervised learning loss L_s is calculated by doing a linear transformation of the embedding representation ($\tilde{Y}$) to try and predict the real label (Y) and taking either the RMSE loss for the LOS outcome or the binary cross-entropy loss for mortality classification. L_s ensures that the embedding space created by the autoencoder is a good linear predictor of the outcome of interest, which is important for the subsequent adapted synthetic control method. Overall, an embedding representation is created that considers a patient's past and is representative of their state on that day.

TABLE 1 Datasets statistics.

Statistic	Dataset					
	Overall	Train	Validation	Test	Pneumonia	UTI
Number of stays	22,845	15,991	3,427	3,427	2,473	923
Mortality rate	18.60	18.47	18.30	19.52	24.02	18.96
Mean LOS	5.63	5.62	5.74	5.55	9.05	5.50
LOS standard deviation	4.23	4.24	4.31	4.15	5.19	4.32
Mean length of treatment	4.38	4.38	4.48	4.30	6.95	4.77
Length of treatment standard deviation	3.32	3.32	3.48	3.18	4.28	3.55
Spearman's correlation between LOS and treatment length	0.72	0.72	0.72	0.73	0.73	0.74
Percentage of patients that stopped treatment during their ICU stay	41.56	41.64	41.17	41.55	31.95	26.54

LOS, length of stay; ICU, intensive care unit; UTI, urinary tract infection.

Once the antoencoder is trained and an embedding representation for each antibiotic day in all patient stays have been created, an adapted synthetic control approach (30) is utilised, where the act of stopping or continuing treatment on a particular day is considered an intervention and each patient acts as a singular unit. This method is useful when evaluating an intervention using randomised controlled trials is challenging, as is the case with antibiotic cessation, and hence retrospective observational data are assessed. Synthetic controls have frequently been applied to understand public health interventions (31, 32), but their use within digital health research is limited. In this study, we want to know what are the predicted outcomes if a given patient was to stop vs. continue antibiotics on a given day within their ICU stay. To this extent, two synthetic controls are created, one can be labelled the "stop synthetic control," which is based on subjects who stopped antibiotics on that particular day, and the second labelled the "continue synthetic control," which is created from subjects who continue antibiotic treatment on that particular day. To achieve this for each day (t), two separate donor pools are created based on subjects associated embedding representation and antibiotic treatment status. In other words, those who continue antibiotics on day t are partitioned into the "continue" embedding space while those who stop antibiotics are placed in the "stop" embedding space. In this way, the estimated outcomes for stopping and continuing on day t are driven by representative donors who experienced analogous treatment. To create the stop and continue synthetic controls for a particular patient i, the k most closely related to embedding representations from each relevant donor pool are selected based on a distance metric (in this study $k = 10$ and Euclidean distance were used for both stop and continue estimations). Given that embeddings are representative of the patients' state, those selected donors will be similar, giving a considered insight into potential alternative outcomes under antibiotic temporality. A ridge regression function ($Loss_i^{S,t} = \sum_{d=1}^{D} [z_{i,d}^t - \sum_{j=1}^{k} x_{j,d}^{S,t} w_{i,j}^{S,t}]^2 + \sum_{j=1}^{k} w_{i,j}^{S,t\,2}$ for stop estimations and $Loss_i^{C,t} = \sum_{d=1}^{D} [z_{i,d}^t - \sum_{j=1}^{k} x_{j,d}^{C,t} w_{i,j}^{C,t}]^2 + \sum_{j=1}^{k} w_{i,j}^{C,t\,2}$ for continue estimations, where d are the embedding dimensions, j are the donors, and z represents the particular patient $i's$ embedding for a given dimension and time) is then applied to the subject and their respective stop and continue donor embeddings. This returns two sets of weights ($w_{i,j}^{S,t}$ for "stop" and $w_{i,j}^{C,t}$ for "continue") that minimise the square difference between the subject of interest and the selected units in the donor pools ($Y_{i,j}^{S,t}$ for "stop" and $Y_{i,j}^{C,t}$ for "continue"). The objective of this L2 regularisation is to fairly distribute weights across the donors for stop and continue estimations. Finally, the stop and continue synthetic control outcomes ($\tilde{Y}_i^{S,t}$ and $\tilde{Y}_i^{C,t}$, respectively) for the particular patient i are computed from the weighted average of donor labels. To this extent during outcome estimation for a given patient i, we assume that we know the outcomes for all other patients within the dataset. Overall outcomes are estimated for each patient on each relevant antibiotic day of their stay if they were to stop vs. continue antibiotic treatment. An overview of the model's architecture and this process for stop and continue outcome estimation is shown in Figure 2.

Model development and software

The model was applied on the MIMIC-IV EHR dataset, which was randomly split based on patients' "stay_id" into training, validation, and testing sets (70%, 15%, and 15%, respectively). PyTorch (33) was used to create a bi-directional LSTM recurrent neural network (RNN) with a custom dataset class to extract labels and features. In order to address the mortality class imbalance (Table 1), over-sampling was used during training. To be specific, those cases with positive mortality were replicated three times within the custom dataset class to achieve a more balanced mortality rate of

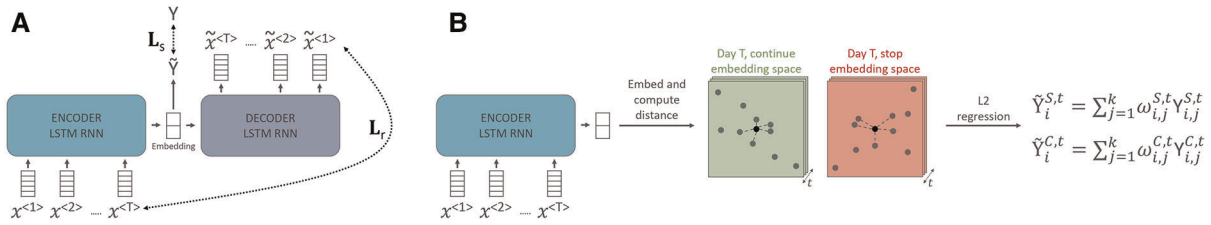


FIGURE 2
Model illustration. (A) The encoder is trained using both a supervised loss (Ls) and reconstruction loss (Lr) (3). (B) To estimate outcomes during testing, an embedding is created for every day of each patient's stay; embedding spaces are partitioned temporally and based on if the patient stopped or continued antibiotics. The closest *k* neighbours are selected as donors from each embedding space and L2 regression returns weights that minimise the square difference between the patient and the donors. A stop and continue synthetic control outcome is estimated as a weighted average of the donors' outcomes.

51.90% within the train dataset. The Adam optimiser (34) was used with binary cross-entropy loss for classification, mean squared error loss for regression, and Ray Tune for hyperparameter optimisation (35). Training utilised 50 epochs, during which the model with the best performance on the validation dataset (RMSE or area under the receiver operating characteristic curve for LOS and mortality prediction, respectively) was selected as the final model. Two separate LSTM autoencoder models were trained on the whole training dataset to create embedding representations relevant to patients' LOS and mortality outcomes. Models were evaluated using functions and metrics from the TorchMetrics, Scikit-learn, and SciPy libraries. Further details of the two models' hyperparameters and their optimisation are shown in the supplementary material (**supplementary Figure S1 and Table S2**).

Model evaluation and metrics

Commonly with the synthetic control method, the delta difference between the single unit and the counterfactual in the pre-intervention period is minimised and the treatment effect is then observed in the post-intervention period. For our research question, this is not possible due to the nature of stopping antibiotics being the final event at one point in time, after which the patient is not applicable to our research population or question. An analogue can be applied for this study where we define "control" and "impact" days that are equivalent to the pre- and post- intervention periods. For estimating outcomes when continuing antibiotics, all the days the patient actually continues antibiotics are "control" days where we expect minimal difference between the true and estimated outcomes. On the other hand, on the single day the patient stops antibiotics, we can assess the "impact" if they were to instead continue. When estimating outcomes upon stopping antibiotics, the reverse is true, whereby each day antibiotics were continued the "impact" of stopping can be assessed and the final day where the patient stops treatment acts as a "control." Note that it is not possible to define this for every patient, given not every individual will stop antibiotics during their ICU stay. The percentage of patients who stopped antibiotic treatment during their ICU stay is shown in **Table 1**. Outcomes are estimated in the same way for impact and control days as discussed in the "Model architecture" subsection. However, for control days, we know the real outcome and so can compare our estimations, while for impact days, the real outcome is unknown. Each day, therefore, acts as both a "control" and "impact" across the two "stop" and "continue" scenario outcome estimations. An outline of this is shown in **Figure 3** and the number of continue and stop donors for each day in the test dataset is illustrated in **supplementary Figure S2**.

For outcome estimation, the mean delta is calculated to evaluate the difference between the real labels and the estimations, through the following formula: $\mu\Delta^S = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}[Y_i^{S,t} - \tilde{Y}_i^{S,t}]]$ for stop estimations and $\mu\Delta^C = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}[Y_i^{C,t} - \tilde{Y}_i^{C,t}]]$ for continue estimations, where T_i is the number of days that the patient receives antibiotics. Minimal difference should be seen on control days where our model aims to reproduce labels, while on impact days you can assess the effect of the unobserved scenario. Statistical analysis can be used to determine if the difference between the true LOS labels and the estimated outcomes are statistically significant. Given the non-normal data distribution, the non-parametric Wilcoxon rank-sum (Mann–Whitney U) test was used with the alpha set at 0.05. Furthermore, the mean absolute percentage error (MAPE) and mean absolute error (MAE) can be calculated through the following notations: $\text{MAPE}^S = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}|Y_i^{S,t} - \tilde{Y}_i^{S,t}|/Y_i^{S,t}]$ and $\text{MAE}^S = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}|Y_i^{S,t} - \tilde{Y}_i^{S,t}|]$, respectively, for stop estimations and $\text{MAPE}^C = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}|Y_i^{C,t} - \tilde{Y}_i^{C,t}|/Y_i^{C,t}]$ and $\text{MAE}^C = (1/n)\sum_{i=1}^{n}[(1/T_i)\sum_{t=1}^{T_i}|Y_i^{C,t} - \tilde{Y}_i^{C,t}|]$, respectively, for continue estimations. Standard ML metrics can also be used to evaluate model prediction

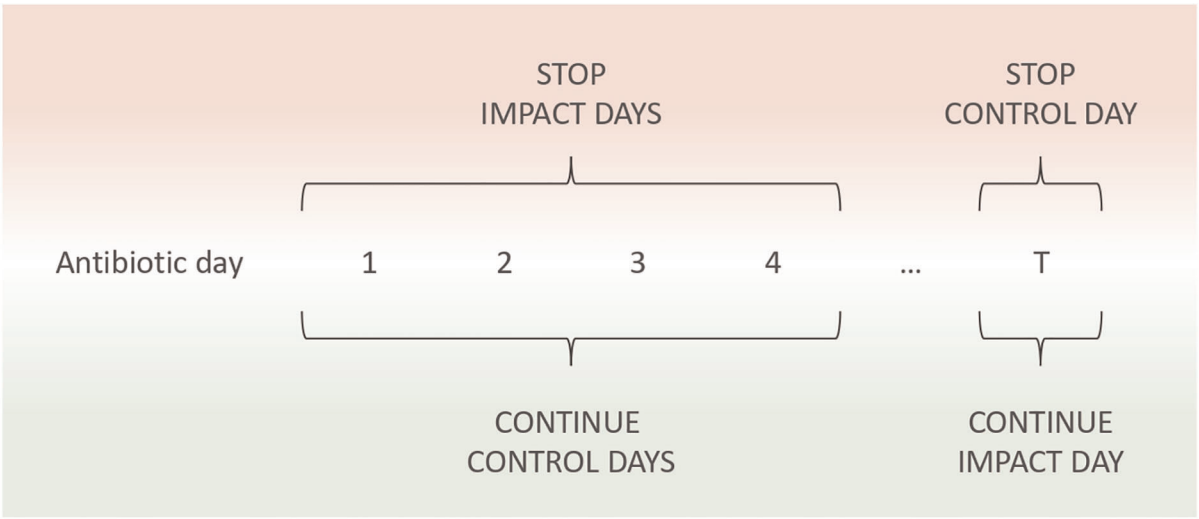


FIGURE 3
Demonstration of the impact and control evaluation process for stop and continue scenarios. An antibiotic day is defined as each day the patient receives treatment as well as the day they stop. After starting antibiotics, each day the patient receives treatment acts as a stop impact and continue control. This continues until antibiotic cessation or ICU discharge. If the patient stops antibiotics during their ICU stay, that initial day where no antibiotics are administered acts as a stop control and a continue impact.

performance. For LOS regression estimation, the RMSE is used, while for the mortality classification task, Area Under the Receiver Operating Characteristic curve (AUROC) is most appropriate given the class imbalance (Table 1), but accuracy, precision, recall, sensitivity, F1 score, and Area Under the Precision Recall curve (AUPRC) can also be calculated. Metrics were calculated as global averages, across all samples, meaning every day of antibiotic treatment within each patients stay is considered equally. 95% confidence intervals were calculated through 1,000 bootstrapped samples on the test set with $n = 1,000$ for mortality metrics and the sum of the squared errors method for LOS RMSE.

To validate our findings beyond the hold out test set, we applied our model to patients who were diagnosed with pneumonia or a urinary tract infection (UTI). The effects of short vs. longer antibiotic treatment regimes have been extensively studied in pneumonia and UTIs. In general, research supports the notion that shorter antibiotic treatments durations are non-inferior to longer ones in these infections, especially for non-complicated cases (19, 36–40). Based on this evidence and the latest antimicrobial prescribing guidelines (41, 42), we defined a long treatment duration as any patient receiving antibiotics for longer than 7 days, and applied our model to estimate their outcomes if they were to instead stop treatment after 7 days. In addition, there is increasing evidence that even shorter courses of antibiotics can be used in such infections, in particular, pneumonia (19, 41). Hence, we investigated the estimated outcomes of those patients who received the standard of care 7 days treatment, for slightly shorter treatment durations (5 or 6 days).

Results

Autoencoder

In total, 18,988 patients, associated with 22,845 unique ICU stays, were included across datasets. Through a linear transformation of a given patient day embedding, outcome estimations could be made on the unseen test set (3427 unique ICU stays). The LOS model achieved an RMSE of 3.88 (95% CI 3.84–3.92), while the mortality estimation model obtained an AUROC of 0.77 (95% CI 0.73–0.80) [accuracy 0.73 (95% CI 0.71–0.75), precision 0.44 (95% CI 0.36–0.46), recall 0.67 (95% CI 0.61–0.72), specificity 0.75 (95% CI 0.72–0.78), F1 0.53 (95% CI 0.46–0.56), and AUPRC 0.55 (95% CI 0.42–0.56)] (Figure 4), indicating that the model was relatively effective at balancing false-positive and false-negative mortality predictions.

Synthetic outcome estimation

LOS and mortality estimation results on the unseen test set are shown in Table 2. For LOS estimation on control days, the mean delta under both stopping and continuing scenarios was 0.24 and 0.42 days, respectively, showing a minimal difference

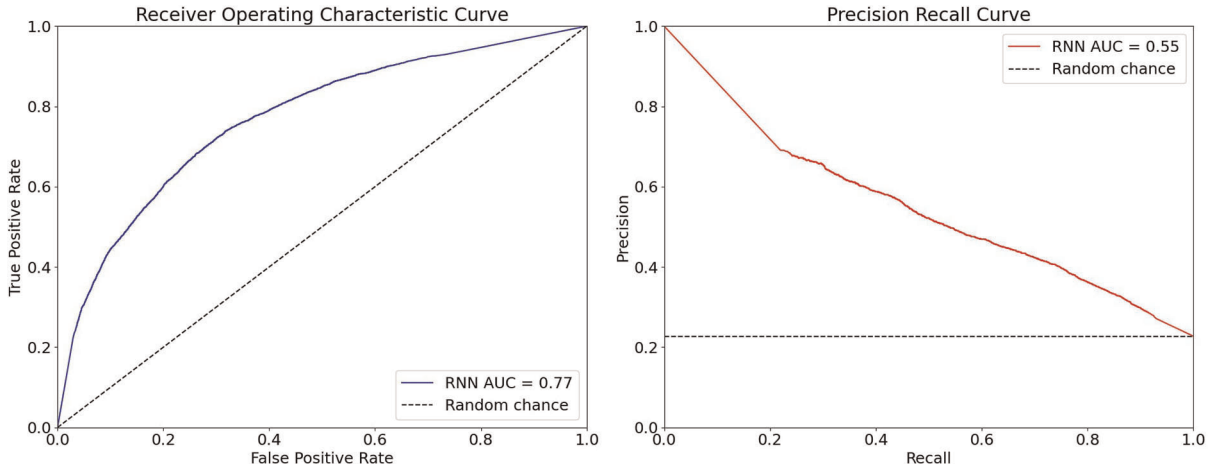


FIGURE 4
ROC and PRC results for the RNN autoencoder on mortality classification.

between predictions and the ground truth labels. Furthermore, a MAPE of 0.26, MAE of 1.32, and RMSE of 1.93 for stop control days show that the corresponding impact estimations are more reliable. On impact days, stopping earlier had a statistically significant shorter LOS (mean difference 2.71 days, p-value <0.01). This indicates that on average LOS estimations for stopping antibiotics earlier are shorter in duration than those when the patient continues antibiotics. For mortality, no impact was observed by stopping or extending antibiotic treatment. Estimations had modest performance with an average AUROC of 0.67 and accuracy of 0.82; however, the model clearly struggled with false-negative predictions.

Estimations were made for each day of each patient's stay within all the extracted data (i.e., train, validation, and testing sets combined) to understand if results would deviate by dataset size. For LOS, reliable estimations were once again obtained (mean stop control difference of 0.33 days and mean continue control difference of 0.42 days). Continuing showed no given impact (mean difference of −0.30 days), while stopping once again showed a significant impact with a mean reduction of 1.87 days. Little difference in mortality estimations was seen between stop and continue controls and impacts (stop impact−0.03, stop control−0.03, continue control−0.05, continue impact−0.05). Mortality predictions were relatively reliable with a mean AUROC of 0.72.

To show the importance of the temporality in our predictions, we created estimations for each antibiotic day of each patients stay, without segregating the embedding space (by time or by antibiotic treatment given they are mutually dependent). The resulting estimations had a mean LOS difference of 2.60 days from the true labels, an RMSE of 5.05, and a statistically significant difference in medians (p-value <0.01).

The performance of the model on subjects towards the edges of the distribution in terms of the correlation between LOS and overall antibiotic treatment length was investigated. Subjects in the 10th and 90th percentiles were selected leading to a smaller Spearman's correlation of 0.35. As expected, given the dataset size (n=686) and donor distribution, results were quite poor with a mean stop control difference of 2.92 days and a mean continue control difference of 2.13 days. The impact of stopping early though was still much greater than the control at 4.36 days mean difference.

Pneumonia and UTIs

A total of 2,473 stays where patients were diagnosed with pneumonia were identified, with a mean LOS of 9.05 days and a mean antibiotic treatment length of 6.95 days. Overall estimation of the results on this whole pneumonia population reflected the wider dataset and are shown in **Table 3**. When focusing on those with long treatment durations and the question of what if they stopped after 7 days of treatment, statistically significant results show that average LOS were 2.82 days shorter when stopping earlier. No difference in estimated mortality was observed; however, estimations were consistent across groups with an average AUROC of 0.75. No significant difference in LOS or mortality was estimated for pneumonia patients who received the standard of care 7 days treatment, if they had slightly shorter treatment durations of 5 or 6 days.

For UTIs, 923 patient stays were selected having a mean LOS and antibiotic treatment length of 5.50 and 4.77 days respectively. Once again, overall estimation results (**Table 3**) were similar to previous findings with trustworthy controls, stopping early being associated with a shorter LOS and no

TABLE 2 Outcome estimation results for patients in the unseen test set.

		LOS				Mortality		
Scenario	Day(s)	Mean delta (days, *p*-value)	MAPE	MAE	RMSE	Mean delta	MAE	AUROC
Stop	Impact	2.71*, <0.01	0.36	3.30	4.80	0.06	0.25	0.66
	Control	0.24, 0.60	0.26	1.32	1.93	0.05	0.15	0.72
Continue	Impact	−2.09*, <0.01	0.77	2.85	3.16	0.05	0.18	0.67
	Control	0.42*, 0.01	0.48	2.72	3.76	0.07	0.24	0.64

*Statistical significance with alpha set at 0.05.
LOS, length of stay; MAPE, mean absolute percentage error; MAE, mean absolute error; RMSE, root mean squared error; AUROC, Area Under the Receiver Operating Characteristic curve.

TABLE 3 Outcome estimation results for patients with pneumonia and UTIs.

				LOS		Mortality	
Infection	Analysis	Scenario	Day (s)	Mean delta (days, *p*-value)	RMSE	Mean delta	AUROC
Pneumonia	Whole dataset	Stop	Impact	3.72*, <0.01	5.87	0.00	0.71
			Control	0.26, 0.47	2.14	0.07	0.76
		Continue	Impact	−2.79*, <0.01	3.65	0.10	0.69
			Control	0.49*, <0.01	4.01	0.05	0.68
	Long treatment durations stopping after 7 days	Stop	Impact	2.82*, <0.01	4.65	−0.03	0.74
			Control	0.43, 0.08	2.11	0.05	0.80
		Continue	Impact	—	—	—	—
			Control	0.41, 0.21	3.47	0.05	0.73
UTI	Whole dataset	Stop	Impact	2.36*, <0.01	4.70	0.14	0.63
			Control	0.36, 0.89	2.04	0.07	0.87
		Continue	Impact	−1.91*, <0.01	3.26	0.03	0.79
			Control	0.38, 0.05	3.82	0.04	0.71
	Long treatment durations stopping after 7 days	Stop	Impact	2.08*, <0.01	4.35	0.30	0.52
			Control	1.04, 0.23	2.42	0.17	0.93
		Continue	Impact	—	—	—	—
			Control	0.26, 0.05	3.48	0.05	0.76

Results are shown for both the whole population and analysis of what if those who received long treatment durations stopped after day 7.
*Statistical significance with alpha set at 0.05.
LOS, length of stay; RMSE, root mean squared error; AUROC, Area Under the Receiver Operating Characteristic curve; UTI, urinary tract infection.

difference in mortality but reliable estimations (AUROC ranging from 0.63 to 0.87). Estimations for stopping after 7 days for those with long treatment durations did show a positive impact in terms of reduced LOS (mean difference 2.08 days, *p*-value <0.01). The stop control where we expect to see minimal difference showed a larger mean deviation of 1.04 days, but statistical analysis showed the medians between control estimations and labels were not significantly different. Mortality estimations here were for the most part dependable; a high predictive performance on stop and continue controls was achieved with an AUROC of 0.93 and 0.78, respectively, but a lower score for the stop impact of 0.52. When analysing those patients who received the standard of care 7 days treatment, for slightly shorter treatment durations (5 or 6 days). A statistically significant result was observed where estimated LOS outcomes were on average longer by 1.45 days if the patients stopped antibiotics slightly earlier (*p*-value <0.01, RMSE 2.72).

Discussion

We demonstrate that our RNN autoencoder and synthetic control-based approach trained on a large ICU EHR dataset can estimate patient outcomes under the alternative scenarios of stopping vs. continuing antibiotic treatment. Results across experiments were consistent, with stop control days often showing the greatest performance indicating our stop impact estimations, which occur on days where the true outcome upon stopping is unknown, are more reliable. The stop impact results from this retrospective study show that stopping antibiotics earlier can be associated with a statistically significant average LOS reduction of 2.71 days. Overall minimal impact on mortality was observed, which is to be expected given death can be caused by a large number of factors beyond those included as model features. **Figure 5** shows some specific illustrative examples of patient LOS and mortality estimations. The pneumonia dataset demonstrated

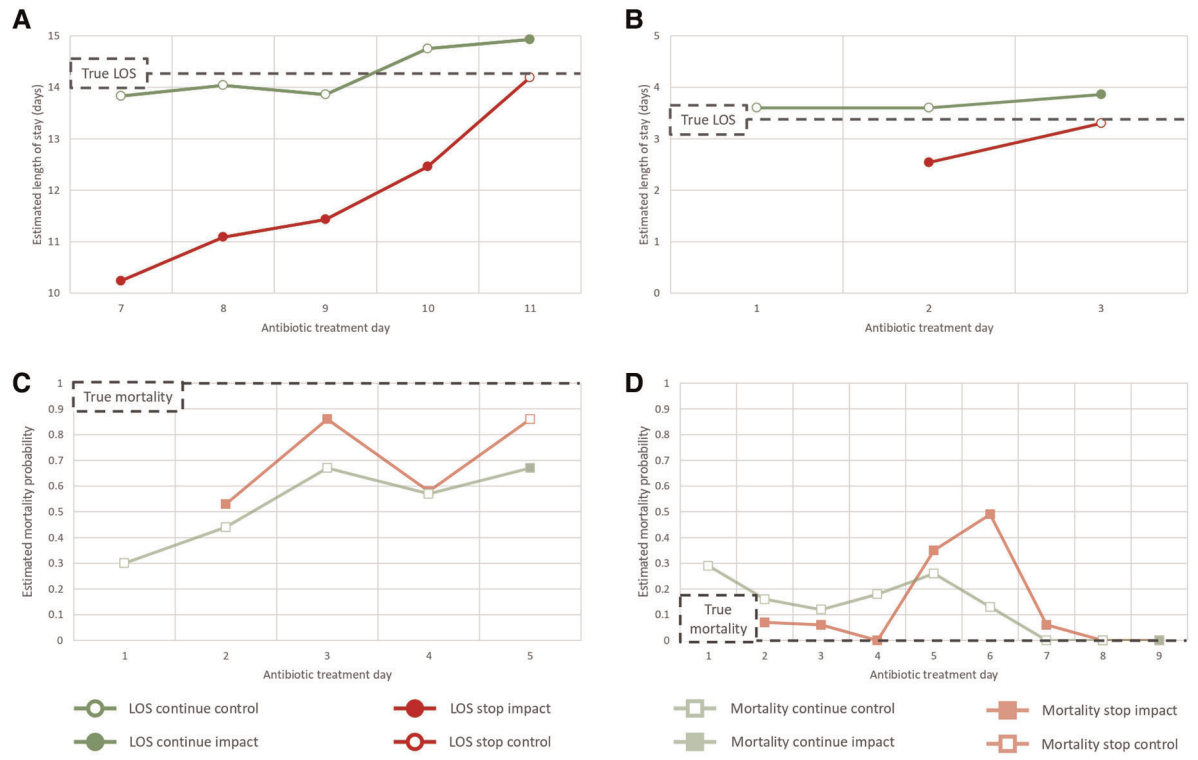


FIGURE 5
LOS and mortality synthetic outcome estimation results for particular patients. These cases were selected as illustrative examples of four distinct patient scenarios: (A) the patient has a long course of antibiotics, (B) the patient has short course of antibiotics, (C) the patient dies, (D) the patient survives. In A/B control estimation results show minimal deviation from the true LOS label while the stop impact estimations have a reduced LOS. Results in C/D indicate mortality estimations are temporally dynamic but with little difference between stop vs. continue.

particularly positive results with overall and stopping on day 7 analysis indicating antibiotic cessation can have a significant impact on LOS in this population (mean difference 3.72 and 2.82 days, respectively). This reflects current clinical thinking that shorter treatments are optimal for this infection (19, 36, 37, 41). However, there is a balance to be made with antibiotic treatment durations. The UTI analysis indicated courses shorter than 7 days may be detrimental to the patient and that the current standard of care treatment duration is likely appropriate. As such, care must be taken to consider the patients and the public's best interests with respect to current infections and the threat of AMR.

Our methodological approach to the problem of antibiotic cessation is novel. This model can in principal assist with individualised antibiotic cessation decisions as it takes into account numerous patient characteristics and the specific treatment scenario with regards to patient outcomes, factors that previously could not be considered together in their entirety. This study has approached the problem of antibiotic cessation from the perspective of making a clinically useful tool designed to support decision-making by estimating direct measures that may influence clinical decision-making under alternative scenarios. We believe it could be useful for prescribing physicians during their daily clinical round to compare between stop and continue estimated outcomes and understand when it is appropriate to cease antibiotic treatment. In particular, this system should help show shorter treatment durations can be safe and support individualised antimicrobial decision-making through hard outcome estimation. From a behaviour change perspective, this approach may provide reassurance to support early cessation of therapy, while promoting improved knowledge and understanding on the issue of antimicrobial optimisation and stewardship (43). It should be noted though that too short a course of antibiotics can cause harm and have negative knock-on effects. As such, the aim of this research is to optimise antimicrobial use and determine the most appropriate antibiotic treatment duration for each individual patient. One significant outstanding question is how clinicians treating a patient would adopt recommendations provided by such a system and if it would influence antimicrobial clinical decision-making. Holistically, we believe antibiotic cessation is a collective, data-driven decision, meaning a CDSS in this area can have a larger influence and acceptance by end users.

However, the degree to which this tool would be accepted and work alongside clinical decision-making behaviour requires investigation.

We have shown that our model is able to reliably estimate alternative patient outcomes depending on their antibiotic treatment status. Based on our results, the size and consistency of the dataset used and, hence, the number of available donors are strongly related to the reliability of outputs. Experiments utilising small datasets often led to poor results given there were not enough suitable patients within a given embedding space to create an appropriate synthetic estimation. On the other hand, there does seem to be a ceiling above which more instances are not necessary. For example, similar results were obtained across the pneumonia, test, and whole datasets even though they had sizes of 2,476, 3,427, and 22,845 patient stays, respectively. As such, we can infer that this method is likely to produce suitable estimations if several thousand patient examples are available. Although this should be reasonable for most clinical scenarios, it does act as a dataset constraint when evaluating less common infections, where potentially more interesting nuanced findings could be made.

The quality of the initial autoencoder model is another significant implication that determines performance. The standard autoencoder model without the synthetic control methods applied achieved higher performance on the LOS prediction task than estimations generated without segregating the embedding space (RMSE of 3.88 and 5.05, respectively). This confirms first that the model has been trained to appropriately represent the patient in the embedding space with respect to their outcome. Second, the temporal aspect of the embeddings assists with synthetic outcome estimations and finally the subsequent synthetic outcome estimation methodology applied ensures that outputs can be clinically applicable with regards to antibiotic treatment. As such, the autoencoder is critical for appropriate temporal representations and subsequent estimations.

It is important to note that there is a high degree of correlation between LOS and overall treatment length in the datasets (**Table 1**, **supplementary Figure S3**). This is to be expected given those patients who are less sick will likely receive fewer antibiotics and leave the ICU sooner. Although the model architecture is designed to account for this, through representative and segregated embeddings, it is still likely that the model "learned" this association causing some confounding. Results on outliers when there is reduced correlation still illustrate that stopping can impact LOS outcomes, even if the predictions themselves are not reliable in this situation given the skewed dataset analysed. Numerous factors influence ICU LOS; hence, even if the model predicts that stopping antibiotics could be neutral or beneficial, other random factors may make this an impossibility. Nevertheless, our results and the strong correlation observed between antibiotic treatment length and LOS in this dataset mean this model can act as a proxy with the ultimate aim of reducing the unnecessary use of antibiotics.

This study has several limitations. We focused on addressing what would happen if antibiotic cessation occurred earlier during a patient's ICU stay. The synthetic control methodology was chosen and adapted as it allows us to address this problem while more traditional causal discovery seems intractable. MAPE and MAE LOS estimation results are in the region of days which could limit clinical utility but are comparable to that of recent research (44). Unlike most synthetic control applications, we do not have an extensive pre-intervention period making confidence in results more challenging. Furthermore, one of our analogues stop "control" days would not be available on a patient-specific level during clinical use due to the nature of cessation occurring after treatment. Other types of interpretability such as being able to investigate selected donors to see if they are clinically meaningful could counteract this. Second, the use of historical EHR data to estimate the synthetic outcome means all our estimations are biased based on past antibiotic prescribing policies. These methodological approaches were required to answer our question of interest but mean that historical approaches towards antimicrobial stewardship govern our model's outputs. The analysis of such a large dataset along with estimations being the weighted average of donors does, however, mitigate this to some extent. In conjunction with this, the analysis presented here is of a macro-scale; however, to realise the potential of this approach for true antimicrobial optimisation, more nuanced, relative, and individualised studies will be required, which we plan to conduct in future. Finally, given the high degree of missingness in the dataset, a number of clinically important features have to be excluded. In particular, research shows that procalcitonin (PCT) and C-reactive protein (CRP) are useful biomarkers for determining when it is safe and appropriate to stop antibiotic therapy (45–48). Neither of these were included as features due to insufficient data. As such, this approach and the subsequent results could potentially be more powerful if applied to a complete dataset focused on a narrow type of infection with defined variables of interest.

In conclusion, we have developed an AI-driven model to estimate patient outcomes if they were to stop or continue antibiotic treatment in the ICU. With further development into a CDSS, we envisage that this can assist clinicians with antimicrobial optimisation and reduce the excessive use of antibiotics to tackle AMR. Future research will investigate which variables promote or hinder cessation and discern the ability of this tool to influence antimicrobial decision-making.

Data availability statement

Publicly available datasets were analysed in this study. These data can be found here: https://physionet.org/content/mimiciv/1.0/.

Author contributions

WB, TR, BH, RW, and DA contributed to study concept and design. WB and BH contributed to data acquisition. WB, BH, and TR contributed to data analysis and accessed and verified the underlying data. WB, TR, and BH contributed to the initial manuscript drafting, discussion of the results, and review of the data. All authors contributed to data interpretation and final revisions of the manuscript. DA, PG, and AH contributed to study supervision. All authors contributed to the article and approved the submitted version.

Funding

WB was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1).

Acknowledgments

The authors would also like to acknowledge (1) the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infection and Antimicrobial Resistance at Imperial College London and (2) The Department for Health and Social Care funded Centre for Antimicrobial Optimisation (CAMO) at Imperial College London. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the UK Department of Health.

Conflict of interest

TR was employed by Sandoz (2020), Roche Diagnostics Ltd (2021), and bioMerieux (2021–2022). These commercial entities were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. All authors declare no other competing interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.997219/full#supplementary-material.

References

1. Nations U. *Political declaration of the high level meeting of the general assembly on antimicrobial resistance: draft resolution/submitted by the president of the general assembly* New York: UN (2016) 6 p.

2. World Health Organization. *Global action plan on antimicrobial resistance*. World Health Organization (2015) 28 p.

3. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. (2022) 399:629–55. doi: 10.1016/S0140-6736(21)02724-0

4. Rawson TM, Moore LSP, Hernandez B, Charani E, Castro-Sanchez E, Herrero P, et al. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? *Clin Microbiol Infect*. (2017) 23:524–32. doi: 10.1016/j.cmi.2017.02.028

5. Hernandez B, Herrero P, Rawson TM, Moore LSP, Evans B, Toumazou C, et al. Supervised learning for infection risk inference using pathology data. *BMC Med Inform Decis Mak*. (2017) 17:168. doi: 10.1186/s12911-017-0550-1

6. Rawson TM, Hernandez B, Moore LSP, Blandy O, Herrero P, Gilchrist M, et al. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. *J Antimicrob Chemother*. (2019) 74:1108–15. doi: 10.1093/jac/dky514

7. Rawson TM, Hernandez B, Wilson RC, Ming D, Herrero P, Ranganathan N, et al. Supervised machine learning to support the diagnosis of bacterial infection in the context of COVID-19. *JAC-Antimicrob Resist*. (2021) 3:dlab002. doi: 10.1093/jacamr/dlab002

8. Hernandez B, Herrero-Viñas P, Rawson TM, Moore LSP, Holmes AH, Georgiou P. Resistance trend estimation using regression analysis to enhance antimicrobial surveillance: a multi-centre study in London 2009–2016. *Antibiotics*. (2021) 10:1267. doi: 10.3390/antibiotics10101267

9. Hernandez B, Herrero P, Rawson T, Moore L, Charani E, Holmes A, et al. Data-driven web-based intelligent decision support system for infection management at point-of-care: case-based reasoning benefits and limitations. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies – HEALTHINF, (BIOSTEC 2017)*. (2017). p. 119–27.

10. Rawson TM, Hernandez B, Moore LSP, Herrero P, Charani E, Ming D, et al. A real-world evaluation of a case-based reasoning algorithm to support antimicrobial prescribing decisions in acute care. *Clin Infect Dis*. (2021) 72:2103–11. doi: 10.1093/cid/ciaa383

11. Tamma PD, Miller MA, Cosgrove SE. Rethinking how antibiotics are prescribed: incorporating the 4 moments of antibiotic decision making into clinical practice. *JAMA*. (2019) 321:139–40. doi: 10.1001/jama.2018.19509

12. Langford BJ, Morris AM. Is it time to stop counselling patients to "finish the course of antibiotics"? *Can Pharm J*. (2017) 150:349–50. doi: 10.1177/1715163517735549

13. Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, et al. Understanding the mechanisms, drivers of antimicrobial resistance. *Lancet*. (2016) 387:176–87. doi: 10.1016/S0140-6736(15)00473-0

14. Spellberg B. The new antibiotic mantra—"shorter is better". *JAMA Intern Med*. (2016) 176:1254–5. doi: 10.1001/jamainternmed.2016.3646

15. Curran J, Lo J, Leung V, Brown K, Schwartz KL, Daneman N, et al. Estimating daily antibiotic harms: an umbrella review with individual study meta-analysis. *Clin Microbiol Infect*. (2022) 28:479–90. doi: 10.1016/j.cmi.2021.10.022

16. Vaughn VM, Flanders SA, Snyder A, Conlon A, Rogers MA, Malani AN, et al. Excess antibiotic treatment duration and adverse events in patients hospitalized with pneumonia. *Ann Intern Med*. (2019) 171:153–63. doi: 10.7326/M18-3640

17. Spellberg B, Rice LB. Duration of antibiotic therapy: shorter is better. *Ann Intern Med*. (2019) 171:210–1. doi: 10.7326/M19-1509

18. Yahav D, Franceschini E, Koppel F, Turjeman A, Babich T, Bitterman R, et al. Seven versus 14 days of antibiotic therapy for uncomplicated gram-negative bacteremia: a noninferiority randomized controlled trial. *Clin Infect Dis*. (2019) 69:1091–8. doi: 10.1093/cid/ciy1054

19. Royer S, DeMerle KM, Dickson RP, Prescott HC. Shorter versus longer courses of antibiotics for infection in hospitalized patients: a systematic review and meta-analysis. *J Hosp Med*. (2018) 13:336–42. doi: 10.12788/jhm.2905

20. Wald-Dickler N, Spellberg B. Short-course antibiotic therapy—replacing Constantine units with "shorter is better". *Clin Infect Dis*. (2019) 69:1476–9. doi: 10.1093/cid/ciy1134

21. Hanretty AM, Gallagher JC. Shortened courses of antibiotics for bacterial infections: a systematic review of randomized controlled trials. *Pharmacotherapy*. (2018) 38:674–87. doi: 10.1002/phar.2118

22. Janssen RME, Oerlemans AJM, Van Der Hoeven JG, Ten Oever J, Schouten JA, Hulscher MEJL. Why we prescribe antibiotics for too long in the hospital setting: a systematic scoping review. *J Antimicrob Chemother*. (2022) 77(8): dkac162. doi: 10.1093/jac/dkac162

23. Charani E, McKee M, Ahmad R, Balasegaram M, Bonaconsa C, Merrett GB, et al. Optimising antimicrobial use in humans: review of current evidence and an interdisciplinary consensus on key priorities for research. *Lancet Reg Health Eur*. (2021) 7:100161. doi: 10.1016/j.lanepe.2021.100161

24. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect*. (2020) 26:584–95. doi: 10.1016/j.cmi.2019.09.009

25. Pandolfo AM, Horne R, Jani Y, Reader TW, Bidad N, Brealey D, et al. Understanding decisions about antibiotic prescribing in ICU: an application of the Necessity Concerns Framework. *BMJ Qual Saf*. (2022) 31:199–210. doi: 10.1136/bmjqs-2020-012479

26. Rawson TM, Charani E, Moore LSP, Hernandez B, Castro-Sánchez E, Herrero P, et al. Mapping the decision pathways of acute infection management in secondary care among UK medical physicians: a qualitative study. *BMC Med*. (2016) 14:208. doi: 10.1186/s12916-016-0751-y

27. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (2021)[Dataset]. doi: 10.13026/s6n6-xd98

28. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*. (2000) 101:e215–20. doi: 10.1161/01.CIR.101.23.e215

29. Qian Z, Zhang Y, Bica I, Wood A, van der Schaar M. SyncTwin: treatment effect estimation with longitudinal outcomes. In: *Advances in Neural Nnformation Processing Systems 34 (NeurIPS 2021)*. Vol. 34. Vancouver Canada: Curran Associates, Inc. (2021). p. 3178–3190.

30. Abadie A, Gardeazabal J. The economic costs of conflict: a case study of the Basque country. *Am Econ Rev*. (2003) 93:113–32. doi: 10.1257/000282803321455188

31. Bouttell J, Craig P, Lewsey J, Robinson M, Popham F. Synthetic control methodology as a tool for evaluating population-level health interventions. *J Epidemiol Community Health*. (2018) 72:673–8. doi: 10.1136/jech-2017-210106

32. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ*. (2016) 25:1514–28. doi: 10.1002/hec.3258

33. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Vol. 32. Vancouver Canada: Curran Associates, Inc. (2019). p. 8024–8035.

34. Kingma DP, Ba J. *Adam: a method for stochastic optimization* (2014). Available from: https://arxiv.org/abs/1412.6980

35. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training [Preprint] (2018). Available at: *http://arxiv.org/1807.05118*.

36. Dimopoulos G, Poulakou G, Pneumatikos IA, Armaganidis A, Kollef MH, Matthaiou DK. Short- vs long-duration antibiotic regimens for ventilator-associated pneumonia: a systematic review and meta-analysis. *Chest*. (2013) 144:1759–67. doi: 10.1378/chest.13-0076

37. Pugh R, Grant C, Cooke RPD, Dempsey G. Short-course versus prolonged-course antibiotic therapy for hospital-acquired pneumonia in critically ill adults. *Cochrane Database Syst Rev*. (2015) (8):CD007577. doi: 10.1002/14651858.CD007577.pub3

38. Drekonja DM, Trautner B, Amundson C, Kuskowski M, Johnson JR. Effect of 7 vs 14 days of antibiotic therapy on resolution of symptoms among afebrile men with urinary tract infection: a randomized clinical trial. *JAMA*. (2021) 326:324–31. doi: 10.1001/jama.2021.9899

39. de Gier R, Karperien A, Bouter K, Zwinkels M, Verhoef J, Knol W, et al. A sequential study of intravenous and oral fleroxacin for 7 or 14 days in the treatment of complicated urinary tract infections. *Int J Antimicrob Agents*. (1995) 6:27–30. doi: 10.1016/0924-8579(95)00011-V

40. Peterson J, Kaul S, Khashab M, Fisher AC, Kahn JB. A double-blind, randomized comparison of levofloxacin 750 mg once-daily for five days with ciprofloxacin 400/500 mg twice-daily for 10 days for the treatment of complicated urinary tract infections and acute pyelonephritis. *Urology*. (2008) 71:17–22. doi: 10.1016/j.urology.2007.09.002

41. National Institute for Health and Care Excellence. *Pneumonia (hospital-acquired): antimicrobial prescribing NICE guideline [NG139]*. (2019). Available from: https://www.nice.org.uk/guidance/ng139

42. National Institute for Health and Care Excellence. *Urinary tract infection (lower): antimicrobial prescribing NICE guideline [NG109]*. (2018). Available from: https://www.nice.org.uk/guidance/ng109

43. Pauwels I, Versporten A, Vermeulen H, Vlieghe E, Goossens H. Assessing the impact of the Global Point Prevalence Survey of Antimicrobial Consumption and Resistance (Global-PPS) on hospital antimicrobial stewardship programmes: results of a worldwide survey. *Antimicrob Resist Infect Control*. (2021) 10:138. doi: 10.1186/s13756-021-01010-w

44. Rocheteau E, Liò P, Hyland S. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. *Proceedings of the Conference on Health, Inference, and Learning, CHIL '21*. 2021 April 8 – 10; New York, NY: Association for Computing Machinery (2021). p. 58–68. Available at: https://doi.org/10.1145/3450439.3451860

45. Schuetz P, Wirz Y, Sager R, Christ-Crain M, Stolz D, Tamm M, et al. Procalcitonin to initiate or discontinue antibiotics in acute respiratory tract infections. *Cochrane Database Syst Rev*. (2017) 2017:CD007498. doi: 10.1002/14651858.CD007498.pub3

46. Rhee C. Using procalcitonin to guide antibiotic therapy. *Open Forum Infect Dis*. (2016) 4:ofw249. doi: 10.1093/ofid/ofw249

47. Oliveira CF, Botoni FA, Oliveira CRA, Silva CB, Pereira HA, Serufo JC, et al. Procalcitonin versus C-reactive protein for guiding antibiotic therapy in sepsis: a randomized trial. *Crit Care Med*. (2013) 41:2336–43. doi: 10.1097/CCM.0b013e31828e969f

48. Coelho L, Póvoa P, Almeida E, Fernandes A, Mealha R, Moreira P, et al. Usefulness of C-reactive protein in monitoring the severe community-acquired pneumonia clinical course. *Crit Care*. (2007) 11:R92. doi: 10.1186/cc6105

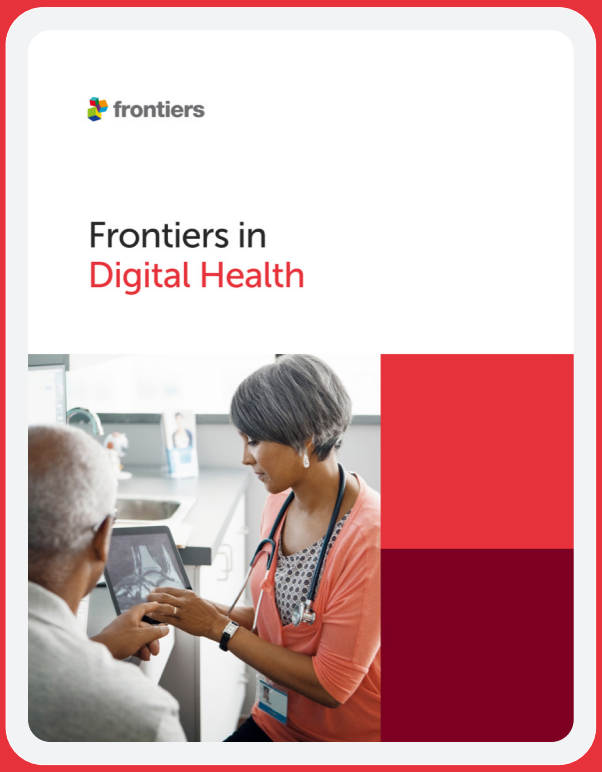
Frontiers in
Digital Health

Explores digital innovation to transform modern healthcare

A multidisciplinary journal that focuses on how we can transform healthcare with innovative digital tools. It provides a forum for an era of health service marked by increased prediction and prevention.

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