frontiers | Frontiers in Digital Health

REVIEW
published: 03 May 2022
doi: 10.3389/fdgth.2022.765406

12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context

Stephane Doyen[1]* *and Nicholas B. Dadario*[2]

[1] Omniscient Neurotechnology, Sydney, NSW, Australia, [2] Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, United States

The healthcare field has long been promised a number of exciting and powerful applications of Artificial Intelligence (AI) to improve the quality and delivery of health care services. AI techniques, such as machine learning (ML), have proven the ability to model enormous amounts of complex data and biological phenomena in ways only imaginable with human abilities alone. As such, medical professionals, data scientists, and Big Tech companies alike have all invested substantial time, effort, and funding into these technologies with hopes that AI systems will provide rigorous and systematic interpretations of large amounts of data that can be leveraged to augment clinical judgments in real time. However, despite not being newly introduced, AI-based medical devices have more than often been limited in their true clinical impact that was originally promised or that which is likely capable, such as during the current COVID-19 pandemic. There are several common pitfalls for these technologies that if not prospectively managed or adjusted in real-time, will continue to hinder their performance in high stakes environments outside of the lab in which they were created. To address these concerns, we outline and discuss many of the problems that future developers will likely face that contribute to these failures. Specifically, we examine the field under four lenses: approach, data, method and operation. If we continue to prospectively address and manage these concerns with reliable solutions and appropriate system processes in place, then we as a field may further optimize the clinical applicability and adoption of medical based AI technology moving forward.

Keywords: artificial intelligence, machine learning, deep learning, medical software, cloud computing, neural network, medicine

OPEN ACCESS

Edited by:
Enrique Castro-Sánchez,
University of West London,
United Kingdom

Reviewed by:
Niamh Lennox-Chhugani,
International Foundation for Integrated Care (IFIC), United Kingdom
Jawahar Kalra,
University of Saskatchewan, Canada

****Correspondence:***
Stephane Doyen
stephane.doyen@o8t.com

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

Received: *27 August 2021*
Accepted: *11 April 2022*
Published: *03 May 2022*

Citation:
Doyen S and Dadario NB (2022) 12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context.
Front. Digit. Health 4:765406.
doi: 10.3389/fdgth.2022.765406

INTRODUCTION

The powerful applications of artificial intelligence (AI) have long been promised to revolutionize the healthcare field. AI has been met with a surge of interest in the scientific and medical communities due to the increasing number of patients receiving healthcare services and the concomitant increases in complexity of data, which is now available, but often uninterpretable by humans alone. These technologies demonstrate the ability to identify malignant tumor cells on imaging during brain surgery (1), unravel novel diseases into explainable mechanisms of viral

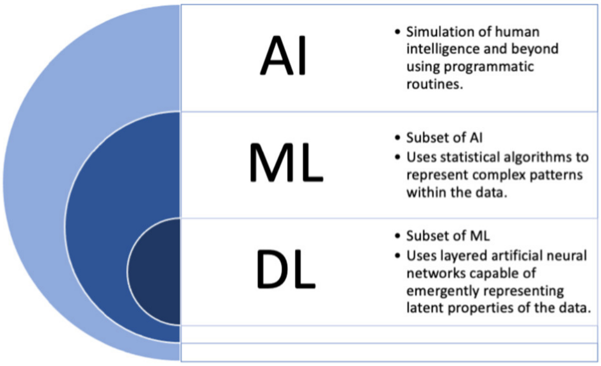

AI
• Simulation of human intelligence and beyond using programmatic routines.
ML
• Subset of AI
• Uses statistical algorithms to represent complex patterns within the data.
DL
• Subset of ML
• Uses layered artificial neural networks capable of emergently representing latent properties of the data.

FIGURE 1 | The umbrella of "AI". AI, artificial intelligence; DL, deep learning; ML, machine learning.

mutations for therapeutic design (2), predict the progression of neurodegenerative diseases to begin earlier treatments (3), and assist with the interpretation of vast amounts of genomic data to identify novel sequence patterns (4), among a number of many other medical applications. Ultimately, the applications of AI in medicine can generally be grouped into two bold promises for healthcare providers: (1) the ability to present larger amounts of interpretable information to augment clinical judgements while also (2) providing a more systematic view over data that decreases our biases. In fact, one could argue that improving our ability to make the correct diagnoses if given an opportunity can be seen as a key duty and moral of the medical field in principle (5). Given this enormous potential, it is unsurprising that Big Tech companies have matched the enthusiasm of scientific experts for AI-based medical devices by investing substantial efforts and funding in their product development over recent years (6). Unfortunately, despite these exhilarating observations and promises, we have yet to truly harness the full potential of AI as a tool in current practices of medicine.

Many loosely utilize the label "AI" for the medical applications described above, but providing more focused definitions for a few important terms may be beneficial moving forward in this work. The term "AI" in fact more broadly refers to the idea of machines being able to intelligently execute tasks in a manner similar to human thinking and behavior. To date, such applications are not currently found in the field. Instead, applications where machines learn from data can more accurately be labeled as machine learning (ML) based solutions. For simplicity, we will use these two terms interchangeably throughout the current manuscript, and outline commonly used terms in **Figure 1** which are considered under the umbrella of "AI."

While not newly described, ML based applications have resurfaced interest in the medical community with the rise of the recent SARS-CoV-2 pandemic (2). International collaborative efforts from Data Scientists have attempted to take advantage of differences in disease prevalence across the world as a way of utilizing early access to data to improve the quick diagnosis and prognostication of patient outcomes in soon-to-be overwhelmed hospitals. In fact, novel approaches had immediately begun to tackle concerns on vaccine developments early in the pandemic according to possible viral mutations that allow escape from the human immune system. By framing SARS-CoV-2 protein sequence data in the context of linguistic rules used in the human natural language space, ML algorithms may be able to present to us *interpretable* mechanisms of how a virus mutates while retaining its infectivity, similar to how a word change in a human language may dramatically alter the meaning of a sentence without changing its grammar (2). Indeed, thousands of exciting models with increasing collaboration and data sharing had been immediately proposed throughout the pandemic to improve the prognostication and clinical management of COVID-19 patients (7, 8). While few clinical applications were found to demonstrate a true clinical impact on patient outcomes for the current pandemic, it is important to note that ML applications have elsewhere demonstrated important clinical applications in similar contexts (9), such as for improving the allocation of limited resources (10), understanding the probability of disease outbreak (11), and the prediction of hospital stay and in-hospital mortality (12). Therefore, while these models continue to demonstrate enormous potential to manage large scale clinical scenarios, further work is still necessary to ensure they can be quickly and effectively leveraged for clinical translation, a concern which has been encountered in the field previously.

Leading digital companies have often pioneered the implementation and excitement of AI technologies in current medical practices. IBM Watson for Oncology took years of advanced development and training with physicians from Memorial Sloan Kettering, ultimately teaming up with the MD Anderson Cancer Center in 2013, another leading cancer center in the United States, with promises of improving oncological diagnoses and therapy decision making. However, this collaboration was terminated by MD Anderson in 2017 due to failure of meeting their oncological and patient goals with this novel software (13, 14). Many believe increasing technical advancements in computational abilities will reinvigorate the potential horizons and trust of these technologies in the medical field moving forward. A new deep learning system made by Google promises to detect 26 skin conditions with accuracy comparable to US board-certified dermatologists (15), yet recent work has already come under substantial controversy due to underperformance based on underlying inter-individual differences in demographics, such as gender and skin color (16). Nonetheless, recent public data continues to demonstrate increased interest in medical AI software as seen with continued surges in AI investments in drug design and discovery, attention in large governmental plans, and focuses on AI in scientific careers for medical applications compared to each previous year (17).

The limitations mentioned above are not to say that we should abandon the field of AI-based medical technologies as a whole. The ability for human intuition alone to map the brain for instance, would take insurmountable amounts of time and effort compared to that which is now possible with AI-based technology (18–20). As original goals previously hoped that these technologies would resemble human based intelligence, it is important to remember *To Err is Human*. Instead, as with all emerging fields, further work is necessary to optimize the clinical applicability of these tools as we move forward to minimize

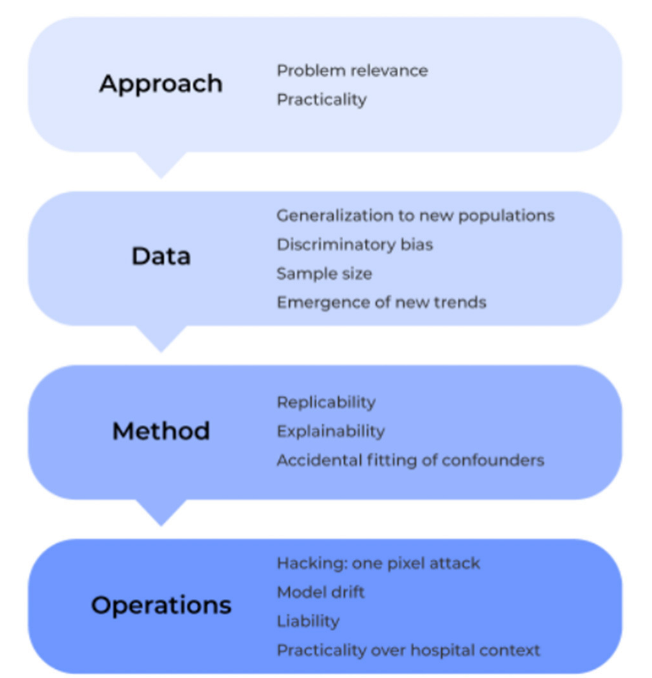


FIGURE 2 | The 12 plagues of AI in healthcare. The development of AI software and deployment into the medical community can be generally grouped into four processes: approach, data, method, and operation. In each component, a number of common problems can be encountered which must be prospectively considered or managed in real-time to ensure efficient and accurate model performance in high-risk medical environments. Here, we show the most common pitfalls of AI-based medical solutions within these components and discuss them throughout the current manuscript.

unnecessary errors and harm. Unfortunately, there is limited information available in the literature that presents a clear guide of the common problems that will be encountered with AI-based software, and more specifically how to overcome them moving forward in medical practices. Such a gap possibly reflects the separation of expertise and focus between medical professionals and data scientists all together.

To address this gap, we provide a clear guide below on the most common pitfalls to medical AI solutions based on the previous literature. Specifically, we examine the field under four lenses: approach, data, method and operation (**Figure 2**). Within these core components, we outline a number of different issues that must be considered in the development and application of machine learning technologies for medical based applications, and we elucidate how to identify, prevent, and/or solve them moving forward to ultimately create safer, more trustworthy, and stronger performing models.

APPROACH: WHAT IS THE PLAN?

Problem 1. Relevance: Cutting Cake With a Laser

Due to a rise in the amount of open-source and advanced statistical software, it has become increasingly easy to develop highly elegant and powered computational models. However, when created with only technical solutions in mind, models can easily be created to solve an non-existent or irrelevant problem. In turn, the ultimate users of the technology, such as the practicing physician in the clinic, will have no interest in the solution being answered by a specific model. Problems with a model's relevance can render even the most elegant application of data science irrelevant (**Figure 3**).

A common example of this can be seen with the increased ability we now have to detect mental illness based on improved and publicly available maps of the brain connectome (21, 22). If a ML specialist received a dataset including data on patients with Schizophrenia, their first thought may be to build a model to detect Schizophrenia. However, current medical practices are already highly capable of such detection and would rather see other issues provide more fruitful avenues for ML applications, such as predicting modulatory treatment responses for Schizophrenia (23). Therefore, viewing such issues from a pure computer science or statistical standpoint is inherently hindering the potential for a current project.

Instead, the ultimate users of the technology should be included at the very beginning of the development of the model. Research and Development goals should be grounded in what has already been suggested in the medical field as important avenues of future work for clinical improvements. For instance, rather than attempting to predict an illness which can already be clearly identified in clinical practice, ML tools can reduce the complexity of patient information presented in a specific pathological states (24, 25) and then present statistical irregularities that can be used by physicians to make more informed decisions for clinical treatment (26, 27). Ultimately, it must be remembered that AI is a powerful *tool* that can be leveraged to answer a difficult questions, the usefulness of the tool is a function of the appropriateness of the question being asked.

Problem 2. Practicality: Not Everyone Has a Cyclotron

Similar to the concerns of *relevance* mentioned above, advances in ML abilities have also created concerns of *practicality*. This refers to building a model which has limited practical applications in the environment of interest due to logistical constraints that were not considered outside of the environment that the model was originally created in, such as requiring more computation than necessary or that which can be feasibly run in a clinic. Reasonable solutions must commonly consider the current state of the field and the technical constraints of the model proposed.

To create and train advanced ML models, disparate data is often harmonized from varying sources and formats to create a single large dataset of valuable information. Data harmonization is particularly important in medicine given that small amounts of data on specific topics are often managed by acquiring data from varying records and from a variety of centers, all of which also commonly utilize different electronic health record (EHR) systems (28). Therefore, to build a model that aims to compare specific patients to a template/control group, algorithms must first harmonize large amounts of a patient's data from varying

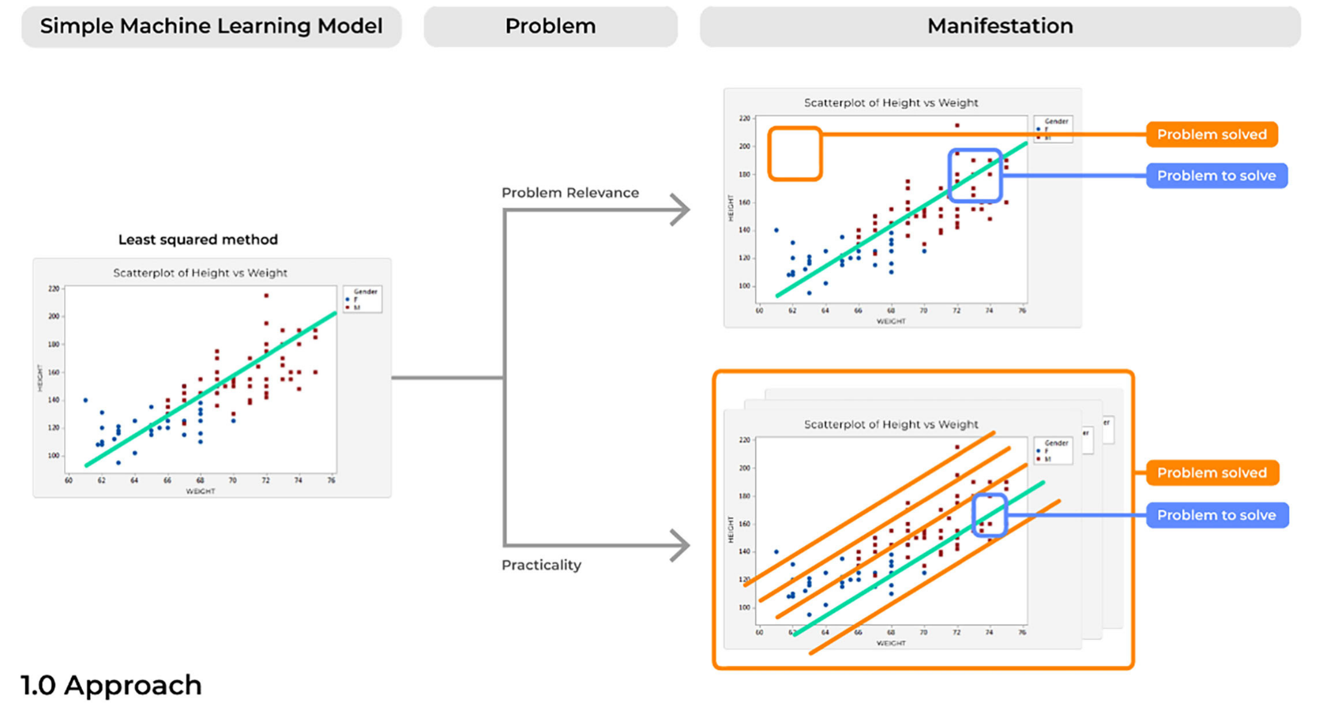


FIGURE 3 | Concerns with the approach to ML product development.

datasets to perform a comprehensive comparison. However, to do this, many algorithms require datasets to be harmonized as *large cohorts*. Unfortunately, outside of the lab, these methods can become impractical when single ($N = 1$) patients present to the clinic. Therefore, these algorithms can fail unless additional practical solutions are present. Importantly, improved super computers may hold the computing power to execute a highly complex model in a lab for a single patient, but a physician may not be able to analyze a patient's data with these algorithms on a less powerful hospital-issued laptop in a time critical environment where it must perform at the highest level. On that same note, it is easy to optimistically consider the intricate problems that can be tackled with ML in the future due to the improved abilities of modern computational algorithms to digest highly complex data; however, medical data is often not available or too limited for specific topics at the current time. When just utilizing the limited available data obtained specifically from academic studies to train a model, illusory results may be obtained given these data are cleaner than that which would be obtained from the actual target field. Therefore, investing time in an approach that is too far into the future may inherently cause difficulties in model preparation due to the lack of available feedback from the clinical field of interest.

To prospectively mitigate problems of *practicality*, implementation should always be considered from the very start of any analytical solution. This will prevent any waste of Research and Development efforts given that alternative solutions were already considered. A notable increase in recent efforts has been put forth by cloud providers and hardware manufactures to provide development frameworks which bridge the gap between an approach to a problem and the hardware to support it (29). Ultimately, including the intended users in the initial development stages will also provide insight on *practicality* during model development as the goal environment will have already been considered.

DATA: WHAT ARE WE USING AND FOR WHAT PURPOSE?

Problem 3. Sample Size: Looking at the World Through a Keyhole

A concern inherent to most analytical solutions includes issues of sample size. When creating and training a model using a limited sample size, inflated results may be demonstrated when actually testing it against a control sample. Subsequently, when introduced into a different, clinical environment, the model's accuracy can lessen, and it can fail. This is because the small sample size may cause a model to present results that are merely due to chance. Consideration of this is paramount to acquire reliable results (**Figure 4**).

In an age of increasingly available data, ML algorithms must be trained on adequately sized and diverse datasets. While this will likely require the harmonization of data from multiple centers, this point is imperative if we are to believe our models are robust enough to consistently provide reliable results at different time points and in a number of different environments. A recent

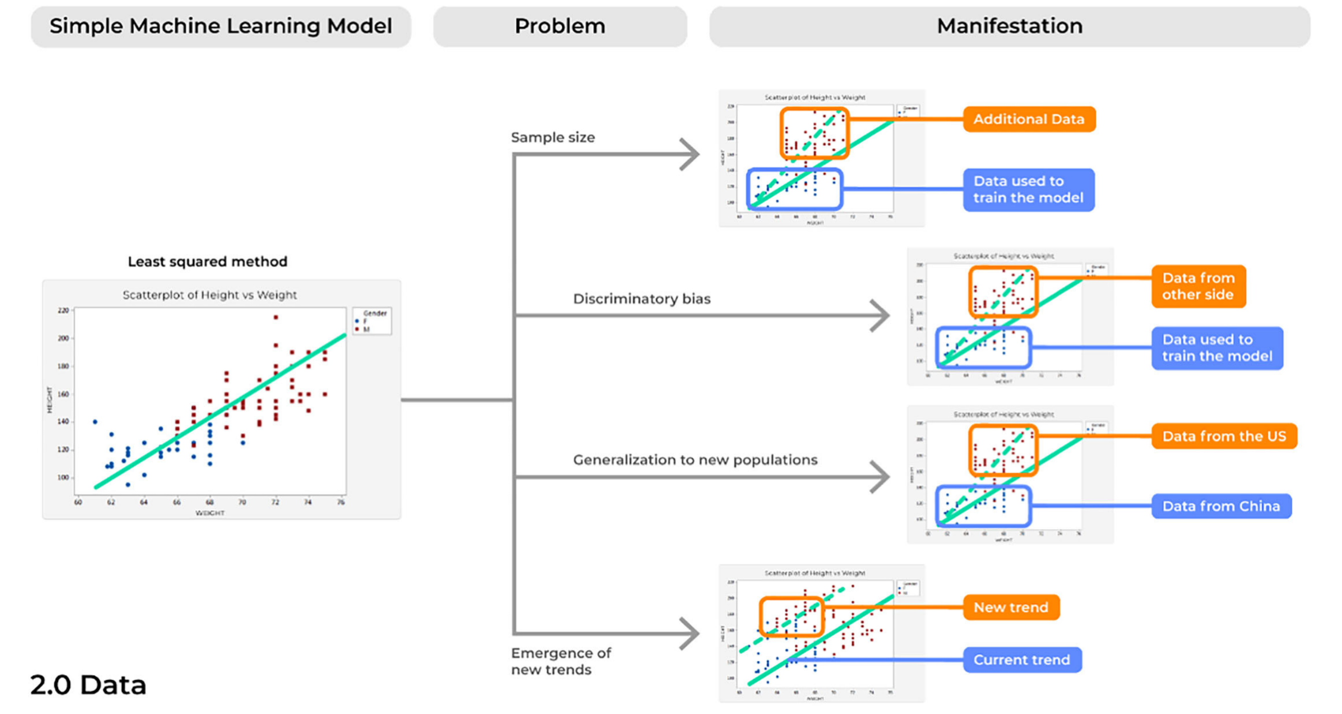


FIGURE 4 | Concerns with the data a model is being trained and tested on.

systematic review assessed 62 comprehensive studies detailing new ML models for the diagnosis or prognosis of COVID-19 utilizing chest x-rays and/or computed tomography images and found a common problem among studies was in the limited sample size utilized to train models (8), with other reviews noting similar observations (30). Amongst the 62 models assessed, more than 50% of diagnostic focused studies (19/32) utilized <2,000 data points. Unlike rare brain tumors which may require years of small amounts of data collection, there have been more than 2 million cases of COVID-19 to date (31). Therefore, it is unsurprising that developing a model to understand a complex biological phenomenon on only 2,000 data points can lead to unreliable results in the intended environment (8).

As in all aspects of medical research, newer ML models must be trained on a large and diverse dataset to provide reliable results. A clear solution is to implement an active and ongoing effort to acquire more data from a variety of sources. ML models, unlike physical medical devices, can be continually updated and improved to provide more reliable and accurate performances. As more data becomes available, models can be retrained with larger sample sizes and then updated for the best performance in the field. Ultimately, this may require a model to be pulled from a clinical practice, retrained and tested based on more data which is now available, and then placed back into the environment of interest. However, with the increased demand for more data, one must consider the possibility that an ML modeler may acquire and include data for which they have no patient consent. Once the new data is incorporated into the model and the model is input into the field, it is very difficult to identify which data was utilized. A possible solution to these patient consent concerns with increasing data is implementing in each image a certificate stating patient consent, such as through non-fungible tokens (NFTs) (32).

Problem 4. Discriminatory Biases: Rubbish in, Rubbish Out

Perhaps the most commonly discussed problem concerning AI technologies is their potential to exacerbate our current societal biases in clinical practice. Even more alarmingly, developing models with historical data may perpetuate our historical biases on a variety of different patients in ways we have since improved from.

Issues concerning discriminatory bias may manifest as a model demonstrating high performance on a single sample of patient data, but then failing on different subsets of individuals. For instance, an increasing focus in ML based applications has been to create algorithms capable of assisting dermatologists in the diagnosis and treatment of diseases of the skin (33). While racial inequalities to healthcare delivery are becoming more and more documented across the world, recent work has suggested one of the major sources of these inequalities stems from the lack of representation of race and skin tone in medical textbooks (34). Medical schools in recent years have immediately begun to address these concerns by updating textbook images for increased inclusivity, yet deep learning (DL) image-based classifier algorithms often continue to use low quality datasets for training, which commonly contain unidimensional data (e.g., mostly lighter skin images) (35).

In turn, these algorithms will perform better on the image type it is trained on, and subsequently propagate biases that were represented in the original datasets. This cycle ultimately provides the chance for profound failures with specific groups of peoples. In a study examining three commercially available facial recognition technologies (Microsoft, Face++, and IBM) based on intersectional analyses of gender and race, differences in error rates of gender classification were demonstrated to be as high as 34.7% in darker-skinned females compared to 0.8% in lighter-skinned males (36). Ultimately, even if some diseases are more common in specific races or genders, such as melanoma in non-Hispanic white persons, all patients with a variety of skin types should be included for the potential benefits of these algorithms in the future (35). While these concerns may be more obvious, even the location of the image acquisition center can bias a model's performance. This is due to the demographical makeup of the surrounding community in which images were trained and tested. Thus, when the model performs in a different environment, its performance will drastically differ based on the new demographics encountered.

Improved patient related factors must be considered when building datasets for the training of a model. By building models on data from a variety of different sites with increased awareness of the specific populations included, we may begin to mitigate the potential biases in our results. Ultimately, improvements in DL algorithms remain relatively nascent, and there has been an increased focus on classification performance across gender and race, providing us with impetus to ensure that DL algorithms can be successfully used to mitigate health care disparities based on demographics (35–37).

Problem 5. Generalization to New Populations: From the Few, but Not the Many

Expanding on concerns of discriminatory biases, problems with generalization may occur due to the expansion of global software markets. Aside from differences in gender and skin color alone, a model may fail based on datasets trained on individuals from a single population due to underestimating the differences in population driven variability.

Most consider that using multi-centric data improves the external validity of a model's results given that it tests samples of individuals from a variety of locations. However, if all the centers providing data are within a single country, such as in the United States for instance, how will these models perform in China where there are unique individual characteristics that are concomitantly shaped by differences in the environment? For example, one of the most accepted neurobiological models of language suggests a dominant left-lateralized system. However, many of these models were based on participants that speak English or are from the United States (38), while other studies including Chinese participants suggest a right-lateralized white matter system related to learning Mandarin (39). In fact, other studies have suggested these results also expand to non-Chinese subjects who learned Mandarin as a non-native language, such as European subjects (40), suggesting that differences in white matter connectivity may be more pronounced for some tonal languages. However, it is also reasonable to conjecture based on the effects of these subtle differences that there are likely additional underlying inter-individual differences not being considered in this paradigm outside of just tonal languages alone (41). Without consideration of differences across separate datasets (42), unexpected performances in ML-based brain mapping software could jeopardize market expansion into different areas outside of where the model was originally developed.

To prevent and manage issues of generalization, a number of solutions exists. First, similar to what was described above, data must be accumulated from a variety of sources. However, to provide the most generalizable results, these sources must span several different sites inside and outside of the country of origin where the model was developed. Surely, inter-individual factors must be considered during production to improve the robust ability of a model in different environments. Nonetheless, it is also likely that site-specific training should also be considered as an optimal avenue to tailor models based on the specific populations where a model is going to be implemented. Then, external validation testing in separate adequately sized datasets (43) can ensure that an algorithm can model data from different sites similarly to that which it trained on.

Importantly, improved collection of multi-site data simultaneously raises concerns of patient anonymity, patient agency and informed consent. Fortunately, a great deal of progress has been demonstrated with methods of federated learning to deal with the bias of models when trained with homogenous populations (44). Federated learning methods improves the maintainability of data anonymity when sharing patient data across numerous sites, thus allowing for improved research collaboration and model performance across heterogenous populations (44). However, given the ability for various ML systems to re-identify individuals from large datasets, a key improvement in the future suggested by Murdoch (45) will likely also include recurrent electronic informed consent procedures for new uses of data and further emphasis on the respect for the ability of patients to withdraw their data at any time.

Problem 6. Emergence of New Trends: Surfacing Creatures From the Depth

This problem is likely the most relevant to the current conditions of the world with the recent SARS-CoV-2 pandemic. As such, problems related to the *emergence of new trends* refers to when a new trend emerges in the data that the initial model was not built to account for, thus altering the new statistical comparisons being made between variables.

Previously, ML techniques have been commonly applied to predict changes in seasonal diseases, such as influenza (46), to further allow hospitals to appropriately prepare for medical supply needs, such as bed capacity, and to appropriately update both vaccine developments and citizens themselves of prevalent circulating strains. This is because many viruses commonly mutate and produce a variety of strains each year, yet vaccines can only account for a number of the most prevalent strains. In such paradigm, ML tools can be applied to estimate which strains will be most common in upcoming seasons with high accuracy to be included in upcoming seasonal vaccines (46). However, unexpected changes can occur in the environment, such as a new pandemic, which drastically alters the environmental landscape and therefore changes the way two variables may be modeled based on new environmental parameters. If there is not an ongoing monitoring system in place, these models can lead to potential harm as results are no longer reliable. Similarly, medical devices are constantly being altered and upgraded to improve their diagnostic and visualization abilities, such as for functional magnetic resonance imaging (fMRI) scanners. However, magnetic field inhomogeneity between different scanners, such as a 3 Tesla vs. a newer 7 Tesla, could lead to differences in relative blood oxygen level-dependent (BOLD) signal intensity, and therefore contains poor inter-scanner reliability (47). As such, when utilizing brain mapping software on an individual patient with different scans, erroneous brain network anomalies may arise and can lead to inappropriate neurosurgical treatments just merely due to the inability of a model to account for differences in functional magnetic resonance imaging (fMRI) scanners utilized.

Models in production should be created with a set of test reflective environmental data to ensure expected performances *in situ*. Furthermore, alongside changes in current clinical practices, models must be continually monitored and tested with new data to assess for reliability and validity. This continual external validation testing with separate adequately sized datasets than which it was trained on provides a necessary avenue for improvement as the field of healthcare and the environment itself is continually changing.

METHOD: HOW DOES THE TECH APPROACH PLAY OUT?

Problem 7. Reproducibility: Bad Copies

Concerns of replicability are not a newly discussed phenomena for a number of different fields that require the processing of large amounts of data (48). However, failure of a model to demonstrate the same results time and time again presents a profound risk of inconsistencies in the delivery and quality patient care, and therefore should be considered in future ML developments (**Figure 5**).

To ensure an ML algorithm applied in the healthcare setting is fully reproducible, some (49) have suggested that a study should produce the same results according to (1) technically identical conditions (related to code and dataset release), (2) statistically identical conditions (related to differences in sampled conditions still yielding the same statistical relationships), and (3) conceptually identical conditions (related to how the results are reproduced in accordance with pre-defined descriptions of the model's effects). When these methods of reproducibility are not met, the one who created the model would be unable to replicate its results on subsequent runs. Furthermore, when others are attempting to assess the model, possibly to improve its applicability, they too will be unable to obtain the reported effects by the original authors. As such, recent methodological and reporting standards have been proposed to address these issues, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis OR Diagnosis (TRIOPD), and its recent statement for ML-prediction algorithms (TRIPOD-ML) (50, 51).

In addition to the obvious potential improvements in patient safety following more rigorous evaluations of clearly reported methodology, improved reporting of ML algorithms can also provide an important way to advance the field as a whole. A number of different Data Scientists may spend countless hours in designing complex ML models to address an imminent question, such as what we saw for COVID-19, yet increasing errors will commonly be identified from just the smallest differences across algorithms (8). Instead, if models and datasets are clearly reported following a study, then others can appropriately assess these models and collectively improve upon them to produce more robust pipelines. This will ultimately improve our ability to bring these tools to clinical practice as a model becomes more accurate without repeating the same mistakes. The increased requirements for adherence to rigorous, ML reporting guidelines across many major peer-reviewed journals is a promising improvement moving forward.

Problem 8. Explainability: The Black Box Problem

One of the largest concerns of AI-based devices in medicine concerns physicians' lack of trust for model performance (52). Unfortunately, as ML models have increased in complexity, this improvement has often been met with a trade-off in explainability, in which there is increasing uncertainty regarding the way these models actually operate (53).

This problem is often described as a model operating in a "black box," in which irrespective of model performance, very little can be elucidated about why a model made a specific decision. A common example of this can be seen with a highly powered ML technique known as deep learning (DL). DL applications can maintain hundreds of stacked representations across hundreds of layers, a relationship that no human can truly accurately comprehend in full detail. However, a number of important improvements can be made in the field as we improve concerns of lack of explainability, to which a whole field has been dedicated known as Explainable Artificial Intelligence (XAI) (53). Ultimately, ML tools are capable of taking highly dimensional data and quickly making accurate decisions in highly time-critical medical scenarios, a feat that humans may never physically nor cognitively be capable of performing (54). However, if we could explain the various decisions being executed by a certain model and the specific features being analyzed to produce a certain outcome (24), physicians can better interpret these results based on logic and previous knowledge. Then, healthcare providers may not only be able to better trust these algorithms, but providers may also continually improve the model's performance when the system presents an error that is likely based on a specific wrong answer possibly being executed

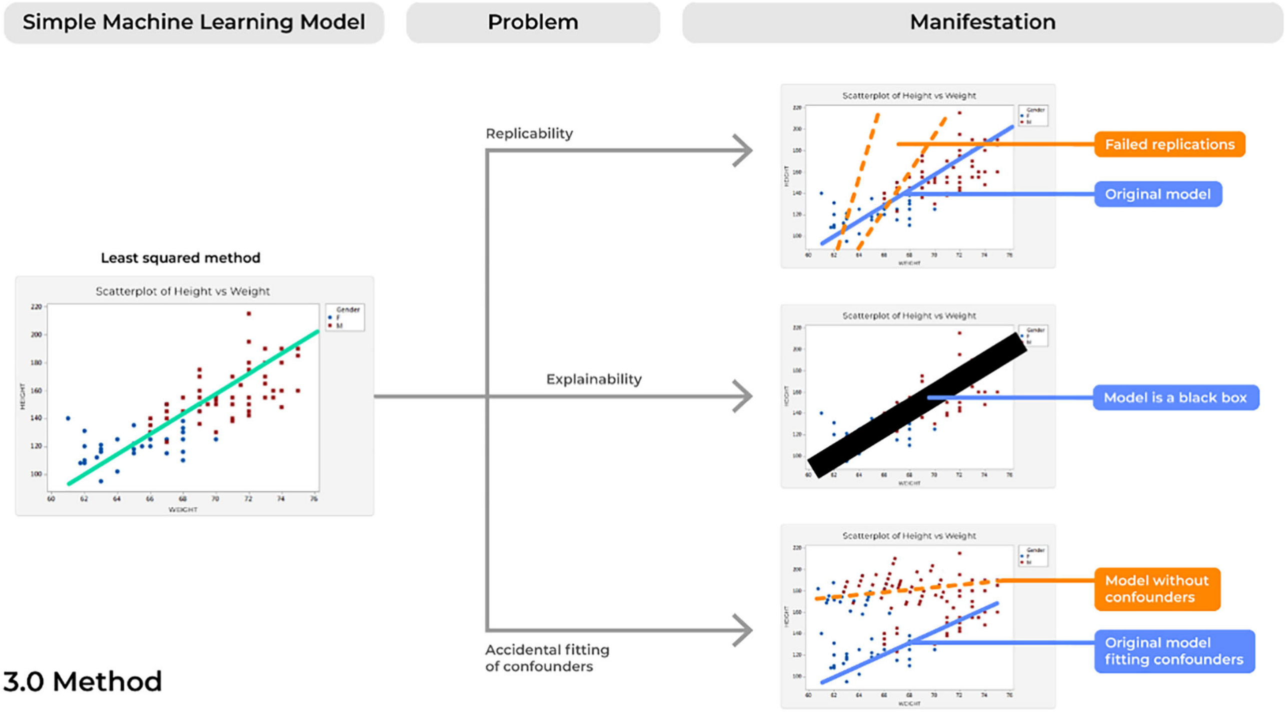


FIGURE 5 | Concerns of the methodology utilized to develop an ML model.

in a portion of a decision tree. In fact, since these models are highly capable of detecting novel patterns in large amounts of data which are invisible to the human eye (4), interpretable and explainable models may also unlock new insights in the scientific world that spur further improved ML developments in the future, creating a positive reinforcing cycle of innovation.

Outside of the trust of a practicing healthcare provider, the patient themself, if diagnosed by a ML tool to have a malignant skin lesion, may too require an interpretable and justifiable reason why specific results were provided, such as *why* the tumor was diagnosed as malignant. Thus, it is important that a clinician is able to interpret the decisions made by a specific algorithm, but this also raises concerns of violating patient-physician relationships and liability for AI technology in general (55). A core component of the Hippocratic Oath requires physicians to do no harm and act upon their greatest judgement for improved patient care. With the incorporation of machine-learning guided clinical care, failure to understand a model's decision making can shift fiduciary duties away from the physician and hurt the patient-physician alliance (56). Furthermore, if a model provides a piece of information that leads to a poor outcome for a patient, is it the *machine's* fault or is it the *healthcare provider's* medical error? Unsurprisingly, promotion of interpretability of a model is outlined as a main principle within the recent World Health Organization (WHO) guidelines on *Ethics & Governance of Artificial Intelligence for Health* (57). Both the model and provider must be able to clearly elucidate these findings to the patient if we are to truly incorporate ML into standard medical practices.

Movement toward white-box, also called glass-box, models provides a solution to address concerns of explainability. These models can often be seen with linear (58) and decision-tree based models (24), although a number of other applications are increasingly being developed (53). In fact, DL based networks make up the majority of the highly sought after radiological-AI applications for the medical field (1), such as the systems that can diagnose brain cancer during surgery. Such networks provide the enthusiasm for the recent large scale efforts in the field to improve the explainability of advanced ML techniques (59). Specifically, by utilizing white box models as first line modeling techniques, one can ensure findings are being appropriately made based on ground truths provided by current scientific knowledge. For example, a number of recently developed practical approaches have been introduced using input vector permutation to better understand how specific inputs impact the predictions of a model and may be particularly useful to gain insight into how models make specific decisions (60, 61). Explainable AI approaches, such as deconvolution methodology, can be applied to more complicated models, such as convolutional neural networks (CNNs) and ensembles, to improve the interpretability of the more complex models (62). However, further research is needed in the field of explainable AI to better understand model-specific techniques that can be leveraged to ultimately improve the transparency of these models in the healthcare setting.

Problem 9. Accidental Fitting of Confounders: Guilt by Association

ML tools are able to digest highly complex datasets by continually assessing and scanning different features until optimal performance is achieved. As such, concerns of accidentally fitting confounders can easily surface and a model that was thought to be capable of predicting an outcome is instead making a prediction based on factors unrelated to that outcome of interest. If so, these models can produce not only unreliable results in clinical practice, but can also present profound risks of patient harm, such as by under- or over-estimating specific diagnoses.

An example of this problem can be seen with a model that is purported to show great performance in detecting autism. However, if not carefully assessed for confounders, one may miss that the model is actually detecting head motion. For instance, patients with autism often move more in fMRI scans and this can cause head motion artifacts that compromise fMRI data due to altering voxel and stable state magnetization. Ultimately, this will cause scans to show false regions of increased/decreased brain activity that are misused to diagnose autism (63). If head motion is not corrected for, the performance of these models will collapse (64). Unfortunately, these children may have already received unnecessary treatments (65) that resulted in increased financial burden (66) and possibly decreased treatments for other diagnoses (67). Alarmingly, the literature presents a number of additional examples of this problem that have may have gone unnoticed in certain ML algorithms.

First, an ML specialist must have a strong understanding of the data being modeled. Then, when actually developing the model, one should carefully explore and rule out any concerns for confounders. In addition to previous descriptions of "white-box" models, improved understanding of the features being mapped may allow further appropriate critical evaluations of model performances and in turn lead to increased trust in the medical community.

OPERATION: IN THE FIELD

Problem 10. Model Drift: Like a Rolling Stone

For many of the reasons discussed above, over time a model will likely begin to make an accumulating number of errors. This could be due to issues with *model drift*, in which a model that was deployed into production many years ago would begin to show performance decay over time (**Figure 6**). Different than problems with *the emergence of a new trend*, *model drift* represents a multifactorial issue that likely reflects the relationship between two variables changing with time, ultimately causing a model to become increasingly unstable with predictions that are less reliable over time.

Generally, the training of ML models follows an assumption of a stationary environment; however, two types of model drift based on non-stationary environments have been described, including: (1) virtual concept drifts and (2) real concept drifts (68). Virtual drifts refer to when the statistical characteristics or marginal distributions of the actual data changes according to a change in time without the target task itself also adjusting similarly (e.g., the regression parameters). Real drifts refer to situations when the relationships between two or more variables in a model are based on a function of time, such that parameters in which the model was trained now becomes obsolete at different points in time (e.g., Pre-Covid vs. Post-Covid) (69). Without considering the possibility of a model drifting, a model can begin to predict outcomes in an unexpected way, which in a healthcare setting could immediately represent incorrect diagnoses being made.

To account for model drift, both active and passive methods have been proposed, of which the later represents the easiest solution to implement (68). Active methods refer to the methodology for detecting this drift and then self-adjusting its parameters to retrain the system to account for this shift, such as by forgetting old information and then updating based on new data (70). However, this methodology is more practical when data is available as a continual stream that will allow a model to continually adapt to recent data inputs. Differently, passive learning methods are reliable in that the performance of a model will be continually or periodically monitored by developers, such as through each release cycle, thus ensuring consistent and reliable results according to the model's original results. As more data becomes available, passive methods could allow users to adapt the model and retrain it based on new data and updated scientific knowledge. Thus, this method could allow for more transparency over time concerning the model's performance, avoiding scenarios where a model may make decisions on new relationships that are non-interpretable or even scientifically unsound.

Problem 11. Practicality Over Hospital Context: Will the IT Department Say Yes?

Systems should be developed according to the environment in which they will be deployed. While this may seem intuitive, there are a number of strict requirements that technology must follow in a healthcare setting that may not be accounted for, especially with cloud-based computing software. Thus, a system should be developed based on how hospitals are organized and specifically how healthcare providers will plan to use these models.

A key concern can be seen with the Health Insurance Portability and Accountability Act of 1996 ("HIPAA") (71)–as well as other patient and individual privacy standards across the globe. Only those who require the handling of patient of data at a given time for the ultimate care of the patient are permitted access to patient data. Therefore, a cloud-based computing system that leaks data to the cloud presents a clear violation. This is not to say cloud-based software cannot be used in medicine given that internet-based methodology demonstrates several beneficial ways to increase the capacity of a hospitals operating system. In fact, current EHR systems represent the standard for the digital storage, organization, and access to healthcare records, and thus cloud-based computing will likely become standard IT infrastructure in the future. Nonetheless, specific rules and

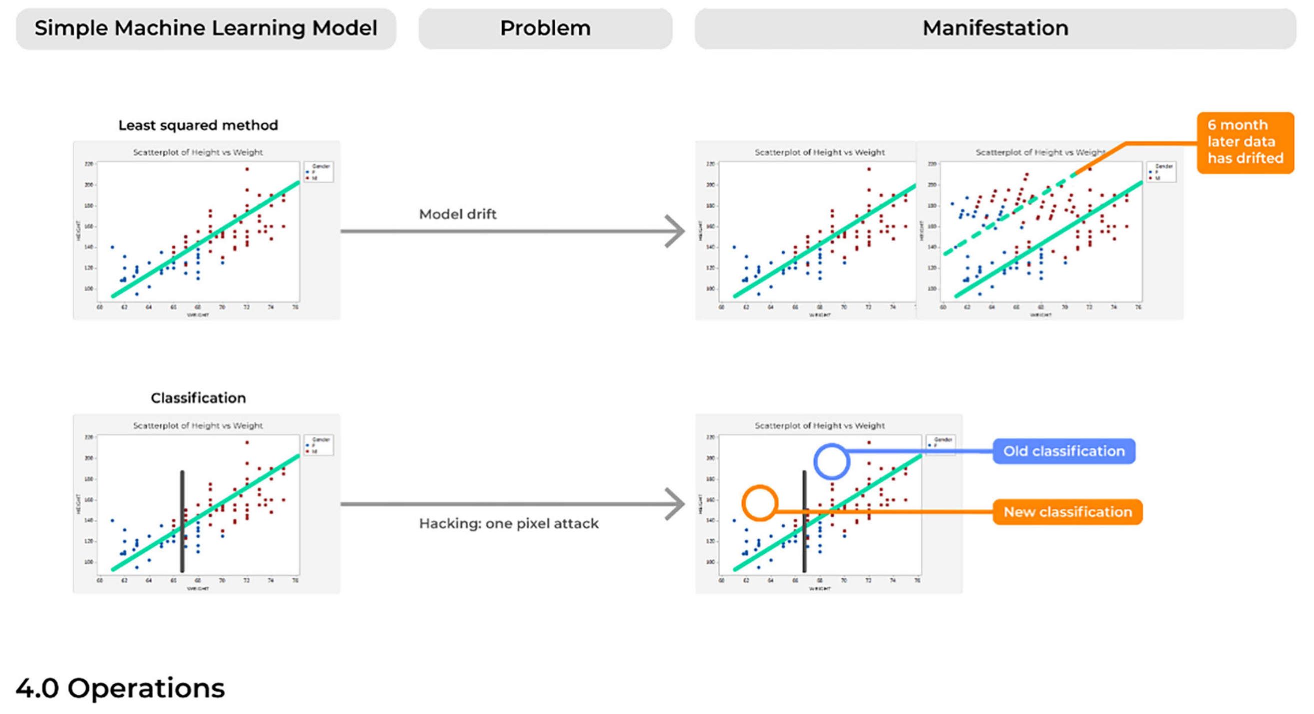
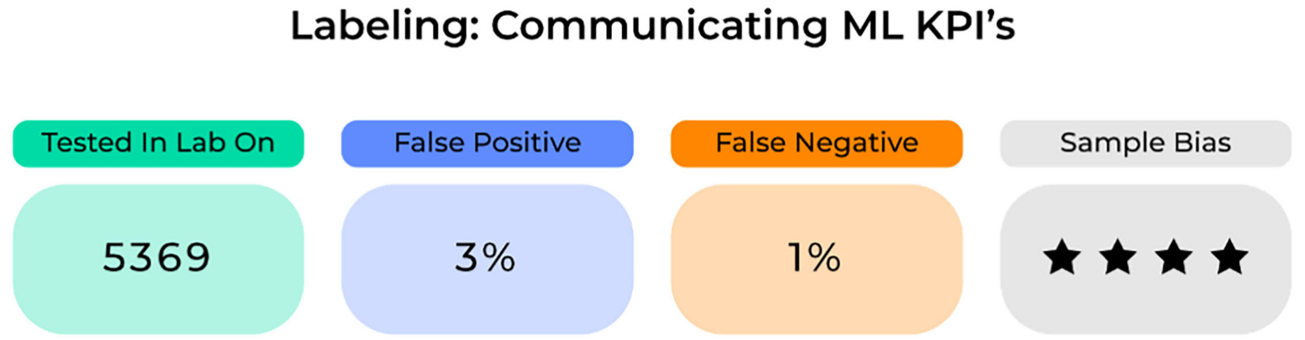

Simple Machine Learning Model
Problem
Manifestation
Least squared method
Scatterplot of Height vs Weight
Model drift
Scatterplot of Height vs Weight
Scatterplot of Height vs Weight
6 month later data has drifted
Classification
Scatterplot of Height vs Weight
Hacking: one pixel attack
Scatterplot of Height vs Weight
Old classification
New classification
4.0 Operations

FIGURE 6 | Concerns with the operations of an ML-system.



Labeling: Communicating ML KPI's

Tested In Lab On	False Positive	False Negative	Sample Bias
5369	3%	1%	★★★★

FIGURE 7 | Standard labeling system according to Key Performance Indicators (KPIs).

regulations must be considered prospectively to adjust a specific system to the HIPPA and IT requirements of a given healthcare system (72, 73). Furthermore, as mentioned previously, if the system implemented is too computing heavy, the model itself may become impractical as it can take hours to run on a less-powered healthcare provider's laptop.

To consider impending concerns of meeting the rigorous standards and requirements contained in the hospital context, developers should meet with the end users and product stakeholders at the beginning of production. In turn, this will allow a clear delineation of the current restraints of an environment that will allow developers to prospectively include user requirements in solution designs.

Problem 12. Hacking: One Voxel Attack

Despite the novelty of advanced ML systems that are highly capable of managing complex data relationships, it must be remembered that ML systems are inherently IT systems which can be similarly fooled and hacked by outsider users.

One of the most common applications of AI is for image classification of radiologic scans. Deep neural networks are highly capable of analyzing imaging scans, allowing them to determine if a scan presents an image of a malignant or benign tumor (1) or can even differentiate between different types of highly malignant tumors often within a time frame unimaginable for humans. Nonetheless, the ability to fool AI models is a long-understood threat, possibly accomplished just by rotating the imaging scan

(74). One particular well-known threat is described as the "one-pixel attack," referring to the ability to drastically fool a neural network by just changing a single pixel in the image being analyzed (75). In turn, this causes the model to classify the image as being of a different class than what is actually represented in the image. Ultimately, this single form of hacking merely suggests the vulnerable nature of ML systems, and also contributes to the truth that we do not always fully understand how a model may be working. Therefore, when a model is failing, we may not be fully aware of this failure. As such, there are profound concerns of similar cyber attacks on ML software in the medical field, especially given the often mere dichotomous classifications asked for by providers with these image-based classification methods (e.g., malignant or benign). Such attacks also present enormous danger to the field of AI itself, which following an attack–could spur long-periods of mistrust with the medical community.

A number of methods have been proposed to prevent the damage from these adversarial attacks. Re-training the model with robust optimization methodology can increase the resistance of a model to these attacks. Increase detection methods to identify attacks may also be appropriate (76). Other methods have also been similarly described, but it remains uncertain the degree to which these methods are better than others for a given scenario. Nonetheless, what is certain is that the integrity and robustness of an AI system must be rigorously examined against known attacks to achieve further safety and trust with applications in the medical field.

FUTURE DIRECTIONS: A STANDARD LABELING SYSTEM FOR MEDICAL AI

Addressing each of the concerns above provides a way to rigorously create a robust model that performs safely and accurately in a field full of potential concerns. However, one way to further advance the field and improve the widespread adoption of these robust technologies is through a standard labeling system that can accurately detect and then convey anomalies in an ML-based system's performance and quality (77).

Common to the most successful business plans are the use of key performance indicators (KPIs) as a way to document success and efficiency. KPIs demonstrate the achievement of measurable landmarks toward reaching company and consumer goals. For an ML model, standard labeling could display KPIs possibly related to the (1) sample it trained and tested on, (2) its quantitative accuracy including information on false positive and negatives, and (3) its risk of specific biases (**Figure 7**). Importantly, these KPIs need to be clearly defined and continuously updated in order for healthcare providers to appropriately examine and incorporate specific ML-based systems into standard clinical practices. Ultimately clinicians will need to assess a model's success to understand where and when to apply it in a given scenario. To make this decision, it will require the use of accurate performance metrics for each model on the target population.

An important concern where standard medical labeling may be of use is in determining medical liability. While it is justified to create incentives for risk control based on the environment of application, the degree of responsibility is less clear cut, and who is ultimately responsible: the *machine* or the *healthcare provider*? While the ethics of AI are outside the scope of this paper, further objective information on a model's performance for a given population with a standard labeling system, possibly contained in the legal section of an ML system, will ultimately improve our objective insight into their performance abilities. In turn, this can give healthcare providers more complete information on whether or not to incorporate these advanced systems in specific scenarios or not.

CONCLUSION

Advancements in the field of artificial intelligence have promised a number of exciting and promising applications for the medical field to improve the quality and delivery of health care services. While there have been remarkable advances in previous years, these applications have yet to fully demonstrate their true potential in clinical applications due to failures in demonstrating reproducible and reliable results as well as the general mistrust of these technologies in the medical community. We outline many of the problems that future developers will likely face that contribute to these failures, specifically related to the approach, data, methodology, and operations of machine learning based system developments. If we continue to prospectively address and manage these concerns with reliable solutions and appropriate system processes in place, then we as a field may further optimize the clinical applicability and adoption of medical based AI technology.

AUTHOR CONTRIBUTIONS

SD: writing, reviewing, editing, and conceptualization. ND: writing, reviewing, and editing. Both authors contributed to the article and approved the submitted version.

REFERENCES

1. McAvoy M, Prieto PC, Kaczmarzyk JR, Fernández IS, McNulty J, et al. Classification of glioblastoma versus primary central nervous system lymphoma using convolutional neural networks. *Sci Rep.* (2021) 11:15219. doi: 10.1038/s41598-021-94733-0
2. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. *Science.* (2021) 371:284–8. doi: 10.1126/science.abd7331
3. Pan D, Zeng A, Jia L, Huang Y, Frizzell T, Song X, et al. (2020). Early detection of alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning. *Front Neurosci.* 14:259. doi: 10.3389/fnins.2020.00259
4. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* (2015) 16: 321–32. doi: 10.1038/nrg3920
5. Medicine, National Academies of Sciences and Medicine. *Improving Diagnosis in Health Care.* Washington, DC; The National Academies Press (2015).

6. Amoroso A. *How Big Tech Investing for the Future.* (2021). Available online at: https://www.icapitalnetwork.com/insights/blog/how-big-tech-is-investing-for-the-future/ (accessed August 15, 2021).
7. Institute TAT. Data science and AI in the age of COVID-19. *Reflections on the Response of the UK's Data Science and AI Community to the COVID-19 Pandemic, The Alan Turing Institute* (2021). Available online at: https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf
8. Roberts MD, AIX-COVNET, Driggs M, Thorpe J, Gilbey M, Yeung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell.* (2021) 3:199–217. doi: 10.1038/s42256-021-00307-0
9. Mhasawade V, Zhao Y, Chunara V. Machine learning and algorithmic fairness in public and population health. *Nat Machin Intell.* (2021) 3:659–66. doi: 10.1038/s42256-021-00373-4
10. Snyder JJ, Salkowski N, Wey A, Pyke J, Israni AK, Kasiske BL, et al. Organ distribution without geographic boundaries: a possible framework for organ allocation. *Am J Transplant*. (2018) 18:2635–40. doi: 10.1111/ajt.15115
11. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. *Nature*. (2013). 496:504–7. doi: 10.1038/nature12060
12. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. (2018) 1:18. doi: 10.1038/s41746-018-0029-1
13. Herper M, Anderson Benches MD. *IBM Watson In Setback for Artificial Intelligence in Medicine.* (2017). Available online at: https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/?sh=5456395a3774 (accessed August 1, 2021).
14. Ross C, Swetlitz I. *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show.* (2018). Available online at: https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf (accessed August 1, 2021).
15. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3
16. Kayser-Bril N. *Google Apologizes After its Vision Ai Produced Racist Results.* (2021). Available online at: https://algorithmwatch.org/en/google-vision-racism/ (accessed August 1, 2021).
17. Zhang DS, Mishra E, Brynjolfsson J, Etchemendy D, Ganguli B, Grosz B, et al. The AI index 2021 annual report. *arXiv[Preprint]*. (2022). arXiv: 2103.06312. Available online at: https://arxiv.org/ftp/arxiv/papers/2103/2103.06312.pdf
18. Baker CM, Burks JD, Briggs RG, Conner AK, Glenn CA, Sali G, et al. A connectomic atlas of the human cerebrum-chapter 1: introduction, methods, and significance. *Oper Neurosurg*. (2018). 15:S1–9. doi: 10.1093/ons/opy253
19. Briggs RG, Lin YH, Dadario NB, Kim SJ, Young IM, Bai MY, et al. Anatomy and white matter connections of the middle frontal gyrus. *World Neurosurg.* (2021) 150:e520–9. doi: 10.1016/j.wneu.2021.03.045
20. Palejwala AH, Dadario NB, Young IM, O'Connor K, Briggs RG, Conner AK, et al. Anatomy and white matter connections of the lingual gyrus and cuneus. *World Neurosurg*. (2021) 151:e426–37. doi: 10.1016/j.wneu.2021.04.050
21. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. *Nature*. (2016) 536:171–8. doi: 10.1038/nature18933
22. Dadario NB, Brahimaj B, Yeung J, Sughrue ME. Reducing the cognitive footprint of brain tumor surgery. *Front Neurol.* (2021) 12:711646. doi: 10.3389/fneur.2021.711646
23. Ren H, Zhu J, Su X, Chen S, Zeng S, Lan X, et al. Application of structural and functional connectome mismatch for classification and individualized therapy in Alzheimer disease. *Front Public Health*. (2020) 8:584430. doi: 10.3389/fpubh.2020.584430
24. Doyen S, Taylor H, Nicholas P, Crawford L, Young I, Sughrue ME. Hollow-tree super: A directional and scalable approach for feature importance in boosted tree models. *PLoS ONE*. (2021) 16:e0258658. doi: 10.1371/journal.pone.0258658
25. O'Neal CM, Ahsan SA, Dadario NB, Fonseka RD, Young IM, Parker A, et al. A connectivity model of the anatomic substrates underlying ideomotor apraxia: a meta-analysis of functional neuroimaging studies. *Clin Neurol Neurosurg.* (2021) 207:106765. doi: 10.1016/j.clineuro.2021.106765
26. Poologaindran A, Profyris C, Young IM, Dadario NB, Ahsan SA, Chendeb K, et al. Interventional neurorehabilitation for promoting functional recovery post-craniotomy: A proof-of-concept. *Sci Rep*. (2022) 12:3039. doi: 10.1038/s41598-022-06766-8
27. Stephens TM, Young IM, O'Neal CM, Dadario NB, Briggs RG, Teo C, et al. Akinetic mutism reversed by inferior parietal lobule repetitive theta burst stimulation: can we restore default mode network function for therapeutic benefit? *Brain Behav*. (2021). 11:e02180. doi: 10.1002/brb3.2180
28. Colubri A, Hartley MA, Siakor M, Wolfman V, Felix A, Sesay T, et al. Machine-learning prognostic models from the 2014-16 ebola outbreak: data-harmonization challenges, validation strategies, and mHealth applications. *EClinicalMedicine*. (2019) 11:54–64. doi: 10.1016/j.eclinm.2019.06.003
29. Powell K. NVIDIA and King's College London Announce MONAI open source AI framework for healthcare research. *Domain-Optimized, PyTorch-Based Project Aids Researchers Developing AI in Healthcare.* (2020). Available online at: https://blogs.nvidia.com/blog/2020/04/21/monai-open-source-framework-ai-healthcare (accessed November 20, 2015).
30. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. (2020) 369: m1328. doi: 10.1136/bmj.m1328
31. Times TNY. *Tracking Coronavirus in New York: Latest Map and Case Count.* (2021). Available online at: https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html (accessed August 20, 2021).
32. Kostick-Quenet K, Mandl KD, Minssen T, Cohen IG, Gasser U, Kohane I, et al. How NFTs could transform health information exchange. *Science*. (2022) 375:500–2. doi: 10.1126/science.abm2004
33. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* (2018) 138:1529–38. doi: 10.1016/j.jid.2018.01.028
34. Louie P, Wilkes R. Representations of race and skin tone in medical textbook imagery. *Soc Sci Med*. (2018) 202:38–42. doi: 10.1016/j.socscimed.2018.02.023
35. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatology*. (2018) 154:1247–48. doi: 10.1001/jamadermatol.2018.2348
36. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Vol. 81*. PMLR (2018). p. 77-91. Available online at: https://proceedings.mlr.press/v81/buolamwini18a.html
37. Krishnan A, Almadan A, Rattani A. Understanding fairness of gender classification algorithms across gender-race groups. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2020).
38. Frost JA, Binder JR, Springer JA, Hammeke TA, Bellgowan PSF, et al. Language processing is strongly left lateralized in both sexes: evidence from functional MRI. *Brain*. (1999) 122:199–208. doi: 10.1093/brain/122.2.199
39. Qi Z, Han M, Garel K, Chen ES, Gabrieli JDE. White-matter structure in the right hemisphere predicts Mandarin Chinese learning success. *J Neurolinguistics*. (2015) 33:14–28. doi: 10.1016/j.jneuroling.2014.08.004
40. Crinion JT, Green DW, Chung R, Ali N, Grogan A, Price GR, et al. Neuroanatomical markers of speaking Chinese. *Hum Brain Mapp*. (2009) 30:4108–15. doi: 10.1002/hbm.20832
41. Wang XD, Xu H, Yuan Z, Luo H, Wang M, Li HW, et al. Brain hemispheres swap dominance for processing semantically meaningful pitch. *Front Hum Neurosci*. (2021). 15:621677. doi: 10.3389/fnhum.2021.621677
42. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. (2019) 2:e191095. doi: 10.1001/jamanetworkopen.2019.1095
43. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol.* (2019) 20:405–10. doi: 10.3348/kjr.2019.0025

44. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. (2021) 27:1735–43. doi: 10.1038/s41591-021-01506-3
45. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics*. (2021) 22:122. doi: 10.1186/s12910-021-00687-3
46. Hayati M, Biller P, Colijn C. Predicting the short-term success of human influenza a variants with machine learning. *bioRxiv*. (2019). doi: 10.1101/609248
47. Zhao N, Yuan LX, Jia XZ, Zhou XF, Deng XP, He HJ, et al. Intra- inter-scanner reliability of voxel-wise whole-brain analytic metrics for resting state fMRI. *Front Neuroinform*. (2018). 12:54. doi: 10.3389/fninf.2018.00054
48. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci*. (2017) 18:115–26. doi: 10.1038/nrn.2016.167
49. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M, et al. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med*. (2021) 13:eabb1655. doi: 10.1126/scitranslmed.abb1655
50. Collins GS, Reitsma JB, Altman DG, Moons KG, Group T. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. *Circulation*. (2015) 131:211–9. doi: 10.1161/CIRCULATIONAHA.114.014508
51. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. (2019) 393:1577–9. doi: 10.1016/S0140-6736(19)30037-6
52. Asan, O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res*. (2020) 22:e15154. doi: 10.2196/15154
53. Linardatos PV, Papastefanopoulos S, Kotsiantis. Explainable AI: a review of machine learning interpretability methods. *Entropy*. (2020) 23:18. doi: 10.3390/e23010018
54. Titano JJM, Badgeley J, Schefflein M, Pain A, Su M, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. (2018) 24:1337–41. doi: 10.1038/s41591-018-0147-y
55. Zech H. Liability for AI: public policy considerations. *ERA Forum*. (2021) 22:147–58. doi: 10.1007/s12027-020-00648-0
56. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med*. (2018) 378:981–983. doi: 10.1056/NEJMp1714229
57. Guidance W. *Ethics Governance of Artificial Intelligence for Health*. World Health Organization (2021). Available online at: https://www.who.int/publications/i/item/9789240029200
58. Fung P, Zaidan H, Timonen JV, Niemi A, Kousa J, et al. Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. *J Aerosol Sci*. (2021) 152:105694. doi: 10.1016/j.jaerosci.2020.105694
59. Li X, Xiong H, Li X, Wu X, Zhang X, Liu J, et al. Interpretable deep learning: Interpretations, interpretability, trustworthiness and beyond. *arXiv[Preprint]*. (2021). arXiv: 2103.10689. Available online at: https://arxiv.org/pdf/2103.10689.pdf
60. Lundberg SM, Lee SI. *Unified Approach to Interpreting Model Predictions*. Curran Associates, Inc. (2017).
61. Radečić D. *LIME: How to Interpret Machine Learning Models With Python. Explainable Machine Learning at Your Fingertips*. (2020). Available online at: https://towardsdatascience.com/lime-how-to-interpret-machine-learning-models-with-python-94b0e7e4432e (accessed November 20, 2021).
62. Buhrmester V, Munch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Mach Learn Knowl Extr*. (2021) 3:966–89.
63. Hashem S, Nisar S, Bhat AA, Yadav SK, Azeem MW, Bagga P, et al. Genetics of structural and functional brain changes in autism spectrum disorder. *Transl Psychiatry*. (2020) 10:229. doi: 10.1038/s41398-020-00921-3
64. Power JD, Schlaggar BL, Petersen SE. Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage*. (2015) 105:536–51. doi: 10.1016/j.neuroimage.2014.10.044
65. Danielson ML, Bitsko RH, Ghandour RM, Holbrook JR, Kogan MD, Blumberg SJ, et al. Prevalence of parent-reported ADHD diagnosis associated treatment among U.S. Children and adolescents. *J Clin Child Adolesc Psychol*. (2016) 47:199-212. doi: 10.1080/15374416.2017.1417860
66. Pelham WE, Foster EM, Robb JA. The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *J Pediatr Psychol*. (2007) 32:711–27. doi: 10.1093/jpepsy/jsm022
67. Eslami T, Almuqhim F, Raiker JS, Saeed F. Machine learning methods for diagnosing autism spectrum disorder and attention- deficit/hyperactivity disorder using functional and structural MRI: a survey. *Front Neuroinform*. (2021) 14:575999. doi: 10.3389/fninf.2020.575999
68. Straat MF, Abadi C, Göpfert B, Hammer, Biehl M. Statistical mechanics of on-line learning under concept drift. *Entropy*. (2018) 20:775. doi: 10.3390/e20100775
69. Ditzler G, Roveri M, Alippi C, Polikar R. Learning in nonstationary environments: a survey. *IEEE Comput Intell Mag*. (2015) 10:12–25. doi: 10.1109/MCI.2015.2471196
70. Faria ER, Gonçalves IJCR, de Carvalho ACPLF, Gama JM. Novelty detection in data streams. *Artif Intell Rev*. (2016) 45:235–69. doi: 10.1007/s10462-015-9444-8
71. Tariq RA, Hackert PB. *Patient Confidentiality*. StatPearls. Treasure Island (FL), StatPearls Publishing Copyright © 2021, StatPearls Publishing LLC (2021).
72. Chard K, Russell M, Lussier YA, Mendonça EA, Silverstein JC. A cloud-based approach to medical NLP. *AMIA Annu Symp Proc*. (2011) 2011:207–16.
73. Schweitzer EJ. Reconciliation of the cloud computing model with US federal electronic health record regulations. *J Am Med Inform Assoc*. (2012) 19:161–5. doi: 10.1136/amiajnl-2011-000162
74. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. (2019) 363:1287–89. doi: 10.1126/science.aaw4399
75. Sipola T, Kokkonen T. *One-Pixel Attacks Against Medical Imaging: A Conceptual Framework. Trends and Applications in Information Systems and Technologies*, Cham, Switzerland, Springer International Publishing (2021).
76. Xu H, Ma Y, Liu HC, Deb D, Liu H, Tang JL, et al. Adversarial attacks and defenses in images, graphs and text: a review. *Int J Autom Comput*. (2020) 17:151–78. doi: 10.1007/s11633-019-1211-x
77. Zhao NJ, Zhu R, Liu D, Liu M, Zhang D. Label-less: a semi-automatic labelling tool for kpi anomalies. In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications* Paris, France (2019).

Conflict of Interest: SD is a co-founder, chief data scientist, and shareholder for Omniscient Neurotechnology.

The remaining author declares 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.