Curr Treat Options Cardio Med (2024) 26:29–45
DOI 10.1007/s11936-023-01032-0

Mitigating Bias in Clinical Machine Learning Models

Julio C. Perez-Downes, DO[1]
Andrew S. Tseng, MD[1]
Keith A. McConn, BA[2]
Sara M. Elattar, MD[1]
Olayemi Sokumbi, MD[3]
Ronnie A. Sebro, MD, PhD[4]
Megan A. Allyse, PhD[5]
Bryan J. Dangott, MD[6]
Rickey E. Carter, PhD[5]
Demilade Adedinsewo, MD, MPH[1,*]

Address
[*,1]Department of Cardiovascular Medicine, Mayo Clinic, Jacksonville, 4500 San Pablo Rd, S. Jacksonville, FL 32224, USA
Email: adedinsewo.demilade@mayo.edu
[2]Alix School of Medicine, Mayo Clinic, Rochester, MN, USA
[3]Departments of Dermatology and Laboratory Medicine & Pathology, Mayo Clinic, Jacksonville, FL, USA
[4]Department of Radiology, Mayo Clinic, Jacksonville, FL, USA
[5]Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL, USA
[6]Department of Laboratory Medicine and Pathology, Mayo Clinic, Jacksonville, FL, USA

Published online: 10 February 2024

Keywords Artificial intelligence · Bias · Digital technology · Health disparities · Machine learning

Abbreviations *AI* Artificial intelligence · *AIM-AHEAD* Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity · *CONSORT-AI* Consolidated Standards of Reporting Trials-Artificial Intelligence · *COVID-19* Coronavirus disease-19 · *ECG* Electrocardiogram · *EGM* Electrogram · *ML* Machine learning · *NAII* National artificial intelligence initiative · *PRIME* Proposed requirements for cardiovascular imaging-related machine learning evaluation · *SPIRIT-AI* Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence · *US* United States · *US FDA* United States Food and Drug Administration · *WHO* World Health Organization

Abstract

Purpose of review Identifying the risk for and addressing bias in clinical machine learning models is essential to reap its full benefits and ensure health equity. We provide a review of the machine learning landscape in clinical medicine, highlight ethical concerns with a particular focus on algorithmic bias, and offer a framework for mitigating bias.

Recent findings Machine learning, the computational framework that supports artificial intelligence, now plays a significant role in everyday life and its potential role in clinical medicine continues to increase exponentially. Multiple machine learning models have demonstrated outstanding performance, surpassing human abilities with specific tasks, and are poised to revolutionize clinical research and practice over the next few years. While machine learning can augment clinician's diagnostic capabilities, support clinical decision-making, and improve health care efficiency, they are not infallible. One key concern with the use of machine learning models is algorithmic bias, which if present poses a non-trivial risk to patient care particularly if algorithms are used in a population different from that used to create the algorithm. Recommendations and methods to identify and mitigate algorithmic bias to ensure responsible development of machine learning models are summarized.

Summary With the anticipated widespread adoption of machine learning in medicine, significant ethical concerns remain, particularly the risk for bias. Researchers, model developers, and end users need to be aware of the potential for bias, its associated risk, and methods to guard against it prior to deploying it for clinical use.

Opinion statement

This review article summarizes the use of machine learning in clinical medicine and evaluates bias in this context. We also discuss existing mechanisms and systems geared towards mitigating bias and propose additional recommendations to ensure responsible model development and deployment.

Introduction

With technological advancements over the last decade, artificial intelligence (AI) and machine learning (ML) are increasingly prevalent in everyday life [1] and will continue to play an important and expanding role in the way we live, work, and play. Advancements in AI technology were facilitated by the availability of large amounts of digital data and contemporary computer chip processors with more efficient computing capabilities [2]. In today's landscape, advanced AI/ML models can analyze, synthesize, and generate solutions to common problems that exceed human-level performance. The development of ML models in the field of medicine continues to expand, as evidenced by the exponential increase in the number of scientific publications on this topic [3]. Applications include the ability to predict the likelihood of cardiac dysfunction using data from an electrocardiogram [4] to rapid detection of cancer from radiographic images [5].

A key drawback of ML models in medicine for biomedical research and in clinical care algorithms is the potential to introduce biases reflective of data used to train the model in what has been referred to as algorithmic bias [6••]. Algorithmic bias may go unrecognized if inappropriate metrics of success are chosen, such as model validation only in populations similar to the training sample. Another form of bias can occur when ML models are trained to answer the wrong question, where training endpoints do not accurately match the intended prediction or outcome. The reality of the risk associated with biased models has been demonstrated with facial recognition algorithms [7–9] performing poorly among dark skinned females and the use of commercially developed and marketed recidivism prediction models by law enforcement agencies which have repeatedly overestimated threats for adults of African descent living in the USA [10, 11]. Algorithmic bias has also been observed with ML models in clinical medicine with poorer performance noted among women and patients from racial and ethnic minority groups [12•, 13•]. In this review article, we provide a brief overview of ML and AI models, review various uses in clinical research and healthcare, discuss algorithmic bias, and offer potential solutions to addressing bias with ML models intended for use in medicine.

Machine learning in clinical research and medicine

AI is a broad term used to describe the ability of machines or computers to perform functions that typically require human intelligence. AI relies on machine learning to develop generalized algorithms that enable near human, or in some cases super-human, levels of performance on these tasks. Machine learning refers to a class of methods that enable machines to learn generalized, specific, and or complex associations from data. Frequently in medicine, this learning process enables artificial intelligence algorithms that were trained on large datasets to make predictions or classifications through learned patterns or features in data [14]. There are various types of ML which can be coarsely differentiated by how they are trained with respect to information about the "truth" and the complexity of the model architecture used in the algorithm. For the training aspects, approaches are generally defined as either supervised or unsupervised. Supervised learning presents to the algorithm not only the data intended to make the prediction but also information about the true status of the individual case, which is often referred to as the "label." Unsupervised approaches withhold the label and instead ask the algorithm to identify combinations or profiles of data that are similar within the profile yet distinct across profiles effectively generating empirical labels for the data without human guidance.

In terms of machine learning architectures, ordinary logistic regression is an example of a very simple model architecture whereas a convolutional neural network may have a very complex architecture. With the advent of high-performance computing, utilizing graphical processing units (GPUs), advances in how models are optimized ("trained") have been accomplished

that have facilitated training models with thousands (in some cases hundreds of thousands) of model parameters. In the case of neural networks, as the model architecture grows in complexity and the number of successive layers (chained calculations defined by the network) increases within the model architecture, the modeling framework takes on the name deep learning to try to express the scale of the model's computational framework [15]. Some examples of these ML types are discussed later in this article.

AI use in medicine is often classified as augmented/assistive intelligence and autonomous intelligence. The former encompasses task-specific and domain-specific AI systems developed to assist clinicians with clinical decisions and patient care whereas the latter refers to an AI system that does not require clinician interpretation to make patient care recommendations [16]. It is believed that fully autonomous algorithms are unlikely to replace health care providers, but rather clinicians will interact with these across a continuum of automation [16, 17]. Generalized autonomous intelligence for broad use in medical settings, with no input from a managing physician in all phases of patient care, currently does not exist.

Current applications of machine learning in clinical medicine

Physiologic signals

This involves analyzing biological signals and using these to predict specific outcomes. Examples of these signals include surface electrocardiograms (ECG) and intracardiac electrograms (EGM) which capture cardiac electrical activity from outside and within the heart, respectively, electroencephalograms (electrical activity from the brain) (EEG), electromyogram (muscle electrical activity) (EMG), and actimetry (limb activity/displacement) [18]. Of these, ECGs have been extensively studied. Using deep learning, several studies have demonstrated the ECG's ability to detect multiple cardiovascular pathologies which supersedes human interpretation of the ECG signals. These include detection of left ventricular dysfunction [4, 19, 20], pregnancy-related cardiomyopathy [21], silent atrial fibrillation [22, 23], hypertrophic cardiomyopathy [24], cardiac amyloidosis [25], valvular heart disease [26, 27], and cardiac allograft rejection [28].

Medical images

Human interpretation of medical images is a deeply embedded practice in multiple medical specialties. The use of AI to aid in the extraction of useful information from medical images for localization, segmentation, registration, classification and prediction purposes, or image refinement to augment clinical interpretation is emerging in importance [29]. Its use cuts across several medical fields and subspecialties. In the field of cardiology, this technique has been demonstrated with AI-generated cardiac ultrasound image annotations for assessment of left ventricular ejection fraction [30, 31].

In the field of diagnostic radiology, ML algorithms, specifically deep learning, have been extensively utilized to help improve diagnostic accuracy and efficiency, with brain, breast, eye, chest, musculoskeletal, and abdominal imaging [29]. For example, during the coronavirus-19 (COVID-19) pandemic prior to the development of a rapid reverse transcriptase polymerase chain reaction (RT-PCR) test, deep learning was used to analyze computed tomography (CT) images of the chest in patients with suspected COVID-19 [17]. This model had an accuracy of 96%, AUC of 0.95, and sensitivity of 89% to accurately differentiate COVID-19 from other pneumonias [17]. A revolutionary utilization of AI has emerged in the field of breast imaging, where computer-aided diagnosis is utilized as standard of care to facilitate improvement of cancer detection rates at earlier stages than previously done [32]. Additional applications include efficient triaging of studies that necessitate prompt evaluation, improvement of image quality which facilitates diagnoses, and potentially enabling a more accurate assessment of disease progression [5, 33].

In the field of dermatology, a specialty that relies heavily on pattern recognition, ML is playing a groundbreaking role in diagnostics and assessments. The large clinical, dermatoscopic, and histopathologic image databases have enabled dermatologic studies focusing on early diagnosis of cutaneous disorders. A landmark study in the use of ML in dermatology demonstrated competence comparable to board-certified dermatologists in identifying most common skin cancers and in identifying the deadliest skin cancer, malignant melanoma [34]. Although there is enormous potential for ML to expand the reach of dermatologic care access, the lack of enough images with diverse skin tones limits the accurate training of algorithms and represents a substantial bias in available datasets. A recent systematic review of publicly available skin cancer image datasets revealed both poor reporting and poor representation of Fitzpatrick skin type. In a review with available skin type information from three datasets with 2436 images, only ten images were Fitzpatrick skin type V and only a single image was from skin type VI [35]. Similarly, in the International Skin Imaging Collaboration: Melanoma Project, which is one of the largest and often-used, open-source, public-access archives of pigmented lesions, the patient data comprise predominantly fair-skinned individuals in the USA, Europe, and Australia [36, 37]. This bias is of significance especially when considering the varied presentation of skin cancer in skin of color populations. For instance, although cutaneous melanoma incidence is highest among non-Hispanic White persons, non-White individuals have been observed to present with later stage melanoma at diagnosis and have lower overall survival outcomes emphasizing the need for early detection through ML in non-White persons [38]. If ML models are inadequately trained on darker skin types, even the most advanced algorithm will likely perform poorly with images in skin of color [39]. Aware of this limitation, there is ongoing intentional effort for image repositories around the world to include photos of darker skin types to ensure algorithms are trained to meet the dermatologic needs of all patients while avoiding the exacerbation of existing disparities in dermatologic care for patients with skin of color.

The potential applications of machine learning in digital pathology (DP) are extensive with research and industry applications already showing

promising results in spatial analysis and immuno-oncology [40]. Clinically, there are many opportunities for semi-automated workflows to provide more consistent pathology results; however, DP remains a young field and clinical deployments are currently limited to early adopters [41–43]. The global regulatory environment has also played a role in the adoption and penetration of DP with European DP regulatory approvals occurring a few years before the USA [43]. Multiple factors including economic, regulatory, and technical difficulties limit slide scanning and digitization in specialties such as hematology and cytology. Consequently, available digitized slides for ML model development only represent a small fraction of pathology slides worldwide with the potential for bias in the datasets. While this is unintentional, ML algorithms developed with these limited datasets may face challenges with scalability. In addition, algorithms developed from images scanned by one DP vendor may not perform well when presented with images from another DP vendor. Over time, we believe widespread adoption of DP and curation of joint data repositories will enrich DP datasets by increasing absolute numbers available for training, case variety, and diversity to support development of robust ML models.

Acoustic signals

This involves the analysis of sounds for diagnostic purposes. Examples in medicine include the use of heart sounds (phonocardiograms), lung sounds, and voice-based sounds. This has been demonstrated with automated detection of valvular heart disease [44, 45], improved classification of lung auscultatory sounds [46], and non-invasive diagnosis of COVID-19 from cough recordings [47].

Text processing

One of the more common examples in clinical and non-clinical environments is text processing, and it refers to the analysis and interpretation of text (numeric and words) and speech with ML where model outputs are either used to augment diagnostic capacity or assist with patient care by answering medical questions. These types of models have demonstrated utility with disease or clinical outcome prediction [48–50] and identification of disease phenotypes [51, 52].

More recently, significant advancements have been made with generative AI to produce human-like responses to text or speech-based inputs. Generative AI algorithms have been trained on data that are largely found openly available online. The training of these models extends the concepts of natural language processing to learn not only the basic elements of speech but also the predictable patterns of word usage in the context of how a topic is summarized or reported on. In general, there is a predictable flow for how a recipe is written online or a scientific article is composed. Generative AI learns these structures and can assemble new works based on the underlying probability structure estimated from many examples. ChatGPT (Chat Generative Pre-trained Transformer) is an example of this, released by OpenAI

in November 2022 with a refined version GPT-4 released in March 2023 [53]. In one study, ChatGPT was shown to provide higher quality and more empathetic responses to patient questions when compared to physicians [54]. While impressive, the sources of data used to train the model are not always accurate. The saying "do not believe everything you read online" is taking on new meaning in the era of generative AI. Furthermore, at least in the context of science, there are some notable differences present in text generated by generative AI [55]. Another concern is that the performance of large language models (LLMs) like ChatGPT seems to decrease or decay with time. It is unclear whether this is due to changes made in the algorithm to speed up convergence since there are a large number of users or whether this is related to the training of the model on progressively less accurate data [56].

Algorithmic bias

A commonly cited limitation with deep learning, is its "black box" nature. The complexity in the model architecture and the long series of internal calculations make it challenging to clearly identify which specific features in an input (image, signal, or dataset) are being used for model prediction. While there are tools such as saliency maps, gradient-weighted class activation maps (Grad-CAMS) [57], and Shapley Additive explanations (SHAP explainers) that help identify components of the input data, these are often approximations of the entire modeling process. Furthermore, the algorithms can only learn from what they are given. As a result, these systems are highly dependent on the training datasets from which it learns to make predictions. ML models may demonstrate bias inherent in the underlying dataset, resulting in predictions that may contribute to healthcare disparities related to race, sex, or socioeconomic status [58]. In a classical statistical context, "extrapolation" of the model beyond the data was a common warning given to all people learning modeling. This same warning applies to ML; however, the concept of extrapolation is far more nuanced given the complexity of the data and the resulting algorithm. Another challenge with newly created ML models using contemporary or retrospective data is that the model is trained to recapitulate the outcomes seen during the time period when the data was obtained. For example, a model trained to predict college acceptance using data from the 1960s would very likely show that male sex is a strong predictor of acceptance. These ML models are therefore at risk of encrusting temporal societal biases in their predictions. Bias can also occur when ML models are trained to answer the wrong question, i.e., predicting a biased proxy variable believed to represent the actual outcome of interest. These types of bias are collectively referred to as algorithmic bias [6••]. Examples of bias inherent in training datasets include specific variables or features that favor a specific racial group based on past discriminatory practices [59] or underrepresentation of certain groups or individuals as demonstrated with commercial facial recognition algorithms, which showed near perfect discrimination among light skinned males but high error rates among dark skinned females [8].

The high cost of algorithmic bias has been demonstrated multiple times in non-healthcare domains which have led to unfair hiring practices [60, 61] and erroneous identification or penalization of individuals by the criminal justice system [10, 11]. Due to the potential for serious adverse health care consequences, it is critical that ML models developed for use in clinical care are thoroughly evaluated for bias and intentional steps taken to mitigate it in a systematic fashion.

Mitigating bias and responsible artificial intelligence

Ethics in machine learning

Addressing ethical issues surrounding the clinical integration of AI/ML is essential to ensuring that these technologies are translated into broader use in a just manner. Multiple ethical challenges have been identified with the use of AI/ML for health care: algorithmic bias, privacy, cybersecurity, data ownership, accountability, autonomous systems, the digital divide, impact on labor and employment, commercialization, governance, and impact on climate change (Fig. 1). In this article, our discussions will focus on mitigating algorithmic bias.

Efforts directed towards mitigating bias in AI and ML models are often referred to as responsible artificial intelligence. This broadly encompasses the following domains: *inclusivity*—ensuring women and racial/ethnic minority groups are adequately represented in training datasets; *specificity*—ensuring that appropriate and specific training targets are selected when developing ML models; *transparency*—ensuring standard reporting to include information regarding training data, model annotation, and interpretability; and

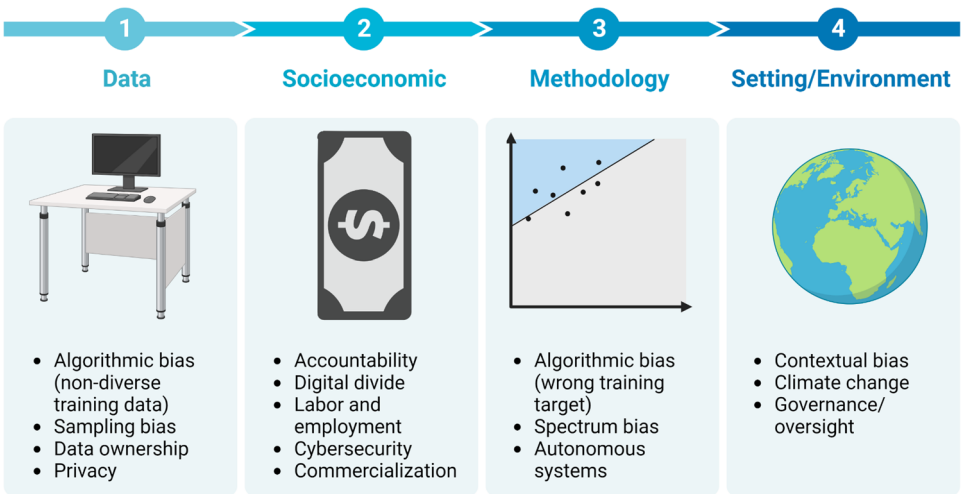

Fig. 1 Ethical challenges in machine learning for clinical research and practice. This figure illustrates four key overarching ethical issues in machine learning—data, socioeconomic, methodology, and environment-related challenges. It also lists examples within each category that machine learning model developers and end-users need to be aware of to adequately evaluate bias and establish steps to mitigate it. Illustration created with Biorender.com.

validation—conducting rigorous testing/auditing, validation studies (internal and external), and clinical trials as appropriate prior to deploying ML models for use in clinical care [62, 63] (Fig. 2).

There are a few governing organizations that have provided legal frameworks to regulate ML models and ensure ethical concerns are addressed. The European Union's AI act aims to stratify AI applications by levels of risk and accordingly, either ban or regulate by conformity assessment [64]. To date, there are proposed bills introduced in the US congress to address the utilization and implementation of AI [65]. While this national legislation is debated and modified, there is a patchwork of state and local legislation addressing this gap. New York City's Local Law 144 [66] (which requires bias audits of AI-enabled tools used for employment decisions) is an example of this [60, 67]. In addition, the Blueprint for an AI Bill of Rights, a non-binding framework released by the White House in October 2022, details five principles that seek to guide the design and implementation of AI, including (1) safe and effective systems, (2) algorithmic discrimination practices, (3) data privacy, (4) notice and explanation, and (5) human alternatives, consideration, and fallback [60]. Other proposed ethical guardrails include UNESCO's Recommendation on the Ethics of Artificial Intelligence and the United States Intelligence Community's Artificial Intelligence Ethics Framework [68, 69].

Guidelines and recommendations

Guidelines are essential to facilitate equitable development and validation of ML models and inform developers in promoting transparency in the design and reporting of AI algorithms [1–3]. As the role of AI/ML in clinical

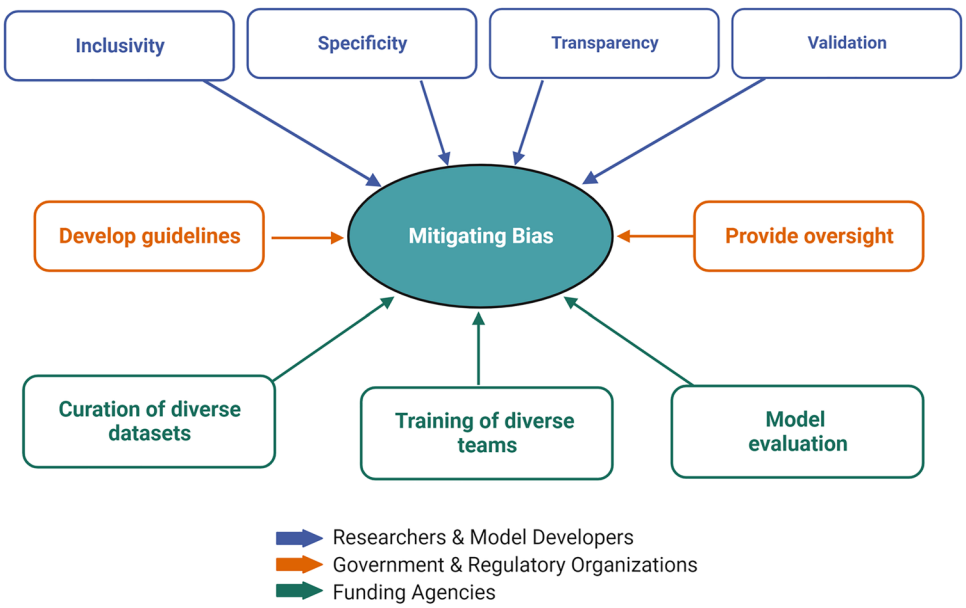

Fig. 2 Framework for mitigating bias in clinical machine learning models. Illustration created with Biorender.com.

medicine continues to expand, it is critical that human autonomy is preserved and that appropriate guidelines are developed and adopted for responsible utilization of this emerging technology. As of July 2022, 521 AI/ML-enabled devices had received US FDA approval with the majority being in the fields of radiology and cardiology [60, 70, 71]. At this time, many regulatory guidelines remain in development by a number of governmental authorities which aim to critically evaluate applications of AI/ML in medicine and ensure its trustworthiness [60, 66]. One challenge is that the stewards—governmental authorities, and regulatory staff often lack the technical expertise to evaluate these models adequately and appropriately.

The World Health Organization (WHO) is among the first to develop and publish a guidance document and propose a framework for governance of AI/ML for health [72]. It highlights the following 6 ethical principles: (1) protecting autonomy, (2) promoting human well-being, human safety, and the public interest, (3) ensuring transparency, explainability, and intelligibility, (4) fostering responsibility and accountability, (5) ensuring inclusiveness and equity, and (6) promoting artificial intelligence that is responsive and sustainable [72, 73].

Framework for addressing bias and ensuring responsible AI

Researchers and ML model developers

Intentional efforts to ensure responsible AI at the model development phase (inclusivity, specificity, transparency, and validation) often lie with researchers and developers of ML models.

To address *inclusivity*, a few use cases are described. For example, while AI/ML holds promise in improving healthcare delivery and lowering costs in low-middle income countries (LMIC), one key limitation is the unavailability of high-quality data, from LMIC countries, needed to train AI/ML models in an equitable manner that represents the characteristics and unique aspects of the population [74, 75]. It is important for researchers, AI developers, and local health systems to invest in curating digital training datasets for this purpose, especially if these models are intended for LMIC use. In the USA, these translate to ensuring inclusion of diverse racial backgrounds and in some cases consider oversampling of racial and ethnic minority groups and patient populations for which these models are intended for use [74–76]. Questions to consider include the following:

- Will this study include the appropriate population that would be representative of the target population (i.e., avoid sampling bias)?
- Will AI/ML model development utilize techniques and methods to minimize overfitting and other potential programming-related biases?

Fundamentally, researchers must first ask the right question and design a study that is appropriate to answer the question, i.e., *specificity*. During model development and in the study design phase, the potential for bias must be

meticulously considered and pre-emptively addressed. Relevant questions to consider here are as follows:

- Will the study design be adequate to address the clinical question (e.g., inclusion of the appropriate spectrum of disease severity, i.e., spectrum bias)?
- Is the model being used and applied in an appropriate population for which it was developed?
- Will the model be useful in a different setting, e.g., LMICs and poor resource settings with limited access to technology (contextual bias) [77]?
- Will there be equitable access to the model by all populations?

With regard to *transparency*, two different guidelines for AI related study protocols and reporting AI clinical trial interventions have been developed. These are based on the 2013 Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT 2013) and the Consolidated Standards of Reporting Trials (CONSORT 2010) statements. These updated guidelines are referred to as SPIRIT-AI and CONSORT-AI. Additional questions for developers to consider are the following:

- How will financial incentives influence the implementation of the model?
- How can we balance financial and clinical considerations when marketing an AI/ML-derived product?

Finally with *validation*, this must include internal and external validation. External validation should include evaluation in multiple health systems and settings (inpatient vs. outpatient for example), diverse patient populations, retrospective and prospective evaluations. Implementation studies are also crucial in evaluating the feasibility of incorporating ML tools into current clinical practice and lastly clinical trials which allows the objective evaluation of the ML model's impact on clinical outcomes.

Funding agencies

In September of 2022, the National Institute of Health (NIH) announced it will invest $130 million USD to expand the use of AI/ML in biomedical and behavioral research, the bridge to artificial intelligence (Bridge2AI) program [78]. As part of this effort, the NIH will support ethical data curation and use, build diverse teams/workforce with AI/ML expertise, as well as efforts to reduce bias. It is important to consider the training of individuals from racial and ethnic minority groups to perform AI and ML research. These researchers may be able to detect nuanced biases in the AI/ML models because of cultural differences. This will encourage racial and ethnic minority researchers to contribute to the narrative around the research and how it affects their community.

Additional NIH-related efforts include specific funding opportunities to support the ethical development of AI/ML models in biomedicine [79],

the health equity and researcher diversity program (AIM-AHEAD) [80], the Science Collaborative for Health disparities and Artificial intelligence bias Reduction (ScHARe) platform [81], and the launch of a prize competition titled "bias detection tools in health care challenge" which concluded in March 2023 [82]. It is imperative that the NIH and other research funding agencies continue to support these programs, in addition to promoting and funding efforts to evaluate existing models for bias through validation studies, and the development of novel tools to mitigate bias in AI/ML [81, 82].

Government and regulatory organizations

The national artificial intelligence initiative (NAII) [83] established in 2021 has been tasked with the developing guidance for regulating AI. While this and a bill of rights are still in progress, some US agencies have adopted some guidelines and principles for AI/ML use developed by the department of defense and the office of the director of national intelligence in 2020 to promote trustworthy use of AI in the federal government. In April 2023, four government agencies also released a joint statement on guarding against discrimination and bias in AI systems [68] with plans to use existing civil and consumer rights laws to enforce this.

The American College of Cardiology Innovation Council developed the PRIME checklist for AI/ML-derived algorithms [84] in which one critical component is the requirement to report model-related bias. These are part of the efforts seeking to standardize scientific reporting and evaluation of AI/ML algorithms and systematically evaluate bias. In addition to these efforts, government agencies, policymakers, and regulating bodies need to establish clear regulations and guidelines to ensure that consumer protection standards are in place and that bias and conflicts of interest are adequately addressed.

Conclusion

As novel machine learning algorithms are developed and refined, their use will become increasingly integrated into our daily lives. Its role in medicine will continue to expand by facilitating personalized and precision medicine, holding promise for earlier diagnosis, improved treatment of disease, and health promotion [85]. It is imperative that these systems are developed, utilized, and implemented in a manner that ensures everyone will benefit from the use of these technologies for healthcare. The words of Martin Luther King Jr. could not be more relevant at this time: "Of all the forms of inequality, injustice in health is the most shocking and the most inhuman", as such, it is critical that we are all aware of the significant risk algorithmic bias poses to healthcare and that intentional efforts are put in place to guard against it. Recognizing and addressing bias will not only ensure equitable use of AI/ML models but more importantly facilitate optimal, safe, and efficient health care for all people.

Acknowledgements

Manuscript illustrations were created using BioRender.com.

Author contributions

Doctors J.PD, A.T, K.M, S.E, O.S, R.S, M.A, B.D, R.C, and D.A. all contributed to the writing of the main manuscript. Dr D.A. and J.P.D prepared Figs. 1 and 2. All authors whose names appear on the submission 1) made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; 2) drafted the work or revised it critically for important intellectual content; 3) approved the version to be published; and 4) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

Dr. Adedinsewo's research is supported by the Mayo Clinic Building Interdisciplinary Research Careers in Women's Health (BIRCWH) Program funded by the National Institutes of Health (NIH) K12 HD065987. Dr. Carter is a Scientific Advisor for Anumana, Inc., and part of his effort related to this publication was supported by Grant Number UL1 TR002377 from the National Center for Advancing Translational Sciences (NCATS). Ronnie Sebro is the Deputy Editor for Radiology: Artificial Intelligence, Associate Editor for Journal of Digital Imaging, and BMC Musculoskeletal Disorders. All other authors have nothing to disclose.

Compliance with Ethical Standards

Conflict of interest
The authors declare no competing interests.

Human and animal rights and informed consent
This article does not contain any studies with human or animal subjects performed by any of the authors.

Disclaimer
The views expressed in this article are those of the authors and do not necessarily reflect the official views and policy of the NIH.

References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Rodgers CM, Ellingson SR, Chatterjee P. Open data and transparency in artificial intelligence and machine learning: a new era of research. F1000Res. 2023;12:387. https://doi.org/10.12688/f1000research.133019.1.

2. Li B, Gu J, Jiang W. Artificial intelligence (AI) chip technology review. In: 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). China: Taiyuan; 2019.
3. Mesko B, Gorog M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med. 2020;3:126. https://doi.org/10.1038/s41746-020-00333-z.
4. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25(1):70–4. https://doi.org/10.1038/s41591-018-0240-2.
5. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10. https://doi.org/10.1038/s41568-018-0016-5.
6.•• Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. Algorithmic bias playbook. Federal Trade Commission. 2021. Accessed 1 Aug 2023.
This article provides an in-depth and practical exploration of algorithmic bias, how to detect it, evaluate potential causes, and prevent bias with specific examples in healthcare.
7. Gentzel M. Biased face recognition technology used by government: a problem for liberal democracy. Philos Technol. 2021;34(4):1639–63. https://doi.org/10.1007/s13347-021-00478-z.
8. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of machine learning research. 2018. https://proceedings.mlr.press/v81/buolamwini18a.html.
9. Raji I, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. 2019. p. 429–35.
10. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153–63. https://doi.org/10.1089/big.2016.0047.
11. Flores AW. False Positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. Fed Probat. 2016;80:38. Accessed 20 Jul 2023.
12.• Norori N, Hu QY, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns. 2021;2(10):100347. https://doi.org/10.1016/j.patter.2021.100347.
This article highlights opportunities to use open science tools to address bias in machine learning and artificial intelligence for healthcare. This includes data/code sharing, inclusive algorithms, and participant centered algorithm development.
13.• Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123–44. https://doi.org/10.1146/annurev-biodatasci-092820-114757.
The authors describe ethical considerations at different stages of machine learning model development in healthcare and propose recommendations to address hidden challenges throughout the model development continuum.
14. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med. 2018;24(9):1304–5. https://doi.org/10.1038/s41591-018-0178-4.
15. Latif J, Xiao C, Imran A, Tu S. Medical imaging using machine learning and deep learning algorithms: a review. In: 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). Sukkur, Pakistan; 2019.
16. Bitterman DS, Aerts H, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health. 2020;2(9):e447–9. https://doi.org/10.1016/S2589-7500(20)30187-4.
17. Bai HX, Wang R, Xiong Z, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology. 2021;299(1):E225. https://doi.org/10.1148/radiol.2021219004.
18. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112: 103375. https://doi.org/10.1016/j.compbiomed.2019.103375.
19. Adedinsewo D, Carter RE, Attia Z, et al. Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. Circ Arrhythm Electrophysiol. 2020;13(8):e008437. https://doi.org/10.1161/CIRCEP.120.008437.
20. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med. 2021;27(5):815–9. https://doi.org/10.1038/s41591-021-01335-4.
21. Adedinsewo DA, Johnson PW, Douglass EJ, et al. Detecting cardiomyopathies in pregnancy and the postpartum period with an electrocardiogram-based deep learning model. Eur Heart J Digit Health. 2021;2(4):586–96. https://doi.org/10.1093/ehjdh/ztab078.
22. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861–7. https://doi.org/10.1016/S0140-6736(19)31721-0.

23. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. Lancet. 2022;400(10359):1206–12. https://doi.org/10.1016/S0140-6736(22)01637-3.
24. Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J Am Coll Cardiol. 2020;75(7):722–33. https://doi.org/10.1016/j.jacc.2019.12.030.
25. Grogan M, Lopez-Jimenez F, Cohen-Shelly M, et al. Artificial intelligence-enhanced electrocardiogram for the early detection of cardiac amyloidosis. Mayo Clin Proc. 2021;96(11):2768–78. https://doi.org/10.1016/j.mayocp.2021.04.023.
26. Cohen-Shelly M, Attia ZI, Friedman PA, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. Eur Heart J. 2021;42(30):2885–96. https://doi.org/10.1093/eurheartj/ehab153.
27. Elias P, Poterucha TJ, Rajaram V, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. J Am Coll Cardiol. 2022;80(6):613–26. https://doi.org/10.1016/j.jacc.2022.05.029.
28. Adedinsewo D, Hardway HD, Morales-Lara AC, et al. Non-invasive detection of cardiac allograft rejection among heart transplant recipients using an electrocardiogram based deep learning model. Eur Heart J Digit Health. 2023;4(2):71–80. https://doi.org/10.1093/ehjdh/ztad001.
29. Altaf F, Islam SM, Akhtar N, Janjua NK. Going deep in medical image analysis: concepts, methods, challenges, and future directions. IEEE Access. 2019;7:99540–72. https://doi.org/10.1109/ACCESS.2019.2929365.
30. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020;580(7802):252–6. https://doi.org/10.1038/s41586-020-2145-8.
31. He B, Kwan AC, Cho JH, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature. 2023;616(7957):520–4. https://doi.org/10.1038/s41586-023-05947-3.
32. Masud R, Al-Rei M, Lokker C. Computer-aided detection for breast cancer screening in clinical settings: scoping review. JMIR Med Inform. 2019;7(3):e12660. https://doi.org/10.2196/12660.
33. Cellina M, Ce M, Irmici G, et al. Artificial intelligence in emergency radiology: where are we going? Diagnostics (Basel). 2022;12(12). https://doi.org/10.3390/diagnostics12123223.
34. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056.
35. Wen D, Khan SM, Ji XuA, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health. 2022;4(1):e64–74. https://doi.org/10.1016/S2589-7500(21)00252-1.
36. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018;5:180161. https://doi.org/10.1038/sdata.2018.161.
37. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247–8. https://doi.org/10.1001/jamadermatol.2018.2348.
38. Dawes SM, Tsai S, Gittleman H, Barnholtz-Sloan JS, Bordeaux JS. Racial disparities in melanoma survival. J Am Acad Dermatol. 2016;75(5):983–91. https://doi.org/10.1016/j.jaad.2016.06.006.
39. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv. 2022;8(32):eabq6147. https://doi.org/10.1126/sciadv.abq6147.
40. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol. 2022;35(1):23–32. https://doi.org/10.1038/s41379-021-00919-2.
41. Hanna MG, Ardon O, Reuter VE, et al. Integrating digital pathology into clinical practice. Mod Pathol. 2022;35(2):152–64. https://doi.org/10.1038/s41379-021-00929-0.
42. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology. 2017;70(1):134–45. https://doi.org/10.1111/his.12993.
43. Rizzo PC, Caputo A, Maddalena E, et al. Digital pathology world tour. Digit Health. 2023;9:20552076231194550. https://doi.org/10.1177/20552076231194551.
44. Chorba JS, Shapiro AM, Le L, et al. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. J Am Heart Assoc. 2021;10(9):e019905. https://doi.org/10.1161/JAHA.120.019905.
45. Long Q, Ye X, Zhao Q. Artificial intelligence and automation in valvular heart diseases. Cardiol J. 2020;27(4):404–20. https://doi.org/10.5603/CJ.a2020.0087.
46. Grzywalski T, Piecuch M, Szajek M, et al. Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination. Eur J Pediatr. 2019;178(6):883–90. https://doi.org/10.1007/s00431-019-03363-2.
47. Jordi Laguarta FH, Brian Subirana. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol. 2020;1:275–81. https://doi.org/10.1109/OJEMB.2020.3026928.

48. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86. https://doi.org/10.1038/s41746-021-00455-y.
49. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J Biomed Inform. 2020;101:103337. https://doi.org/10.1016/j.jbi.2019.103337.
50. Artzi NS, Shilo S, Hadar E, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med. 2020;26(1):71–6. https://doi.org/10.1038/s41591-019-0724-8.
51. Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res. 2019;79(21):5463–70. https://doi.org/10.1158/0008-5472.CAN-19-0579.
52. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. 2019;14(12):3426–44. https://doi.org/10.1038/s41596-019-0227-6.
53. Uprety D, Zhu D, West HJ. ChatGPT-A promising generative AI tool and its implications for cancer care. Cancer. 2023;129(15):2284–9. https://doi.org/10.1002/cncr.34827.
54. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. https://doi.org/10.1001/jamainternmed.2023.1838.
55. Desaire H, Chua AE, Isom M, Jarosova R, Hua D. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. Cell Rep Phys Sci. 2023;4(6). https://doi.org/10.1016/j.xcrp.2023.101426.
56. Lingjiao Chen MZ, James Zou. How is ChatGPT's behavior changing over time? 2023. https://arxiv.org/pdf/2307.09009.pdf. Accessed 24 Aug 2023.
57. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy; 2017. Accessed 25 Aug 2023.
58. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544–7. https://doi.org/10.1001/jamainternmed.2018.3763.
59. Buolamwini J. Artificial intelligence has a problem with gender and racial bias. Here's how to solve it. Time. 2019.
60. Blueprint for an AI Bill of Rights. 2022. white house.gov.
61. Mujtaba DF, Mahapatra NR. Ethical considerations in AI-based recruitment. In: 2019 IEEE International Symposium on Technology and Society (ISTAS). Medford, MA, USA; 2019.
62. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA. 2020;324(14):1397–8. https://doi.org/10.1001/jama.2020.9371.
63. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health. 2020;2(5):e221–3. https://doi.org/10.1016/S2589-7500(20)30065-0.
64. White paper on artificial intelligence - a European approach to excellence and trust. 2020.
65. Joint statement of enforcement efforts against discrimination and bias in automated systems. 2023.
66. Commission E. White paper on artificial intelligence- a European approach to excellence and trust. 2020. https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 20 Jul 2023.
67. Protection NCaW. Automated employment decision tools (AEDT). 2021. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page#:~:text=Local%20Law%20144%20of%202021,audit%20is%20publicly%20available%2C%20and. Accessed 28 Jul 2023.
68. Joint statement: Bureau of Consumer Financial Protection DoJ, U.S. Equal Employment Opportunity Commission, and the Federal Trade Commission. Joint statement of enforcement efforts against discrimination and bias in automated systems. 2023. https://files.consumerfinance.gov/f/documents/cfpb_joint-statement-enforcement-against-discrimination-bias-automated-systems_2023-04.pdf. Accessed 30 Jul 2023.
69. UNESCO. Recommendation on the ethics of artificial intelligence. unesdoc. 2022. p. 43. https://unesdoc.unesco.org/ark:/48223/pf0000381137. Accessed 31 Jul 2023.
70. Administration UFaD. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed 30 Jul 2023.
71. Benjamens S, Dhunnoo P, Mesko B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3:118. https://doi.org/10.1038/s41746-020-00324-0.
72. WHO. Ethics and governance of artificial intelligence for health. World Health Organization. 2021. p. 150. https://www.who.int/publications/i/item/9789240029200. Accessed 30 Jul 2023.
73. Ethics and governance of artificial intelligence for health. World Health Organization. 2021 p.150.

74. Richards-Kortum R, Oden M. Engineering. Devices for low-resource health care Science. 2013;342(6162):1055–7. https://doi.org/10.1126/science.1243473.
75. Niezen G, Eslambolchilar P, Thimbleby H. Open-source hardware for medical devices. BMJ Innov. 2016;2(2):78–83. https://doi.org/10.1136/bmjinnov-2015-000080.
76. Castillo EG, Harris C. Directing research toward health equity: a health equity research impact assessment. J Gen Intern Med. 2021;36(9):2803–8. https://doi.org/10.1007/s11606-021-06789-3.
77. Minssen T, Gerke S, Aboy M, Price N, Cohen G. Regulatory responses to medical machine learning. J Law Biosci. 2020;7(1):lsaa002. https://doi.org/10.1093/jlb/lsaa002.
78. Health NIo. NIH launches Bridge2AI program to expand the use of artificial intelligence in biomedical and behavioral research. National Institutes of Health. 2022. https://www.nih.gov/news-events/news-releases/nih-launches-bridge2ai-program-expand-use-artificial-intelligence-biomedical-behavioral-research. Accessed 31 Jul 2023.
79. Health NIo. About ethics, bias, and transparency for people and machines. 2022. https://datascience.nih.gov/artificial-intelligence/initiatives/ethics-bias-and-transparency-for-people-and-machines. Accessed 31 Jul 2023.
80. Health NIo. Artificial intelligence/machine learning consortium to advance health equity and researcher diversity (AIM-AHEAD). NIH. 2023. https://datascience.nih.gov/artificial-intelligence/aim-ahead. Accessed 1 Aug 2023.
81. Disparities NIoMHaH. ScHARe (Science Collaborative for Health Disparities and Artificial intelligence bias REduction). 2023. https://www.nimhd.nih.gov/resources/schare/about-schare.html. Accessed 30 Jul 2023.
82. Science NCfAT. Bias detection tools in health care challenge. 2023. https://ncats.nih.gov/funding/challenges/bias-detection-tools-in-health-care. Accessed 28 Jul 2023.
83. Office NAII. National AI Initiative Act of 2020. Artificial Intelligence Office. 2020. https://www.ai.gov/. Accessed 30 Jul 2023.
84. Sengupta PP, Shrestha S, Berthon B, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. JACC Cardiovasc Imaging. 2020;13(9):2017–35. https://doi.org/10.1016/j.jcmg.2020.07.015.
85. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2023;14(7):8459–86. https://doi.org/10.1007/s12652-021-03612-z.

Publisher's Note