Mod Pathol 38 (2025) 100686

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Advancing Pathology Together

Journal homepage: https://modernpathology.org/

Review Article

Ethical and Bias Considerations in Artificial Intelligence/Machine Learning

Matthew G. Hanna[a,b,*], Liron Pantanowitz[a,b], Brian Jackson[c,d], Octavia Palmer[a,b], Shyam Visweswaran[e], Joshua Pantanowitz[f], Mustafa Deebajah[g], Hooman H. Rashidi[a,b,*]

[a] *Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania;* [b] *Computational Pathology and AI Center of Excellence (CPACE), University of Pittsburgh, Pittsburgh, Pennsylvania;* [c] *Department of Pathology, University of Utah, Salt Lake City, Utah;* [d] *ARUP Laboratories, Salt Lake City, Utah;* [e] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania;* [f] *University of Pittsburgh Medical School, Pittsburgh, Pennsylvania;* [g] *Department of Pathology, Cleveland Clinic, Cleveland, Ohio*

ARTICLE INFO

Article history:
Received 17 August 2024
Accepted 27 November 2024
Available online 16 December 2024

Keywords:
artificial intelligence
bias
computational pathology
ethics
machine learning
pathology

ABSTRACT

As artificial intelligence (AI) gains prominence in pathology and medicine, the ethical implications and potential biases within such integrated AI models will require careful scrutiny. Ethics and bias are important considerations in our practice settings, especially as an increased number of machine learning (ML) systems are being integrated within our various medical domains. Such ML-based systems have demonstrated remarkable capabilities in specified tasks such as, but not limited to, image recognition, natural language processing, and predictive analytics. However, the potential bias that may exist within such AI-ML models can also inadvertently lead to unfair and potentially detrimental outcomes. The source of bias within such ML models can be due to numerous factors but is typically categorized into 3 main buckets (data bias, development bias, and interaction bias). These could be due to the training data, algorithmic bias, feature engineering and selection issues, clinic and institutional bias (ie, practice variability), reporting bias, and temporal bias (ie, changes in technology, clinical practice, or disease patterns). Therefore, despite the potential of these AI-ML applications, their deployment in our day-to-day practice also raises noteworthy ethical concerns. To address ethics and bias in medicine, a comprehensive evaluation process is required, which will encompass all aspects of such systems, from model development through clinical deployment. Addressing these biases is crucial to ensure that AI-ML systems remain fair, transparent, and beneficial to all. This review will discuss the relevant ethical and bias considerations in AI-ML specifically within the pathology and medical domain.

Introduction

Artificial intelligence (AI) is becoming increasingly adopted across various domains, profoundly impacting societal sectors such as criminal sanctions,[1,2] loan offerings,[3] personnel hiring,[4] and health care.[5-7] The combination of enhanced computational capabilities and vast digital data sets has ushered in an unprecedented era of technological advancement. AI tools, including machine learning (ML) algorithms, are now pivotal in the health care domain. From the health care perspective, AI is particularly promising, with its potential to revolutionize diagnostics, treatment planning, and other aspects of patient care. Pathology and

* Corresponding authors.
E-mail addresses: hannamg3@upmc.edu (M.G. Hanna), rashidihh@upmc.edu (H.H. Rashidi).

https://doi.org/10.1016/j.modpat.2024.100686

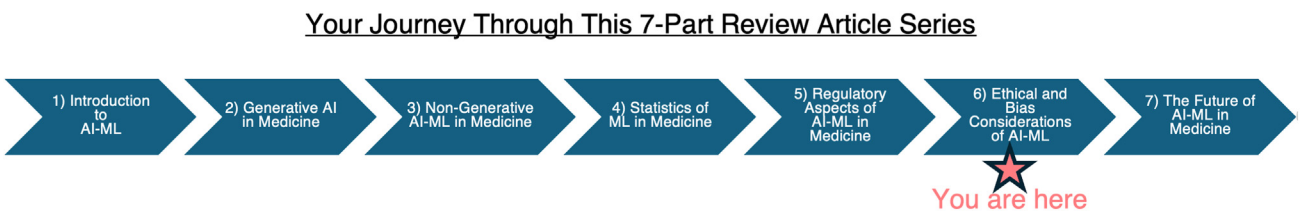

Figure 1.
Your journey through this 7-part review article series. AI, artificial intelligence; ML, machine learning.

laboratory medicine are at the forefront of integrating AI into practice and research, leveraging its capabilities to analyze digitized images, molecular data, laboratory text and tabular data, as well as pathology reports with remarkable accuracy and efficiency. Alongside these advancements come significant ethical considerations and concerns over potential biases inherent to AI models.

One such ethical concern revolves around the privacy and security of patient data that is used to train some of these AI systems. As these systems rely on vast amounts of personal health information, ensuring their robust data protection processes becomes an essential part of maintaining patient trust and complying with regulatory standards.[8] Moreover, biases in AI algorithms, whether due to the data sets used for training a ML model or the architecture of the algorithms themselves, can lead to potential inequities in certain health care delivery settings. Such challenges will need to be addressed, but overcoming those hurdles is no small task and may require a multifaceted approach. It is also important to note that medical practitioners (especially pathologists and laboratory medicine professionals) are data stewards as well and must accordingly play a pivotal role in guiding the ethical development and deployment of our current and future AI-ML technologies. This includes ensuring the development of an environment that prioritizes transparency in the AI lifecycle, addresses biases in AI algorithms to ensure fair representation, conducts rigorous validation studies to assess performance across diverse populations, advocates for the privacy of patient health care data, safeguards informed consent processes, ensuring accountability when something goes wrong, and actively engages in regulatory frameworks that promote fairness and accountability. Moreover, concern exists regarding the extent to which AI systems should be permitted to make autonomous decisions, and whether humans should remain in the loop. Moral and ethical concerns regarding AI also involve the potential impact of this technology on employment (eg, disparate economic and social disruption), cultural values (eg, stifling originality), and the environment (eg, carbon footprint due to large amounts of computational power).

Although AI holds tremendous promise for enhancing pathology and medicine, it must also be implemented responsibly to mitigate potential ethical pitfalls and biases. By embracing ethical principles and fostering collaboration across disciplines, AI can serve as a force for positive change in health care while upholding the values of fairness, transparency, and patient-centered care (Fig. 1).

Ethical Artificial Intelligence

Ethics has been a foundation of medicine at least since the days of Hippocrates. A widely used modern formulation of medical ethics can be found in the Belmont Report. Key principles in medical ethics discussed in the Belmont Report include autonomy, beneficence, nonmaleficence, and justice (Table 1).[9,10]

AI ethics can be defined as "a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies."[11] AI ethics are necessary because a variety of harms to individuals and society might result from the misuse, abuse, bad design, or negative unintended consequences of AI systems. In the context of AI, certain medical ethical principles related to health care and medical research can be extended to the development and use of AI systems within health care. These ethical principles collectively form the guiding framework for ethical decision-making in health care and the ethical development of technologies across scientific and computational domains. A deeper dive into the key principles (autonomy, beneficence, nonmaleficence, and justice) is necessary to better understand their guiding framework.

Table 1
Ethical principles of health care and medical research with extensions to medical artificial intelligence (AI)

Ethical principle	Health care and medical research	Medical AI
Respect for Autonomy (the principle of self-governance)	Physicians and researchers respect an individual's right to make their own decisions regarding their health care and participation in research	AI developers and users ensure that individuals have sufficient control over their interactions with AI
Beneficence (the principle of doing good)	Physicians and researchers act for the benefit of the individual and protect and defend their rights	AI developers and users are responsible for maximizing human benefits
Nonmaleficence (the principle of avoiding harm)	Physicians and researchers do not harm the individual intentionally, negligently, or unintentionally	AI developers and users are responsible for preventing harm and mitigating risks
Justice (the principle of fairness)	Physicians and researchers treat the individual fairly and equitably, regardless of factors such as race, gender, socioeconomic status, or medical condition	AI developers and users are responsible for promoting equity, regardless of factors such as race, gender, socioeconomic status, or medical condition
Accountability	Physicians and researchers take responsibility for their activities	AI developers and users are accountable for ensuring that AI is designed, implemented, and operated ethically, transparently, and reliably

Autonomy

"Respect for persons," which includes autonomy, pertains to an individual's right to make decisions concerning their physical body and derivates thereof (eg, tissue specimens), and their personal health information. Autonomy is typically upheld through the practice of informed consent. In the context of AI, autonomy must be considered for the use of personal health data for developing, training, and validating AI systems, and the application of AI systems in patient care. AI applications in medicine often use data sets that are deidentified to comply with Health Insurance Portability and Accountability Act regulations, but this may not fully meet ethical standards for patient autonomy.[12] Many individuals expect transparency and control over their identified and even deidentified data. Advocating for a means to opt out at will has been demonstrated in the European Union with their "right to erasure (be forgotten)." This could help balance patient autonomy and AI development.[13] Furthermore, as AI technologies become more prevalent in health care, concerns surrounding patient awareness and consent for AI-assisted diagnostics may also arise. Patients may want to understand the extent to which an AI algorithm contributed to the medical interpretation of their patient test results. Ethical and legal clarity is essential to determine whether and when the disclosure of AI involvement in medical decision-making is required.[14-16]

Beneficence and Nonmaleficence

In order to satisfy beneficence (performing a beneficial deed) and nonmaleficence (refraining from doing harm) requirements, our AI applications must also demonstrate that the potential benefits gained from their use markedly outweigh their potential risks (ie, drastically minimizing any potential harm associated with their use). This includes considering benefits and risks not only for patients directly impacted by AI decisions, whether they are physician-assisted or autonomous, but also for those patients whose samples or medical data contributed to AI system development. Additionally, due to the potential for bias in AI systems, it is also important to continuously evaluate their performance characteristics by their ongoing monitoring across diverse patient and diagnostic cohorts, which will further ensure that the benefits gained outweigh the risks.

Justice

Justice in relation to AI in health care demands equitable distribution of costs, risks, and benefits across diverse populations throughout the development, deployment, and potential commercialization processes. Fairness extends to data collection for training AI systems, ensuring that historically marginalized groups are not disadvantaged. Concerns such as exploitation and lack of consent arise when data, such as genetic information, may be used without permission. Clinical use of AI systems should demonstrate equitable performance across a spectrum of demographic groups to ensure performant generalizability, including subgroup analyses to detect and mitigate biases in minority cohorts. Additionally, ensuring that AI systems are designed and trained with considerations for diverse global backgrounds is also essential to harness their potential benefits in underserved regions with limited health care resources and expertise.[17]

Another justice consideration is the perceived fairness of profiting from other people's personal data or professional work products. Social science research has shown that although most patients are willing for their personal health data to be used for academic research, people are much less open to the use of their data for commercial purposes.[18,19] The Henrietta Lacks case demonstrated that some patients expect to share in financial benefit that derives from the use of their tissues, and health data are arguably at least as personal as body tissues.[20] Medical professionals, such as pathologists and radiologists, may likewise consider it unjust if other people profit off of AI models that are based on their specific expertise, for example, trained on their previous pathology or radiology reports. These ongoing controversies are not only confined to medical disciplines alone but also a point of contention in other fields such as screenwriting and acting, as well as journalism.[21-23] Besides the aforementioned, another important concept to discuss which is not explicitly mentioned as one of the core principles in the Belmont Report is "accountability."

Accountability

Accountability in medical AI ethics requires both individual and organizational adherence to rigorous standards. Individuals, including clinicians and ML model developers, must uphold professional codes of ethics tailored to AI applications, ensuring transparency, bias mitigation, and patient privacy. Organizations should establish robust policies and transparent practices that enforce ethical behavior internally and in partnerships, promoting stakeholders to enhance trust and sustainability.[24,25] Regulatory bodies need to play a pivotal role in overseeing compliance with ethical standards, ensuring that AI technologies in health care prioritize patient welfare and align with evolving regulatory frameworks, although the regulation of medical AI is not fully developed (for more details, please refer to article 5 in this series: Regulatory Aspects of AI-ML).[26] Together, these mechanisms aim to foster responsible innovation, mitigate risks, and uphold ethical standards across the health care sector.

Bias in Artificial Intelligence Systems

The absence of bias, or equivalently the presence of fairness, is critical to ensuring that AI systems operate ethically and equitably. In data science, bias refers to any systematic error or deviation from the true value in data collection, preparation, or analysis. This can arise due to various factors such as incomplete data, skewed sampling methods, or errors in data recording. When AI models are trained on biased data, they can inherit and perpetuate these inaccuracies, leading to biased outcomes and medical decisions.[27] Bias in AI refers to systematic and unfair favoritism or prejudice in AI systems, which can lead to discriminatory outcomes. Three broad factors are responsible for biases in AI models: (1) data bias, which is the use of unrepresentative data; (2) development bias, which is the result of the inappropriate use of AI algorithms in model development; and (3) interaction bias, which is the result of improper user interactions with the model.

Data Bias

The reliance on data to train algorithms introduces an omnipresent challenge, as data itself can reflect biases inherent in

societal structures (eg, racism, prejudice, and classism), historical patterns of biases that mediate the practice of medicine and delivery of health care.[28,29] Bias in AI models often lies in the data sets used during the training phase. These data sets may inadvertently encode societal biases, reflecting historical inequalities, or systemic injustices present in the data collection process. Diagnostic algorithms trained on historical patient data in electronic health records may disproportionately include certain demographic groups (eg, gender, race, socioeconomic status, religion, or disability) over others due to disparities in access to health care or differences in how certain conditions are diagnosed across populations. Bias can also manifest in AI models through the design and implementation of algorithms themselves. The choices made in algorithmic design, such as feature selection or model weights, can inadvertently reinforce biases present in the training data. This phenomenon, known as algorithmic bias, can perpetuate and even exacerbate existing disparities when deployed in real-world applications. There are numerous types of biases that can exist within the context of ML, which in turn apply to pathology and medicine (Table 2).

Bias with respect to medical data can stem from various sources. Missing data are a significant issue, as patients may have patient visitations or testing at multiple sites, and their medical data are fragmented across various health records. Additionally, patients with lower health or information technology literacy might not fully engage with patient portals or accurately report outcomes, further leading to deficient records. This can skew data toward more severe cases or certain demographics, influencing clinical decision support systems to make inaccurate predictions or recommendations that do not generalize well across diverse patient groups.[30,31] Standardized reporting in medicine can also facilitate bias mitigation; however, there are known data mismatches for similar testing across different institutions. An example within laboratory medicine is showcased where Logical Observation Identifiers, Names, and Codes for the same laboratory test have been shown to be incorrect at various institutions in about 20% of coagulation and cardiac tests.[32,33] Sample size disparities among patient subgroups within population data sets can also limit the effectiveness of predictive analytics and decision support tools for those groups, potentially neglecting their health care requirements compared with more represented populations. Misclassification or measurement errors further compound biases in medical data. Patients from varying socioeconomic backgrounds, who receive care in nonacademic clinics where data input and clinical reasoning may differ, may also become more particularly vulnerable to such inaccuracies. These discrepancies can perpetuate such disparities because the biased data within these training algorithms will inadequately address the health care needs of such populations. Implicit biases among health care providers can also influence clinical decision support outputs. Addressing these sources of bias in medical data are crucial for developing equitable health care practices. Therefore, the system must employ robust strategies for data collection, validation, interpretability, and ongoing monitoring to ensure that AI-driven clinical decision support systems effectively serve all patient populations.[34]

Development Bias

Addressing bias in AI models requires a concerted effort to acknowledge and mitigate potential unfairness with the training data, validation, model development, and deployment. There should be a rigorous evaluation of training data sets used to train a ML model. For instance, a distributed health network could preferentially select patient samples to develop a ML model from academic hospitals alone, unintentionally leaving out patient samples from rural community hospitals from the training data set. Bias of patient populations who have access to care in metropolitan city centers, where academic hospitals are located,

Table 2
Representative types of bias and related examples in pathology

Type of bias	Description	Examples in pathology
Data bias	Bias in training data due to underrepresentation, overrepresentation, or misrepresentation of pathology cases.	Overrepresentation of certain demographics in diagnostic data sets.
Algorithmic bias	Bias introduced during algorithm design affecting diagnostic accuracy or treatment recommendations.	Skewed prioritization of symptoms or conditions in diagnostic algorithms.
Sampling bias	Bias from nonrandom sampling methods in pathology data, leading to skewed conclusions.	Insufficient gross dissection for mapping of tumor bed to predict response to therapy.
Measurement bias	Bias due to inaccuracies in diagnostic tests or imaging technologies, affecting treatment decisions.	Variations in test sensitivity or reference ranges across demographic groups.
Labeling bias	Bias in the assignment of disease labels or classifications, influencing disease prediction models.	Subjective interpretation of biopsy tumor grading results leading to misclassification of disease/severity.
Prejudice bias	Bias from preconceived notions about certain diseases or patient groups influencing pathology diagnoses.	Stereotypical assumptions about patient demographics influencing diagnosis or management decisions.
Environmental bias	Bias from environmental factors affecting disease prevalence or diagnostic outcomes in certain regions.	Patterns in patient testing in academic hospitals in urban regions compared to community practices in rural settings.
Interaction bias	Bias from complex interactions between different diseases or co-morbid conditions affecting diagnoses.	Serum antibody chemistry interferences in patients with similar diseases influencing test results.
Feedback loop bias	Bias exacerbated by diagnostic feedback loops where historical data or previous diagnoses influence future diagnostic decisions.	Predicting future ancillary studies based on initial diagnoses without supporting evidence.
Representation bias	Bias from inadequate representation of diverse populations in diagnostic data sets, affecting accuracy.	Underrepresentation of minority populations in genetic screening databases.
Temporal bias	Bias due to changes in disease prevalence or diagnostic criteria not reflected in historical diagnostic data.	Evolution of disease classifications leading to discrepancies in historical and current diagnoses.
Transfer bias	Bias from differences between diagnostic practices in training hospitals and those in community settings.	Variations in diagnostic criteria between academic research and community health centers.
Confirmation bias	Bias where initial beliefs or diagnostic decisions influence subsequent interpretations or actions.	Favoring evidence that confirms initial diagnoses or treatment plans without considering contradictory information.

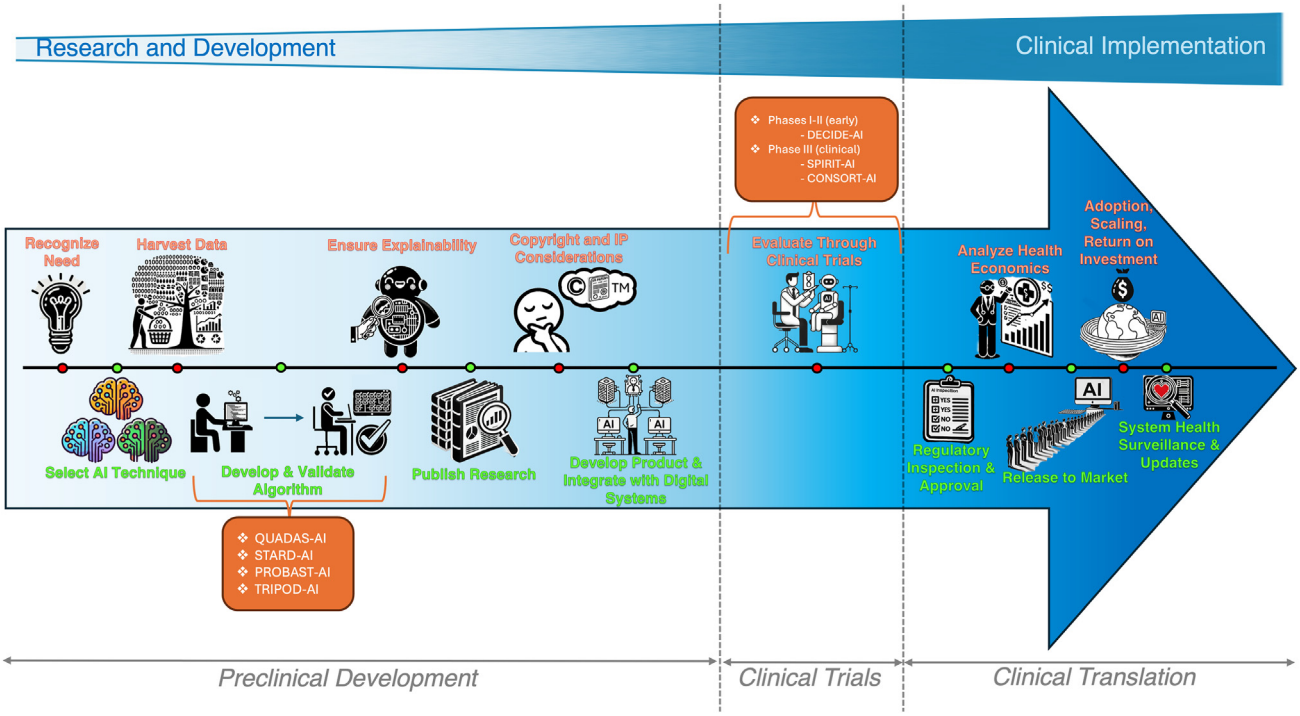

Figure 2.
Guidelines and respective stages of AI in the medical lifecycle. STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies - Artificial Intelligence) provides guidelines for reporting studies that evaluate diagnostic accuracy using AI technologies. It emphasizes transparent reporting of study methods, including participant selection criteria, reference standard procedures, AI algorithm details, and diagnostic performance metrics. TRIPOD-AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Artificial Intelligence) provides guidelines for reporting studies developing and validating AI-driven diagnostic or prognostic models. MINIMAR (MINimum Information for Medical AI Reporting) aims to develop a minimal reporting standard for AI studies in medicine, focusing on essential elements necessary for reproducibility and evaluation. CLAIM (Checklist for Artificial Intelligence in Medical Imaging) and MI-CLAIM (Minimum Information about Clinical Artificial Intelligence Modelling) provide checklists for developing and validating AI models in medical imaging and clinical applications, respectively. MI-CLAIM (Minimum Information for Clinical AI Model) provides guidelines for reporting clinical AI models, focusing on transparency and reproducibility in model development and validation. It emphasizes comprehensive reporting of model inputs, outputs, performance metrics, validation methods, and potential limitations. MI-CLAIM aims to enhance the reliability and interpretability of clinical AI models across health care domains. CONSORT-AI (Consolidated Standards of Reporting Trials for Artificial Intelligence) aims to improve the transparency and quality of reporting in clinical trials that involve AI interventions. It provides guidelines for reporting essential elements such as study design, participant characteristics, intervention details, and outcomes. By standardizing reporting practices, CONSORT-AI facilitates the critical evaluation and replication of AI-driven clinical trials. SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) focuses on enhancing the completeness and transparency of protocols in AI-related clinical trials. It provides recommendations for reporting key elements of trial protocols, including eligibility criteria, randomization procedures, statistical analysis plans, and data handling processes. SPIRIT-AI aims to improve the reliability and reproducibility of AI trial results. DECIDE-AI (Development of the Consolidated Criteria for Reporting Qualitative Research-Artificial Intelligence) provides guidelines for reporting qualitative research studies that involve AI applications. It emphasizes transparency in research methods, data collection, and analysis techniques used in qualitative AI studies. DECIDE-AI aims to ensure that qualitative findings related to AI development and implementation are rigorously documented and communicated. PROBAST-AI (Prediction model Risk of Bias Assessment Tool-Artificial Intelligence) provides guidelines for assessing the risk of bias in prediction models developed using AI techniques. It offers structured criteria for evaluating key domains of bias, including participant selection, predictor variables, model development, and model evaluation. PROBAST-AI aims to improve the robustness and reliability of prediction models by identifying and addressing potential sources of bias. QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence) focuses on assessing the quality of studies that evaluate the diagnostic accuracy of AI technologies. It provides criteria for evaluating study design, participant selection, reference standard procedures, AI algorithm performance, and data analysis methods. QUADAS-AI aims to enhance the reliability and validity of diagnostic accuracy studies involving AI applications. AI, artificial intelligence.

compared with rural-based patients, must be identified to mitigate before they propagate into AI systems. This involves diversifying data sets to ensure representation across different demographic groups and socioeconomic backgrounds, thereby minimizing the risk of underrepresentation or misrepresentation. Additionally, transparency in AI development is crucial. Developers and researchers should document the data sources used, subgroup analytics (eg, race vs ancestry), and the methodologies involved in algorithm design. This transparency enables external scrutiny and fosters accountability, helping identify and rectify biases before clinical deployment. The “Garbage-In, Garbage-Out” principle highlights that the quality of AI outputs is directly dependent on the quality of the input training data. If the data used to train an AI model are biased or flawed, the resulting model will likely exhibit similar biases.[35] Bias in medical data, documented in radiology and other medical domains, has been relatively limited in the pathology literature.[36-39] Bias in medical AI underscores the importance of using high-quality, representative data in the AI development lifecycle, which exemplifies the spectrum of patient populations and disease characteristics that will be used in real-world practice (Fig. 2).

Bias in AI models can manifest in various ways. These models are crucial for various health care applications but may inherit biases from the data sets that they are trained on, such as The Cancer Genome Atlas. Recently, researchers demonstrated that distinct features in whole slide pathology images, derived from pooled digital slides contributed by over 140 hospitals, could be attributed back to their respective submitting institutions.[40] The study selected 8,579 digital slides from The Cancer Genome Atlas, contributed by over 140 medical institutions, and employed DenseNet121 and KimiaNet for feature extraction and cancer type classification. The results showed that features could identify

acquisition sites with significant accuracy (70% for DenseNet121 and over 86% for KimiaNet), indicating the presence of institution-specific patterns. These patterns, although might be less relevant to medical diagnosis, may influence model outcomes such as cancer subtype classification and image search. The study underscores the need for researchers to acknowledge and mitigate biases originating from factors such as whole slide scanner configurations, staining variations, and patient demographics, especially when developing and training deep learning models in digital pathology.

Furthermore, Leo et al[41] highlight the importance of quantitative histomorphometry in evaluating various data sets and downstream models. Quantitative histomorphometry extracts computerized features from digitized tissue slide images to predict disease presence and outcomes. However, variability in laboratory-specific factors such as staining reagents, tissue slice thickness, diagnostic metadata, and slide scanners can compromise feature stability between different laboratories. Their article introduces a preparation-induced instability score and latent instability score to quantify feature variability across and within data sets. Traditional performance metrics such as accuracy and area under the receiver operating characteristic curve are commonly used in feature selection to enhance class differentiation. However, when developing a reliable classifier, it is crucial that the feature selection accounts for both the discriminative power and stability. Feature stability can be shown graphically where its distribution's mean and shape remain consistent across various patient cohorts sharing similar diseases, clinical profiles, or outcomes. The authors continued to show that a prostate cancer detection model performed variably across different data sets from distinct institutions and furthermore across 3 different vendor scanners, with and without color normalization.[41] Furthermore, scanning the same 34 slides on 3 different scanners revealed that Haralick features were notably affected by the scanner model, showing instability in 62% of comparisons. Importantly, feature families that exhibited instability performed significantly worse in classification across different sites compared with within-site classification. These findings underscore the importance of evaluating bias in quantitative histomorphometric features across multiple sites and varying data sets to assess their robustness.

Similarly, if training data used to develop a ML model predominantly include data (eg, images, laboratory results) from a specific demographic group, the model may perform poorly from underrepresented groups. This can lead to disparities in diagnostic accuracy and treatment recommendations for various segments of the population.[42] Ongoing research into techniques such as fairness-aware ML models aims to develop algorithms that explicitly mitigate biases and promote fairness across diverse populations. These techniques incorporate principles of fairness, accountability, and transparency into AI development, ensuring that the benefits of AI technologies are equitably distributed and do not perpetuate societal inequalities. The consequences of biased AI extend beyond individual patient misdiagnoses. Systemic biases in AI can perpetuate existing inequalities and disproportionately affect marginalized communities. Addressing these biases is crucial to prevent harm and promote social justice.

Interaction Bias

Interaction biases may arise when health care providers or patients engage with AI systems in ways that impact the system's performance and impartiality. Automation bias refers to a type of clinician-interaction bias where doctors are not aware that the AI system is less reliable for a particular population. As a result, they trust the system too much and accept inaccurate advice. Privilege bias refers to a form of patient-interaction prejudice that arises when AI systems are not accessible in health care settings where marginalized groups receive treatment, leading to an uneven distribution of benefits provided by AI systems.

Implicit biases among health care providers (eg, differential care, incomplete medical history, limited medical treatments, and specialty referrals) can significantly impact various aspects of health care. In health care delivery, it affects patient-provider communication by influencing how information is exchanged and understood. This, in turn, influences patient-provider relationships, patient satisfaction levels, and perceptions of a physician's patient-centeredness. Implicit bias can also affect patient treatment adherence, as providers' decisions may be influenced by biased perceptions of a patient's likelihood to follow prescribed treatments. In the realm of public health, implicit bias impacts resource allocation decisions, such as the location of testing facilities, distribution of vaccines, and placement of environmental stressors. These decisions can have profound implications for community health outcomes and equity. Within health professions workplaces and medical education, implicit bias manifests in practices related to promotions, compensation, evaluations, awards, and research grants. It contributes to burnout and isolation among health care professionals and impacts the diversity of trainees and the workforce. Biases in recruitment and selection processes can hinder the creation of inclusive learning environments and perpetuate inequalities within the health care workforce.[43] Overall, addressing implicit bias in health care is crucial for promoting equity, improving patient outcomes, fostering inclusive environments, and ensuring fair resource distribution in public health initiatives.

Bias in Medical Artificial Intelligence

In the realm of medical AI, imbalanced population data pose a significant challenge, often resulting in biased algorithms that disproportionately impact historically marginalized communities. The underrepresentation of certain demographic groups within training data sets engenders algorithmic biases, leading to erroneous diagnoses and exacerbating health care disparities. If the training data set predominantly consists of data from a specific demographic or diagnostic group (eg, representation bias), the AI model may learn patterns and features that are specific to that group. As a result, when the model is deployed in diverse populations, it may not perform as well (eg, lack of generalizability) or could exhibit biased predictions as it would not have adequately learned about the underrepresented groups. In medical applications, diagnostic AI models trained on imbalanced data might disproportionately focus on conditions that are prevalent in the majority group represented in the data set. This can lead to underdiagnosis or misdiagnosis of conditions that are less prevalent or differentially presented in minority groups. For example, in prostate cancer, Gleason pattern 3 is the most common ($3 + 3 = 6$), whereas Gleason pattern 5 is the least prevalent. However, Gleason pattern 5 has significant prognostic implications for patients. If models are trained on the typical incidence of prostate cancers, Gleason pattern 5 may be underrepresented, and there may be a lack of performance in detecting Gleason pattern 5, resulting in under-grading the patient's prostate cancer. If undetected, this would affect patient

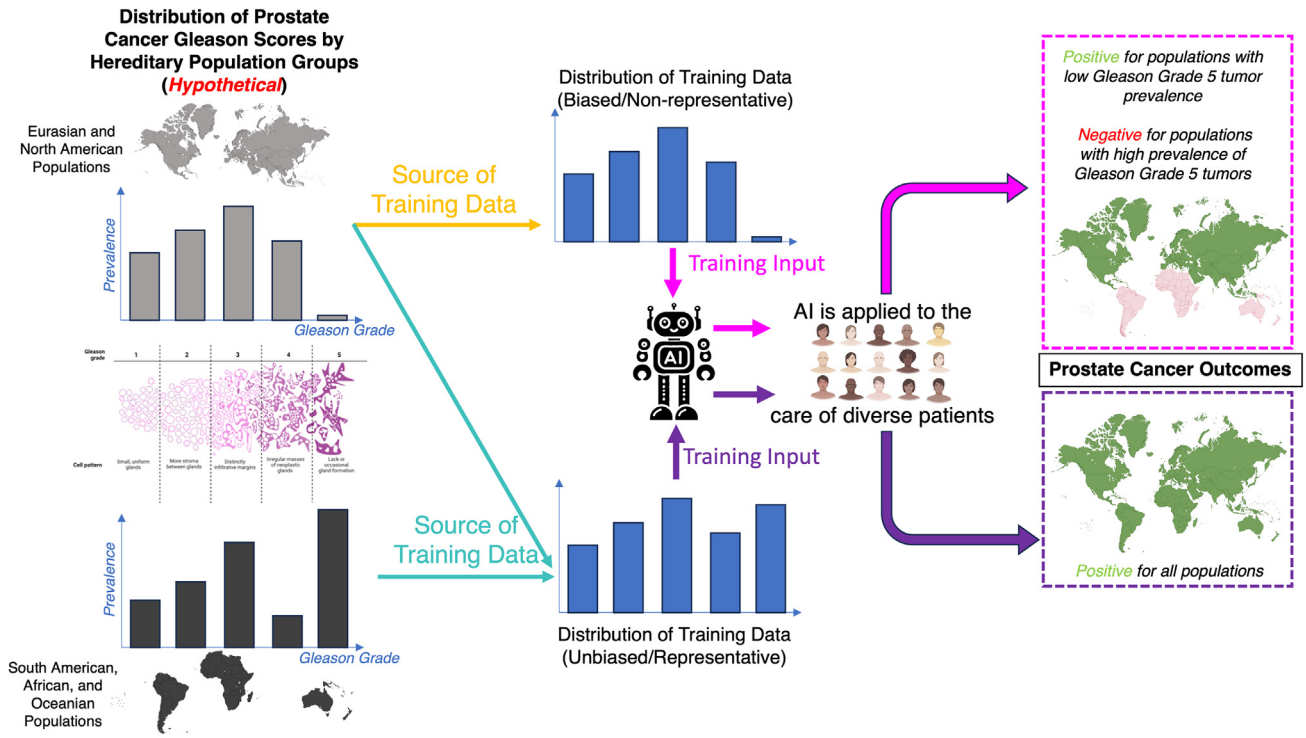

Figure 3.
Representative example of bias introduced into developing an AI algorithm for prostate cancer. AI, artificial intelligence.

management strategies, prognosis, and expectations of disease course.[44,45] Evidence also suggests that prostate cancer survival varies by race, even if the incidence of Gleason patterns are similar.[46] Furthermore, a ML model intended to grade prostate adenocarcinoma could appear to have excellent performance on a normal distribution of patients, as the validation cases are naturally imbalanced and may not include samples representing the higher Gleason pattern. Imbalanced data sets can perpetuate algorithmic biases, where the AI system systematically favors certain groups over others in its predictions or recommendations. This bias can arise due to skewed sampling of data, incorrect assumptions about causality, or indirect correlations that reflect societal or medical biases present in the data collection process. The core concern related to the training of models on imbalanced data are the lack of generalizability when applied to diverse populations. The performance characteristics from such models may not accurately reflect their performance across all demographic groups, leading to unreliable outcomes, model distrust, and potentially harmful decisions in clinical settings (Fig. 3).

Furthermore, biased AI systems can exacerbate disparities in health care outcomes by reinforcing existing inequalities. Vyas et al[47] discuss several ML models in health care that include race and other laboratory testing parameters as inputs. Race, a social construct, is used as a proxy for biology in many ML models although the contribution of social determinants of health and the patient's symptoms may be second to race in the decision-making process.[48] In nephrology, a prime example in laboratory testing is a calculation of the estimated glomerular filtration rate (eGFR), which includes race adjustment. These algorithms result in higher reported eGFR values for African Americans, which suggests better kidney function, and may delay patient care or renal transplantation and lead to worse outcomes.[49-51] Race is also included in many other medical tools predicting patient outcomes, including, but not limited to, The American Heart Association's Guidelines for Heart Failure,[52] The Society of Thoracic Surgeons Short Term Risk Calculator,[53] eGFR, modification of diet in renal disease, and chronic kidney disease epidemiology collaboration equations,[50] Organ Procurement and Transplantation Network: Kidney Donor Risk Index,[54] vaginal birth after cesarean risk calculator,[55] urology kidney stone risk score,[56,57] urinary tract infection risk calculator,[58] rectal cancer survival calculator,[59] National Cancer Institute Breast Cancer Risk Assessment Tool,[60] Breast Cancer Surveillance Consortium Risk Calculator,[61] osteoporosis risk score, fracture risk assessment tool,[62] and pulmonary-function tests.[63] Models may prioritize symptoms or indicators that are more prevalent or easier to detect in the majority group, potentially overlooking symptoms that manifest differently or less frequently in minority groups. This can result in underdiagnosis or misdiagnosis of conditions in underrepresented populations.[64] Patients from underrepresented groups may receive suboptimal treatments or experience adverse outcomes due to algorithms that do not account for their specific medical needs or variations. They can also undermine trust in AI technologies and health care systems if they consistently fail to provide fair and accurate assessments across all patient groups. This underscores the need for inclusive and equitable AI development practices and mitigation of such biases.

Bias Mitigation

Mitigating bias in medical AI necessitates a multidisciplinary approach to mitigate and prevent bias in each phase of the AI developmental lifecycle, which includes problem formulation; data selection, assessment, and management; model development, training, and validation; deployment and integration of models in intended settings; and monitoring, maintenance,

updating, or deimplementation.[65] Embracing diversity in data collection processes, ensuring the representation of historically marginalized communities, and fostering collaborative partnerships between stakeholders are pivotal in advancing equitable health care solutions. To mitigate biases in health care models, it is crucial to address imbalances in training data by the various ML model life cycle stages:

Data Collection and Preparation

Bias mitigation in ML is aimed at ensuring fairness and equity in AI systems, particularly as they increasingly influence decision-making. During data collection and preparation, the focus is on gathering diverse data sets that accurately represent the population and identifying as well as mitigating biases that may already exist in the data. Actively seeking out and including diverse data sets that represent different demographics, geographic locations, and socioeconomic backgrounds is crucial. Implementing techniques such as oversampling of underrepresented classes, under-sampling of overrepresented classes, or generating synthetic data can help balance class distributions in data sets. Additionally, data augmentation, anonymization, and careful handling of missing data help reduce the risk of perpetuating biases during model training. Rigorous exploratory data analysis is also crucial, as it uncovers hidden biases and informs strategies for data preprocessing.

Model Development and Training

In model development, selecting appropriate algorithms and features plays a pivotal role. It is essential to consider fairness and conduct thorough evaluations across different demographic groups to identify any disparate impacts. Techniques such as fairness-aware learning and bias detection algorithms can help mitigate biases that may arise from model selection and training.

Model Evaluation

Performance evaluation to assess model performance across the spectrum of demographic and diagnostic groups helps obviate bias. Bias mitigation in the model evaluation stage should incorporate more than traditional accuracy metrics and include bias and fairness assessments. Evaluating the model's performance across various subgroups (eg, patient populations, diagnostic categories, staining protocols, reference ranges, etc) ensures that predictions are equitable and unbiased. This stage allows for adjustments and refinements to the model to mitigate any identified biases before deployment.

Model Deployment and Ongoing Monitoring

When deploying AI systems, clinical verification and validation as well as ongoing continuous monitoring and maintenance are needed. Regular bias assessment throughout the AI model lifecycle helps to identify and mitigate biases as they arise, including preprocessing steps, algorithmic adjustments, and post-deployment monitoring. Monitoring for biases in real-time and establishing mechanisms for feedback and updates enable ongoing improvements. This includes implementing protocols to address bias-related events promptly and transparently.

Interpretability and Accountability

Developing ML models that provide explainability and self-reflection methods for their predictions enables health care professionals to understand how decisions are made and enhances our abilities to identify potential biases. Such explainability and self-reflection can help enhance transparency and the trustworthiness of our AI systems. Self-reflection (ie, introspection) specifically allows AI models to analyze their own decision-making processes by identifying the biases and factors that influence their outcomes. Explainability, on the other hand, ensures that AI decisions are understandable to stakeholders and end users by providing clear justifications for their prediction outputs or actions. In addition to the above, human collaborations among different teams and stakeholders can also enhance our perspectives and help uncover certain biases that may have been overlooked.

Ultimately, effective bias mitigation requires a holistic approach that incorporates both technical and ethical considerations. By embedding fairness as a core principle from the outset and integrating it into each stage of the ML model development and deployment life cycle, organizations can build AI systems whose contributions minimize harm by enhancing positive impacts and promoting inclusivity. Such an approach in addressing bias in ML models is crucial for ensuring equitable health care outcomes while simultaneously building trust in such AI-driven health care systems. Such a process undoubtedly requires a collaborative effort from health care providers, data scientists, and policymakers whose creative frameworks will prioritize fairness, accountability, and patient well-being.

Findability, Accessibility, Interoperability, and Reusability Principles

The findability, accessibility, interoperability, and reusability (FAIR) guiding principles were developed in the context of scientific data sets to ensure the establishment and enforcement of data stewardship and governance for data management. These principles aim to enhance the quality and utility of data, which can help mitigate bias in AI models.[66] FAIR principles are increasingly being applied to AI models to enhance their usability and integration within various clinical and research applications. The principles help ensure that AI models are managed in a way that promotes discoverability, accessibility, interoperability, and reusability (Fig. 4).

Importance of Findability, Accessibility, Interoperability, and Reusability Principles in Managing Bias in Artificial Intelligence Models

The FAIR principles play a crucial role in managing bias in AI models by fostering transparency, reproducibility, and inclusivity throughout the data lifecycle. *Findability* ensures that diverse and representative data sets are readily discoverable. By making data sources transparent and accessible, researchers can identify and mitigate biases inherent in the data set collection process. This transparency allows for a critical examination of data sources, helping to identify any biases that may have been unintentionally incorporated. *Accessibility* mandates that data should be accessible to a wide range of stakeholders under clear conditions. This principle ensures that biases are not perpetuated through restricted access to data or through opaque data governance

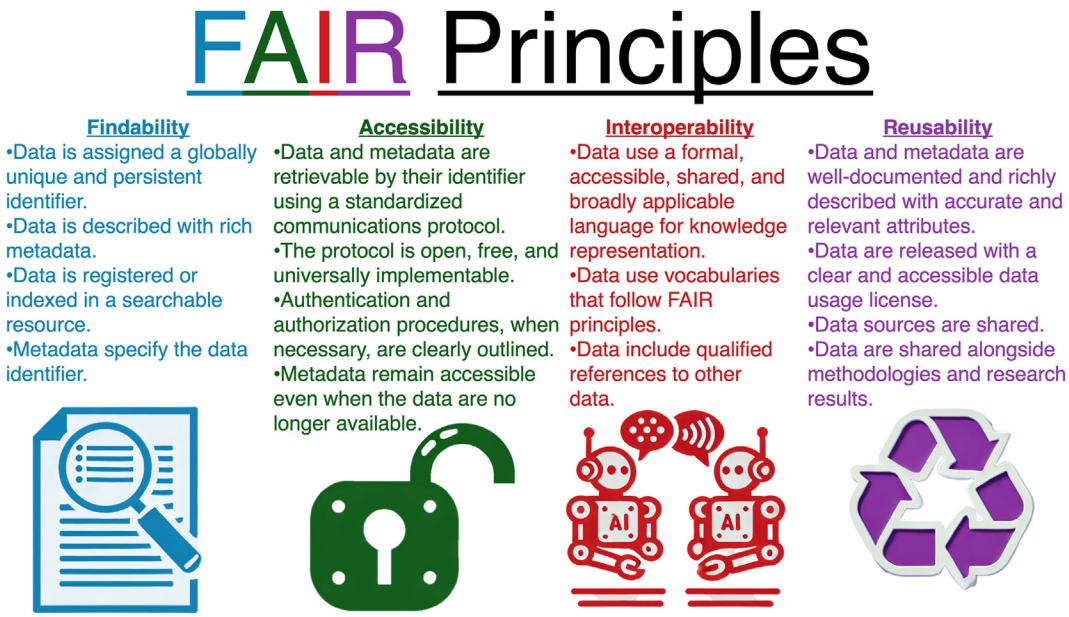

Figure 4.
FAIR principles. FAIR, findability, accessibility, interoperability, and reusability.

practices. By promoting open access and clear protocols for data sharing, the FAIR principles enable researchers to scrutinize data sets and algorithms for bias and discriminatory patterns. *Interoperability* ensures that data can be integrated and analyzed across different platforms and applications. Standardized formats and metadata facilitate the comparison of results from different AI models and data sets. This interoperability allows researchers to assess whether biases identified in 1 data set may also be present in others. *Reusability* encourages the documentation and sharing of data sources, methodologies, and results. By providing clear guidelines for data reuse and replication, the FAIR principles enable researchers to substantiate findings and address biases through rigorous testing and validation processes. Transparency and reproducibility are essential for identifying and mitigating biases in AI models, ensuring fairness and equity in their deployment. In essence, the FAIR principles serve as a framework for promoting ethical AI development by minimizing biases at the foundational level of data management. By enhancing the transparency, accessibility, interoperability, and reusability of data, these principles empower researchers to uncover and address biases, ultimately fostering more equitable and trustworthy AI systems. Implementing the FAIR principles can help minimize bias in AI models by ensuring that data are high-quality, representative, and accessible. This promotes transparency and accountability in AI development, reducing the risk of biased outcomes.[67]

Additionally, there have been several AI checklists (eg, Standards for Reporting of Diagnostic Accuracy Studies-Artificial Intelligence, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Artificial Intelligence, Checklist for Artificial Intelligence in Medical Imaging, MINimum Information for Medical AI Reporting, Development of the Consolidated Criteria for Reporting Qualitative Research-Artificial

Table 3
List of medical artificial intelligence (AI) checklists or guidelines

Guideline/checklist	Purpose	Timeline of medical intervention	Correspondence to FAIR principles
STARD[a]-AI[68]	Reporting standards for studies evaluating diagnostic AI algorithms to ensure transparency.	Evaluation before diagnosis	F: Specifies reporting for diagnostic studies. A: Clear evaluation criteria I: Facilitates algorithm comparison R: Enhances diagnostic tool validation.
TRIPOD[b]-AI[69]	Reporting guidelines for studies developing and validating AI-driven diagnostic/prognostic models.	Development of diagnostic tools	F: Standardizes reporting for model studies. A: Clarity in model development I: Facilitates model comparison. R: Enhances model reproducibility.
PROBAST[c]-AI[69]	Reporting guidelines for assessing the risk of bias and applicability of diagnostic and prognostic AI models.	Development of diagnostic and prognostic tools	F: Advanced search and enriched metadata enhance data set discoverability. A: Clear access mechanisms and open access policies facilitate easy data retrieval. I: Standardized formats and integration support seamless system compatibility. R: Comprehensive documentation and clear licensing ensure effective data reuse.

(*continued on next page*)

Table 3 (*continued*)

Guideline/checklist	Purpose	Timeline of medical intervention	Correspondence to FAIR principles
CLAIM[d]/MI[e]-CLAIM[70,71]	Checklist for developing and validating AI models in medical imaging (CLAIM) and clinical applications (MI-CLAIM).	Development and validation of AI models	F: Provides checklist for model development. A: Clear validation criteria. I: Applies across imaging and clinical models R: Promotes model reuse.
MINIMAR[f,72]	Minimum reporting standard for AI studies in medicine, focusing on essential reproducibility elements.	Reporting of AI studies	F: Specifies essential reporting elements. A: Enhances clarity in study findings. I: Standardizes reporting across studies. R: Facilitates study replication.
PRIME[j,76]	Guidelines for developing AI in medical imaging to ensure robustness and reliability.	Development and validation of AI models	F: Specifies guidelines for AI development. A: Ensures robust development and validation. I: Facilitates comparison of AI models. R: Enhances model reliability.
DECIDE[g]-AI[73]	Development and validation of AI-driven decision support systems in clinical practice.	Integration into clinical decision-making	F: Clear guidelines for system development. A: Transparent validation methods. I: Ensures reliability across systems. R: Supports system evaluation and reuse.
SPIRIT[h]-AI[74]	Guidelines for writing protocols for clinical trials involving AI to ensure consistency.	Clinical trial protocol	F: Specifies protocol details for easy retrieval. A: Clarity in study design. I: Promotes consistency in study protocols. R: Facilitates study replication.
CONSORT[i]-AI[75]	Reporting standards for AI clinical trials to ensure transparency and completeness.	Reporting of AI clinical trials	F: Standardizes reporting for easy discovery. A: Enhances clarity for broader audience. I: Facilitates comparison across trials. R: Promotes reproducibility of trial findings.
QUADAS[k]-AI[77]	To evaluate the quality of studies assessing diagnostic accuracy of AI in healthcare. This tool helps to assess the methodological quality and risk of bias in AI-based diagnostic studies.	Reporting of AI studies	F: Standardizes reporting criteria for diagnostic accuracy. A: Ensures clear comparison for diagnostic accuracy studies. I: Facilitates comparison of diagnostic accuracy across studies. R: Ensures studies with diagnostic accuracy are study findings.

FAIR, findability, accessibility, interoperability, and reusability.

[a] Standards for Reporting of Diagnostic Accuracy Studies - Artificial Intelligence (STARD-AI) provides guidelines for reporting studies that evaluate diagnostic accuracy using AI technologies. It emphasizes transparent reporting of study methods, including participant selection criteria, reference standard procedures, AI algorithm details, and diagnostic performance metrics.

[b] Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Artificial Intelligence (TRIPOD-AI) provides guidelines for reporting studies developing and validating AI-driven diagnostic or prognostic models.

[c] Prediction model Risk of Bias Assessment Tool-Artificial Intelligence (PROBAST-AI) provides guidelines for assessing the risk of bias in prediction models developed using AI techniques. It offers structured criteria for evaluating key domains of bias, including participant selection, predictor variables, model development, and model evaluation. PROBAST-AI aims to improve the robustness and reliability of prediction models by identifying and addressing potential sources of bias.

[d] Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and Minimum Information about Clinical Artificial Intelligence Modelling (MI-CLAIM) provide checklists for developing and validating AI models in medical imaging and clinical applications, respectively.

[e] MI-CLAIM provides guidelines for reporting clinical AI models, focusing on transparency and reproducibility in model development and validation. It emphasizes comprehensive reporting of model inputs, outputs, performance metrics, validation methods, and potential limitations. MI-CLAIM aims to enhance the reliability and interpretability of clinical AI models across health care domains.

[f] MINimum Information for Medical AI Reporting (MINIMAR) aims to develop a minimal reporting standard for AI studies in medicine, focusing on essential elements necessary for reproducibility and evaluation.

[g] Development of the Consolidated Criteria for Reporting Qualitative Research-Artificial Intelligence (DECIDE-AI) provides guidelines for reporting qualitative research studies that involve AI applications. It emphasizes transparency in research methods, data collection, and analysis techniques used in qualitative AI studies. DECIDE-AI aims to ensure that qualitative findings related to AI development and implementation are rigorously documented and communicated.

[h] Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI) focuses on enhancing the completeness and transparency of protocols in AI-related clinical trials. It provides recommendations for reporting key elements of trial protocols, including eligibility criteria, randomization procedures, statistical analysis plans, and data handling processes. SPIRIT-AI aims to improve the reliability and reproducibility of AI trial results.

[i] Consolidated Standards of Reporting Trials for Artificial Intelligence (CONSORT-AI) aims to improve the transparency and quality of reporting in clinical trials that involve AI interventions. It provides guidelines for reporting essential elements such as study design, participant characteristics, intervention details, and outcomes. By standardizing reporting practices, CONSORT-AI facilitates the critical evaluation and replication of AI-driven clinical trials.

[j] PRIME (Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation).

[k] Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) focuses on assessing the quality of studies that evaluate the diagnostic accuracy of AI technologies. It provides criteria for evaluating study design, participant selection, reference standard procedures, AI algorithm performance, and data analysis methods. QUADAS-AI aims to enhance the reliability and validity of diagnostic accuracy studies involving AI applications.

Intelligence, Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence, and Consolidated Standards of Reporting Trials for Artificial Intelligence) for pre-clinical evaluations, clinical trials, and clinical translation that have been developed (Table 3).[68-77] These guidelines and checklists collectively enhance the transparency, reproducibility, and usability of AI research and clinical applications. They are intended to standardize reporting, development, and validation processes for AI models in diagnostics, predictive analytics, and decision support systems. By specifying clear criteria and protocols, these frameworks ensure clarity in study findings, ensure consistency in methodologies, and facilitate comparison across different studies

and applications. Moreover, ethical guidelines (eg, ÉCLAIR) promote responsible AI deployment, ensuring ethical standards are maintained across various health care settings. Together, these efforts support the FAIR principles by improving the findability, accessibility, interoperability, and reusability of AI technologies, thereby advancing their integration and impact in health care and research domains.

Addressing bias in AI-ML models will have profound implications for fairness, justice, and transparency, which can ultimately enhance our patient outcomes. Proactive measures aimed at mitigating bias are indispensable in fostering equitable outcomes and upholding fundamental principles of fairness and equity. A call to action is warranted, urging stakeholders across academia, industry, and policymaking spheres to prioritize responsible development and deployment of all AI technologies. Ethical considerations must underpin each stage of the AI lifecycle, from data collection to algorithmic design and validation, deployment, and ongoing monitoring of such to ensure equitable outcomes for all the patients we are privileged to serve. By addressing biases in data while improving our algorithm design and deployment practices, stakeholders will be able to develop AI systems that are more inclusive, equitable, and aligned with our ethical principles. This approach is critical to harnessing the full potential of AI while minimizing its unintended negative consequences on our patient populations. Additionally, the integration of FAIR principles and inclusive data practices emerges as a cornerstone in this quest for bias mitigation within our AI arena, which fosters transparency, accountability, and inclusivity in data utilization while allowing our stakeholders to chart a path toward a more equitable and just AI ecosystem.

Acknowledgments

The authors are thankful for all those have have contributed to the Computational Pathology & AI Center of Excellence (CPACE).

Author Contributions

This article series is part of the educational effort of Computational Pathology & AI Center of Excellence (CPACE) at the University of Pittsburgh. M.G.H., L.P., B.J., O.P., S.V., M.D., and H.H.R. contributed to the manuscript text contents. All images/figures were constructed by J.P.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

The authors received no funding for this study.

Declaration of Competing Interests

L.P. is a consultant for Hamamatsu, AiXMed and NTP, serves on the advisory board for Ibex, and is a co-owner of Placenta AI and Lean AP. H.H.R. is a creator of MILO and also coinventor of STNG.

Ethics Approval and Consent to Participate

Not applicable for review articles.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used DALL-E via ChatGPT-4o, as well as Adobe Express Online, in order to generate de novo artworks for the creation of certain figures or parts of figures. Additionally, a large language model was used to generate parts of the initial manuscript outline, but the text content of this manuscript is all human generated and did not include generative AI outputs.

References

1. Skeem JL, Lowenkamp C. Risk, race, and recidivism: predictive bias and disparate impact. *Criminol*. 2016;54(4):680–712. https://doi.org/10.2139/ssrn.2687339
2. Angwin J, Larson J, Mattu S, Kirchner L. ProPublica. Machine Bias. Accessed June 21, 2024. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Mukerjee A, Biswas R, Deb K, Mathur AP. Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. *Int Trans Oper Res*. 2002;9(5):583–597. https://doi.org/10.1111/1475-3995.00375
4. Bogen M, Rieke A. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. Upturn: 2018. Accessed June 21, 2024. https://apo.org.au/node/210071
5. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9(2), 010318. https://doi.org/10.7189/jogh.09.020318
6. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377–2378. https://doi.org/10.1001/jama.2019.18058
7. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *Npj Digit Med*. 2023;6(1):113. https://doi.org/10.1038/s41746-023-00858-z
8. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology; 2022. NIST SP 1270.
9. Research USNC for the P of HS of B and B. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research : Appendix*. Department of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research; 1978.
10. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. Oxford University Press; 2001.
11. Leslie D. *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. SSRN; 2019.
12. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*. 2019;10(1):3069. https://doi.org/10.1038/s41467-019-10933-3
13. Right to erasure. 2024. Accessed June 22, 2024. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/individual-rights/right-to-erasure/
14. Chung J, Zink A. Watson H—can i sue you for malpractice—examining the liability of artificial intelligence in medicine. *Asia Pac J Health Law Ethics*. 2017;11:51.
15. Price WN 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765–1766. https://doi.org/10.1001/jama.2019.15064
16. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. *Milbank Q*. 2021;99(3):629–647. https://doi.org/10.1111/1468-0009.12504
17. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167–E179. https://doi.org/10.1001/amajethics.2019.167
18. Dobson R, Wihongi H, Whittaker R. Exploring patient perspectives on the secondary use of their personal health information: an interview study. *BMC Med Inform Decis Mak*. 2023;23(1):66. https://doi.org/10.1186/s12911-023-02143-1
19. Kim J, Kim H, Bell E, et al. Patient perspectives about decisions to share medical data and biospecimens for research. *JAMA Netw Open*. 2019;2(8):e199550. https://doi.org/10.1001/jamanetworkopen.2019.9550
20. Skloot R. *The Immortal Life of Henrietta Lacks. Broadway Paperbacks, an Imprint of the Crown Publishing Group, a Division of*. Random House, Inc; 2011.
21. *Hollywood writers' battle against AI, humans win (for now)*. AP News; 2023. Accessed July 20, 2024. https://apnews.com/article/hollywood-ai-strike-wga-artificial-intelligence-39ab72582c3a15f77510c9c30a45ffc8

22. Robertson K. 8 Daily Newspapers Sue OpenAI and Microsoft Over A.I. *The New York Times*. 2024. Accessed July 20, 2024. https://www.nytimes.com/2024/04/30/business/media/newspapers-sued-microsoft-openai.html
23. Ornstein C, Thomas K. Sloan Kettering's Cozy Deal With Start-Up Ignites a New Uproar. *The New York Times*. 2018. Accessed July 20, 2024. https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html
24. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American Multisociety statement. *Radiology*. 2019;293(2):436–440. https://doi.org/10.1148/radiol.2019191586
25. Petersen C, Berner ES, Embi PJ, et al. AMIA's code of professional and ethical conduct 2018. *J Am Med Inform Asso*. 2018;25(11):1579–1582. https://doi.org/10.1093/jamia/ocy092
26. Allen TC. Regulating artificial intelligence for a successful pathology future. *Arch Pathol Lab Med*. 2019;143(10):1175–1179. https://doi.org/10.5858/arpa.2019-0229-ED
27. Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med*. 1997;9(2):107–138. https://doi.org/10.1016/S0933-3657(96)00367-3
28. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc Natl Acad Sci USA*. 2016;113(16):4296–4301. https://doi.org/10.1073/pnas.1516047113
29. Hogarth RA. The myth of innate racial differences between White and Black People's bodies: lessons from the 1793 Yellow Fever Epidemic in Philadelphia, Pennsylvania. *Am J Public Health*. 2019;109(10):1339–1341. https://doi.org/10.2105/AJPH.2019.305245
30. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517–518. https://doi.org/10.1001/jama.2017.7797
31. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981–983. https://doi.org/10.1056/NEJMp1714229
32. Stram M, Seheult J, Sinard JH, et al. A survey of LOINC code selection practices among participants of the College of American Pathologists Coagulation (CGL) and Cardiac Markers (CRT) proficiency testing programs. *Arch Pathol Lab Med*. 2020;144(5):586–596. https://doi.org/10.5858/arpa.2019-0276-OA
33. Carter AB, de Baca ME, Luu HS, Campbell WS, Stram MN. Use of LOINC for interoperability between organisations poses a risk to safety. *Lancet Digit Health*. 2020;2(11), e569. https://doi.org/10.1016/S2589-7500(20)30244-2
34. 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):1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763
35. Friedman B, Nissenbaum H. Bias in computer systems. *ACM Trans Inf Syst*. 1996;14(3):330–347. https://doi.org/10.1145/230538.230561
36. Fave X, Mackin D, Yang J, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med Phys*. 2015;42(12):6784–6797. https://doi.org/10.1118/1.4934826
37. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52(7):1391–1397. https://doi.org/10.3109/0284186X.2013.812798
38. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J Med Imaging (Bellingham)*. 2015;2(4), 041002. https://doi.org/10.1117/1.JMI.2.4.041002
39. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer. *Transl Oncol*. 2015;8(6):524–534. https://doi.org/10.1016/j.tranon.2015.11.013
40. Dehkharghanian T, Bidgoli AA, Riasatian A, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn Pathol*. 2023;18(1):67. https://doi.org/10.1186/s13000-023-01355-3
41. Leo P, Lee G, Shih NN, Elliott R, Feldman MD, Madabhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: predicting biochemical recurrence from prostate cancer whole slide images. *J Med Imaging (Bellingham)*. 2016;3(4), 047502. https://doi.org/10.1117/1.JMI.3.4.047502
42. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR; 2018:77–91. Accessed April 19, 2022. *https://proceedings.mlr.press/v81/buolamwini18a.html*
43. Vela MB, Erondu AI, Smith NA, Peek ME, Woodruff JN, Chin MH. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annu Rev Public Health*. 2022;43:477–501. https://doi.org/10.1146/annurev-publhealth-052620-103528
44. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *Eur Urol*. 2016;69(3):428–435. https://doi.org/10.1016/j.eururo.2015.06.046
45. Hattab EM, Koch MO, Eble JN, Lin H, Cheng L. Tertiary Gleason pattern 5 is a powerful predictor of biochemical relapse in patients with Gleason score 7 prostatic adenocarcinoma. *J Urol*. 2006;175(5):1695–1699. https://doi.org/10.1016/S0022-5347(05)00998-5. discussion 1699.
46. Fletcher SA, Marchese M, Cole AP, et al. Geographic distribution of racial differences in prostate cancer mortality. *JAMA Netw Open*. 2020;3(3), e201839. https://doi.org/10.1001/jamanetworkopen.2020.1839
47. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383(9):874–882. https://doi.org/10.1056/NEJMms2004740
48. Snipes SA, Sellers SL, Tafawa AO, Cooper LA, Fields JC, Bonham VL. Is race medically relevant? A qualitative study of physicians' attitudes about the role of race in treatment decision-making. *BMC Health Serv Res*. 2011;11:183. https://doi.org/10.1186/1472-6963-11-183
49. Levey AS, Tighiouart H, Titan SM, Inker LA. Estimation of glomerular filtration rate with vs without including patient race. *JAMA Intern Med*. 2020;180(5):793–795. https://doi.org/10.1001/jamainternmed.2020.0045
50. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150(9):604–612. https://doi.org/10.7326/0003-4819-150-9-200905050-00006
51. Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. *JAMA*. 2019;322(2):113–114. https://doi.org/10.1001/jama.2019.5774
52. Peterson PN, Rumsfeld JS, Liang L, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes*. 2010;3(1):25–32. https://doi.org/10.1161/CIRCOUTCOMES.109.854877
53. Shahian DM, Jacobs JP, Badhwar V, et al. The society of thoracic surgeons 2018 adult cardiac surgery risk models: part 1-background, design considerations, and model development. *Ann Thorac Surg*. 2018;105(5):1411–1418. https://doi.org/10.1016/j.athoracsur.2018.03.002
54. Rao PS, Schaubel DE, Guidinger MK, et al. A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation*. 2009;88(2):231–236. https://doi.org/10.1097/TP.0b013e3181ac620b
55. Grobman WA, Lai Y, Landon MB, et al. Development of a nomogram for prediction of vaginal birth after cesarean delivery. *Obstet Gynecol*. 2007;109(4):806–812. https://doi.org/10.1097/01.AOG.0000259312.36053.02
56. Moore CL, Bomann S, Daniels B, et al. Derivation and validation of a clinical prediction rule for uncomplicated ureteral stone—the STONE score: retrospective and prospective observational cohort studies. *BMJ*. 2014;348, g2191. https://doi.org/10.1136/bmj.g2191
57. Wang RC, Rodriguez RM, Moghadassi M, et al. External validation of the STONE score, a clinical prediction rule for ureteral stone: an observational multi-institutional study. *Ann Emerg Med*. 2016;67(4):423–432.e2. https://doi.org/10.1016/j.annemergmed.2015.08.019
58. Shaikh N, Hoberman A, Hum SW, et al. Development and validation of a calculator for estimating the probability of urinary tract infection in young febrile children. *JAMA Pediatr*. 2018;172(6):550–556. https://doi.org/10.1001/jamapediatrics.2018.0217
59. Bowles TL, Hu CY, You NY, Skibber JM, Rodriguez-Bigas MA, Chang GJ. An individualized conditional survival calculator for patients with rectal cancer. *Dis Colon Rectum*. 2013;56(5):551–559. https://doi.org/10.1097/DCR.0b013e31827bd287
60. Breast Cancer Risk Assessment Tool: Online Calculator (The Gail Model). The Breast Cancer Risk Assessment Tool. Accessed June 22, 2024. https://bcrisktool.cancer.gov
61. Tice JA, Miglioretti DL, Li CS, Vachon CM, Gard CC, Kerlikowske K. Breast density and benign breast disease: risk assessment to identify women at high risk of breast cancer. *J Clin Oncol*. 2015;33(28):3137–3143. https://doi.org/10.1200/JCO.2015.60.8869
62. Kanis: Assessment of osteoporosis at the primary... - Google Scholar. Accessed June 22, 2024. https://scholar.google.com/scholar_lookup?title=Assessment+of+osteoporosis+at+the+primary+health+care+level.+WHO+Scientific+Group+technical+report&publication_year=2007
63. Braun L. *Breathing Race into the Machine: The Surprising Career of the Spirometer from Plantation to Genetics*. University of Minnesota Press; 2014.
64. Skiba JH, Bansal AD, Palmer OMP, Johnstone DB. Case report: clinical consequences of adjusting estimated GFR for black race. *J Gen Intern Med*. 2022;37(4):958–961. https://doi.org/10.1007/s11606-021-07179-5
65. Chin MH, Afsar-Manesh N, Bierman AS, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Netw Open*. 2023;6(12):e2345050. https://doi.org/10.1001/jamanetworkopen.2023
66. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1), 160018. https://doi.org/10.1038/sdata.2016.18
67. Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf Serv Use*. 2017;37(1):49–56. https://doi.org/10.3233/ISU-170824
68. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the

STARD-AI Steering Group. *Nat Med*. 2020;26(6):807–808. https://doi.org/10.1038/s41591-020-0941-1
69. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. https://doi.org/10.1136/bmjopen-2020-048008
70. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. https://doi.org/10.1148/ryai.2020200029
71. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. 2024;6(4), e240300. https://doi.org/10.1148/ryai.240300
72. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27(12):2011–2015. https://doi.org/10.1093/jamia/ocaa088
73. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904
74. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2(10):e549–e560. https://doi.org/10.1016/S2589-7500(20)30219-3
75. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374. https://doi.org/10.1038/s41591-020-1034-x
76. 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;13:2017–2035. https://doi.org/10.1016/j.jcmg.2020.07.015
77. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27:1663–1665. https://doi.org/10.1038/s41591-021-01517-0