JAMIA Open, 2025, **8(1)**, ooae155
https://doi.org/10.1093/jamiaopen/ooae155
Review

Review

Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review

Kirubel Biruk Shiferaw, MPH[1,*], Moritz Roloff, MD[1], Irina Balaur, PhD[2], Danielle Welter, PhD[3], Dagmar Waltemath, PhD[1], Atinkut Alamirrew Zeleke, PhD[1]

[1]Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Greifswald D-17475, Germany, [2]Luxembourg Centre for Systems Biology, University of Luxembourg, Belvaux L-4367, Luxembourg, [3]Luxembourg National Data Service, Esch-sur-Alzette L-4362, Luxembourg

*Corresponding author: Kirubel Biruk Shiferaw, MPH, Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Walther-Rathenau-Str. 48, Greifswald D-17475, Germany (s-kishif@uni-greifswald.de)

Abstract

Objectives: The continuous integration of artificial intelligence (AI) into clinical settings requires the development of up-to-date and robust guidelines and standard frameworks that consider the evolving challenges of AI implementation in medicine. This review evaluates the quality of these guideline and summarizes ethical frameworks, best practices, and recommendations.

Materials and Methods: The Appraisal of Guidelines, Research, and Evaluation II tool was used to assess the quality of guidelines based on 6 domains: scope and purpose, stakeholder involvement, rigor of development, clarity of presentation, applicability, and editorial independence. The protocol of this review including the eligibility criteria, the search strategy data extraction sheet and methods, was published prior to the actual review with International Registered Report Identifier of DERR1-10.2196/47105.

Results: The initial search resulted in 4975 studies from 2 databases and 7 studies from manual search. Eleven articles were selected for data extraction based on the eligibility criteria. We found that while guidelines generally excel in scope, purpose, and editorial independence, there is significant variability in applicability and the rigor of guideline development. Well-established initiatives such as TRIPOD+AI, DECIDE-AI, SPIRIT-AI, and CONSORT-AI have shown high quality, particularly in terms of stakeholder involvement. However, applicability remains a prominent challenge among the guidelines. The result also showed that the reproducibility, ethical, and environmental aspects of AI in medicine still need attention from both medical and AI communities.

Discussion: Our work highlights the need for working toward the development of integrated and comprehensive reporting guidelines that adhere to the principles of Findability, Accessibility, Interoperability and Reusability. This alignment is essential for fostering a cultural shift toward transparency and open science, which are pivotal milestone for sustainable digital health research.

Conclusion: This review evaluates the current reporting guidelines, discussing their advantages as well as challenges and limitations.

Lay Summary

As artificial intelligence (AI) continues to play an increasingly central role in health care, its safe and effective integration requires high-quality reporting guidelines that address the specific challenges of implementing AI in clinical settings. This systematic review evaluates the quality of existing AI reporting guidelines in medicine, focusing on their ethical considerations, best practices, and practical recommendations. Using the Appraisal of Guidelines, Research, and Evaluation II tool, we assessed the quality of the guidelines based on 6 key domains: scope, stakeholder involvement, development rigor, clarity, applicability, and editorial independence. While most guidelines performed well in areas such as scope and stakeholder involvement, significant variability was observed in their applicability and development rigor, reflecting the ongoing challenge of translating AI research into real-world health care settings. Additionally, the review highlighted the need for greater attention to reproducibility, ethics, and environmental impacts in medical AI research. Furthermore, the study underscores the importance of aligning guidelines with the Findable, Accessible, Interoperable, and Reusable principles to foster transparency, open science, and collaboration, which are crucial for advancing sustainable digital health research.

Key words: digital medicine; artificial intelligence; machine learning; guidelines; quality; framework; AGREE II; medicine; standard; systematic review; medical informatics.

Introduction

According to the European Union (EU) high-level expert group definition,

> Artificial Intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, reasoning and deciding the best action(s) to take to achieve the given goal.[1]

The expert group also described the technical approaches in AI including machine learning (ML) (such as deep and reinforcement learning), machine reasoning (such as knowledge

Received: October 15, 2024; **Revised:** December 12, 2024; **Editorial Decision:** December 17, 2024; **Accepted:** December 20, 2024

representation), and robotics (such as control, sensors, and actuators). In this work, we use the definition of ML and AI proposed by the EU high-level expert group.[1]

Artificial intelligence has emerged as a promising and yet disruptive technological advancement with the potential to transform health care.[2–4] Studies have shown that AI can improve the diagnostic accuracy, support clinical decisions, predict risk/events, help discover drugs, and support patient management.[5–7] Nonetheless, the ongoing incorporation of AI in clinical settings necessitates the development of current, reliable and robust guidelines, and standard frameworks that consider the evolving challenges of AI implementation in medicine.[8]

Several guidelines for developing and reporting ML models were created by experts worldwide.[9,10] However, an extensive and continuous evaluation of guidelines is still missing to maintain credibility, standardization, quality of care, patient safety, data protection, and ethical research.[11] The agile and ever-evolving challenges in this field impede the process of crafting a gold standard that would cover all aspects of developing and reporting AI studies in the medical domain. For instance, the "hype" in developing and reporting "best-performing" models has recently been challenged by questions regarding reproducibility, explainability, governance, and ethical implications for use in health care.[12] Generative-AI (GenAI) and large language models have already stimulated substantial discourse in science and innovation since 2022.[13]

Reproducibility is one of the most prominent challenges for AI in medicine and science in general.[14] Often general textual descriptions of methods and results are published, with oversimplistic levels of details about the necessary steps in preprocessing, model training and validation, and reporting.[15] A limited use of standardized ML model development and reporting guidelines but also the lack of standardized sharing practices of input data and source code hamper reproducibility.[16,17] From a computational modeling point of view, sharing the model data and code would foster the reusability of the models to answer new research questions or advance the performance of the existing models,[17] a topic that has long been discussed in other fields such as Systems Medicine. A previous review conducted on 72 guidance documents grouped the available guidance resources in 6 phases: (1) data preparation, (2) AI prediction model development, (3) validation, (4) software development (5) impact assessment, and (6) implementation across identified topics.[18] However, the perspective with regard to reporting guideline was missing. A similar review (conducted after the protocol of this review was published) highlighted 26 reporting guidelines grouped into 3 categories (preclinical, translational, and clinical reporting guidelines).[19] The study discussed 14 general reporting guidelines in medical AI research and yet the quality of the guidelines was not assessed.

Due to its complex and sensitive nature, experts and regulatory stakeholders continuously seek up-to-date guidelines when applying AI in medicine. Often, fragments of suggestions and guiding frameworks are developed by different experts, and scientists face the challenge of choosing the appropriate guideline for a specific use case.[20] Thus, the evaluation of existing guidelines would help scientists to identify the best framework to follow in a specific project. Here, we performed a systematic review of available guidelines for ML model development and reporting in health care. We assessed the quality of the guidelines and summarize the ethical frameworks, checklists, best practices, and recommendations. To effectively harness the benefits of AI in medicine, it is essential to not only develop and update guidelines but also to implement well-established datasets and code-sharing concepts. The Findable, Accessible, Interoperable and Reusable (FAIR) guiding principles are the widely accepted approach for scientific data management and stewardship.[21] Their applicability in making software[22] and digital artifacts, such as ML models,[23] FAIR has been shown over the past years, and it is evident that adherence to these principles maximizes research value and fosters open and reproducible science.[24,25]

Methods

A systematic review was conducted following the Preferred Reporting Items for Systematic Review and Meta-Analysis 2020 guidelines.[26] PubMed and Web of Science (WOS) databases were systematically searched. Two reviewers screened titles, abstracts, and full texts for eligibility and performed data extraction based on a predefined data extraction sheet. Quality assessment was performed using the Appraisal of Guidelines, Research, and Evaluation II (AGREE II) tool[27] and discrepancies were resolved through consensus or third-party arbitration. Data synthesis and analysis were conducted using Python.

Protocol and registration

The protocol is published in JMIR Protocols with Digital Object Identifier and International Registered Report Identifier: DERR1-10.2196/47105.[28]

Eligibility criteria

All available guidelines, standard frameworks, best practices, checklists, and recommendations on the topic of reporting AI research in medicine were included irrespective of the study design. Studies were restricted to English language and publications until June 2023.

Search strategy

A systematic literature search was commenced using medical subject headings terms and keywords for medicine, guidelines, and ML (Supplementary Material S1). We used the PubMed and WOS databases and the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network, which is a global initiative working toward improving research value by promoting robust reporting guidelines (http://www.equator-network.org/). Google Scholar search for references in selected papers led to more thorough search results. Then, the search results were uploaded to an online systematic review tool (Rayyan) and then processed with CADIMA,[29] a free web tool facilitating the development of systematic reviews and associated documentation, for further screening and preliminary analysis.

Study selection

After removing duplicates using CADIMA, titles and abstracts were scanned by 2 independent reviewers (K.B.S. and M.R.). The reviewers then performed an independent review of full texts and final decisions on whether to include the article for data extraction were made after discussion.

Data extraction, collection, and management

Two independent reviewers (K.B.S. and M.R.) extracted relevant information (such as study characteristics, study type, aspect, and specific disease/condition of interest if available) from the identified publications using a predefined information extraction sheet.

Quality and risk of bias assessment

Quality, specifically in the context of guidelines, frames the methodological parameters that dictate how other studies should be conducted, reported, and communicated. We assessed the quality of identified guidelines using the AGREE II tool.[27] AGREE II measures the quality of guidelines in 6 fundamental domains including methodological rigorousness and transparency of the guideline development process.[27] Specifically, AGREE II contains 23 items, each rated on a Likert scale rating from 1 (strongly disagree) to 7 (strongly agree) and grouped within the following 6 domains:

- *Domain 1*—Scope and purpose: assesses whether the guideline stated the main target and scope of the intended use of the guideline.
- *Domain 2*—Stakeholder involvement: assesses whether the guideline development process incorporated a representative view of relevant stakeholders including users.
- *Domain 3*—Rigor of development: evaluates the methodological thoroughness followed during the guideline development process.
- *Domain 4*—Clarity of presentation: assesses the clarity of format and language conveyed in the proposed guideline.
- *Domain 5*—Applicability: assesses the presentation of facilitators and barriers to implement the guidelines. The measures need to be considered for the applicability of the guideline.
- *Domain 6*—Editorial independence: assesses the statement with respect to funding bias and competing interests.
- *Overall assessment:* This domain reflects the subjective assessment of the evaluators regarding the overall quality of the guideline and their opinion in recommending the use.

The assessment result and individual scores along with the intraclass correlation (reviewer agreement) can be found in Supplementary Material S2.

Analysis of the guideline quality assessment

To evaluate the risk of bias, 4 independent appraisers performed a quality evaluation of the 11 identified guidelines (details of these guidelines are given in Supplementary Material S3). The rating was calculated by scaling the total as a percentage of the maximum possible scores for a specific domain.[27] For example, domain 1 (scope and purpose) has 3 items, scored from 1 (strongly disagree) to 7 (strongly agree). Hence, the maximum possible score is $7 \times 3 \times 4 = 84$ (where 4 is the number of appraisers), and the minimum possible score is $1 \times 3 \times 4 = 12$. Thus, a domain score is calculated as follows:

$$\text{Domain score} = \frac{\text{Obtained score} - \text{Minimum possible score}}{\text{Maximum possible score} - \text{Minumum possible score}}$$

It is important to note that each domain score is calculated independently, and it is neither recommended to combine domains nor to average the result. Item 11 and 16, which are specific to medical practice guidelines, were adjusted to the median value for all reviewers. AGREE II outline that users can prioritize one domain over others, creating thresholds based on scores for that domain (eg, high-quality guidelines are those with a domain 3 score >70%). Alternatively, a staged AGREE II appraisal can be conducted, where guidelines are first evaluated using the prioritized domain, and only those meeting the threshold are assessed across the other domains.[27] In this study, we prioritized "rigor of development," "stakeholder involvement," and "applicability" to compare and discuss guideline quality. We used intraclass correlation coefficient (ICC) to assess the interrater agreement.

Results

The initial search resulted in 4975 studies from PubMed and WOS databases, with additional 7 studies identified through manual searches in the EQUATOR Network, citation tracking and reviewer recommendation. Two reviewers independently conducted full text reviews of 266 studies and selected 9 studies for detailed data extraction and synthesis (Figure 1). During the peer review process, 2 additional new guidelines were considered for review, making the total studies included 11.

Table 1 indicates the main characteristics of the 11 reporting guidelines. More details about the selected reporting guidelines are presented in Supplementary Material S3.

Quality assessment of the identified guidelines using AGREE II

The AGREE II tool has been designed to evaluate the quality of clinical practice guidelines.[27] We found that the structured and generic framework can be adapted and applied effectively to evaluate nonclinical practice guidelines.[42] The 6 core domains of AGREE II are universally applicable to any set of guidelines or recommendations (details in "Methods"), except for 2 items (items 11 and 16), which are specific for clinical practice guidelines. Its standardized evaluation process permits comparable and consistent evaluation across several guideline types. By using AGREE II in this work, we aim to contribute to evaluating the quality, relevance, and impact of nonclinical guidelines, making them a valuable resource for decision-makers and stakeholders. The primary domains of focus, in order of relevance, were:

1) The rigor of the guideline development process (AGREE II domain 3). We chose it because the thoroughness of the method followed in developing the guideline reflects the quality of the guideline itself.
2) Stakeholder involvement (domain 2), which indicates whether all relevant stakeholders are involved. The premise is that engaging more stakeholders in the development process of a guideline contributes to its quality and usability.
3) Applicability or instruction how the guideline can be used in practice (domain 5), which shows the guideline practicality.

Following the guideline assessment, we calculated the aggregated score for each domain by scaling the total (obtained from the 4 reviewers) as a percentage of maximum possible scores (details in "Methods").

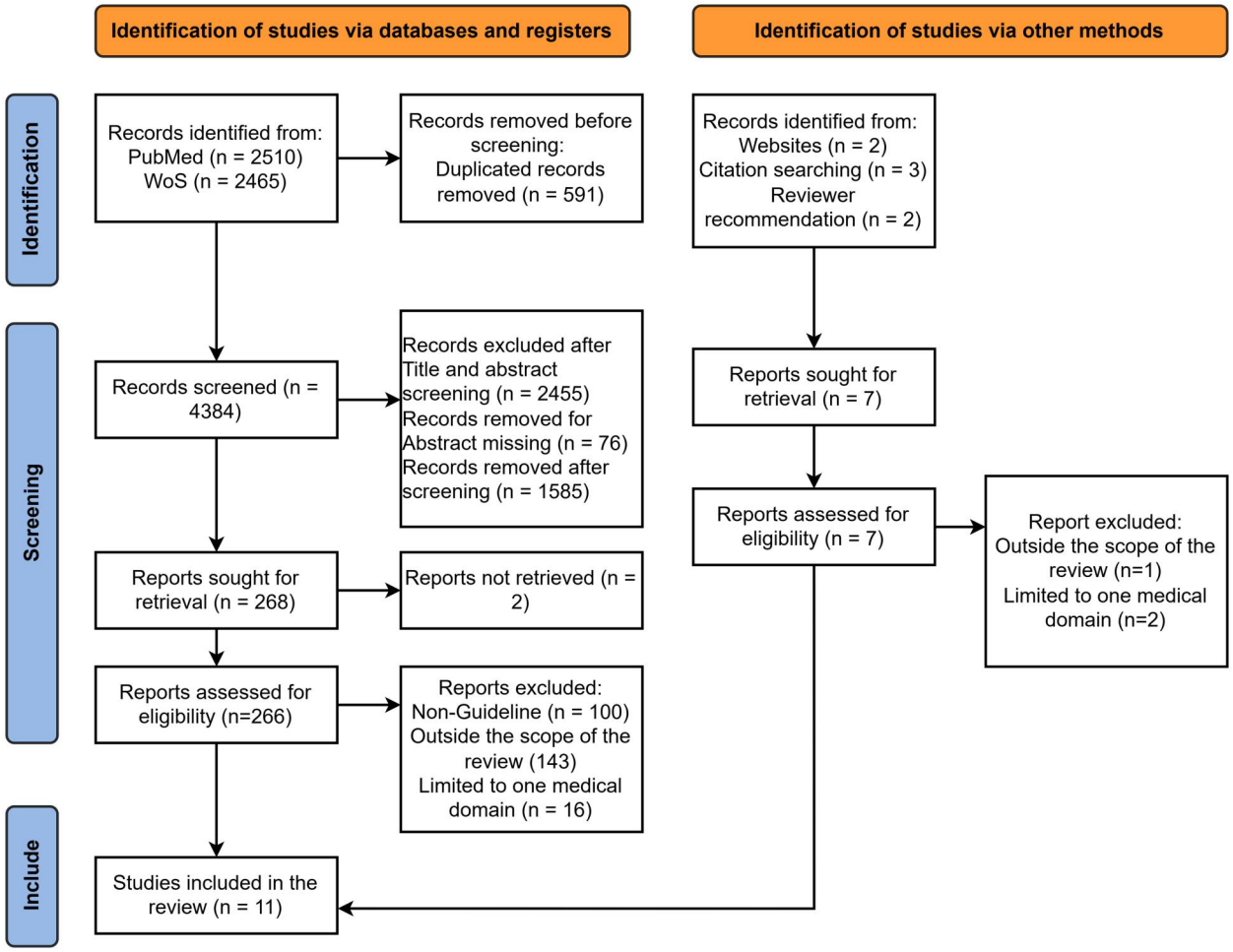

Figure 1. PRISMA flowchart: reporting guidelines of AI-related studies in medicine. Reference date: June 2023. Reports: refer to the full-text articles that are assessed for eligibility after initial screening of records. Records: refer to individual citation of reference that are identified during the literature search. Abbreviation: AI, artificial intelligence.

Our results show that the aggregate scores of scope and purpose/domain 1 (which range from 68.1% to 93.3%) and the editorial independence/domain 6 (range from 75.0% to 97.9%) are the most satisfied criteria of AGREE II across the guidelines. We also observed a clear variability of domain scores across guidelines (Figure 2); the lowest scoring domain refers to domain 5/applicability with 26.0%. When comparing at guideline level, DECIDE-AI and TRIPOD+AI have the highest score across most of the domains. Regarding the "rigor of development" (domain 3), 6 guidelines scored above 70% (TRIPOD+AI, DECIDE-AI, APPRAISE-AI, CONSORT-AI, SPIRIT-AI, and TRIPOD), indicating their higher quality.

TRIPOD+AI and DECIDE-AI are the highest quality guidelines with respect to rigor of development and the involvement of the stakeholders, followed by SPIRIT-AI, APPRAISE-AI, and CONSORT-AI. Moreover, the guideline CLEAR was scored as the highest quality with respect to applicability. See Figure 3 for more details.

We evaluated the interrater agreement regarding the overall and primary domain-level consensus among the 4 evaluators using the ICC.[43] The overall agreement among the 4 independent evaluators regarding the quality of the guidelines was statistically significant ranging from ICC of 0.62 to 0.92 with P-value $< .05$. The details of individual scoring and domain-level ICC can be found in Supplementary Material S2.

Contributions to better reproducibility in AI in medicine and beyond

Reproducibility, described as "the ability of an independent research team to produce the same results using the AI method based on the documentation made by the original research team,"[16] requires an exact representation of all relevant aspects of the study development and realization. This includes the complete information of the used software and source code, the original data as well as the correct documentation of crucial details and precise instructions for the implementation.[44–46] The reproducibility in AI builds trust in the developed models and results.[15,45] Therefore, aiming for reproducibility, focusing on the correct and detailed documentation, and providing the necessary details regarding the source code and data information should be mandatory for every researcher and developer to achieve highly valued and trustful scientific findings.

Given its definition, model reproducibility comes with its challenges related to access to data, code, documentation, and clear instructions. Without the opportunity to access any of the given requirements, researchers fail to reproduce roughly similar findings compared to the original study. The lack of proper

Table 1. Characteristics of the selected reporting guidelines of AI-related studies in medicine.

	Guideline	First author's name, year of publication	Name of the journal	Outcome	Aspect	Standard followed	Domain
1	TRIPOD[30]	Gary S. Collins, 2015	*Circulation*	Checklist of 22 items	Reporting guideline	Yes, Guidance for Developers of Health Research Reporting Guidelines[31]	Diagnostic or prognostic prediction model reporting
2	SPIRIT-AI[32]	Samantha C. Rivera, 2020	*The Lancet Digital Health*[a]	SPIRIT 2013 items () + 15 checklist items	Trial/Protocol registration	Yes, the EQUATOR Network's methodological framework	Reporting of AI trial protocols
3	MG[33]	Romana Haneef, 2022	*Archives of Public Health*	Checklist of 8 core items with 33 subitems and free text items	Developing and reporting guideline	No, but used a "Stepwise approach"	Reporting linked data/ ML studies
4	CONSORT-AI[34]	Xiaoxuan Liu, 2020	*The Lancet Digital Health*[a]	CONSORT + 14 new items	Reporting guideline for AI intervention trials	Yes, the EQUATOR Network's methodological framework	Reporting clinical trials with AI intervention
5	GDRML[35]	Wei Luo, 2016	*Journal of Medical Internet Research*	12 checklist items	Developing and reporting guideline	No/not reported	Developing and reporting ML studies
6	STARD 2015[36]	Patrick M. Bossuyt, 2015	*Radiology*[a]	30 checklist items	Reporting guideline	Not reported but used a stepwise approach (documented in EQUATOR Network) to update the STARD 2003	Reporting of diagnostic accuracy studies
7	DECIDE-AI[37]	Baptiste Vasey, 2022	*The BMJ*	27 checklist items (17 AI specific and 10 generic)	Reporting guideline	Yes, the EQUATOR Network's methodological framework	Reporting of early-stage clinical evaluation of AI systems
8	MI_CLAIM[38]	Beau Norgeot, 2020	*Nature Medicine*	21 checklist items	Reporting guideline	Not reported	Reporting best practice checklist for minimum information about AI modeling
9	CLEAR[39]	Burak Kocak, 2023	*Insights into Imaging*	58 checklist items	Reporting guideline	No/not reported, but used a modified Delphi method for final selection of items	Reporting guideline for radiomics studies
10	TRIPOD+AI[40]	Gary S. Collins, 2024	*The BMJ*	27 checklist items	Reporting guideline	Yes, the EQUATOR Network's methodological framework	Diagnostic or prediction model reporting
11	APPRAISE-AI[41]	Jethro C.C. Kwong, 2023	*JAMA Network Open*	24 checklist items	Reporting guideline	Yes, the Standards for Quality Improvement Reporting Excellence (SQUIRE)	Evaluation of ML model for clinical decision support

Abbreviations: AI, artificial intelligence; ML, machine learning.
[a] Published in multiple journals.

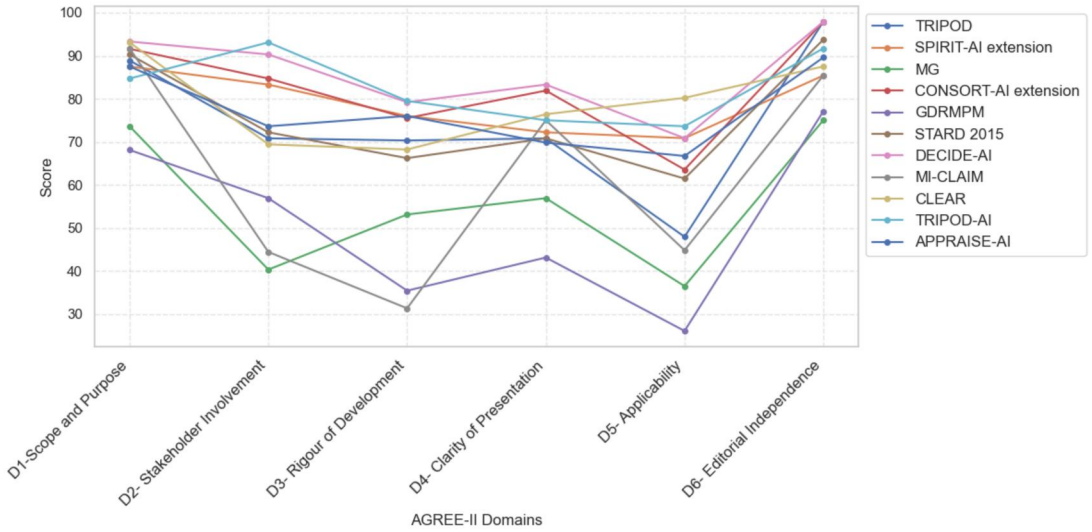

Figure 2. Parallel coordinate plot of reporting guidelines' aggregate evaluation score across the 6 AGREE II domains.

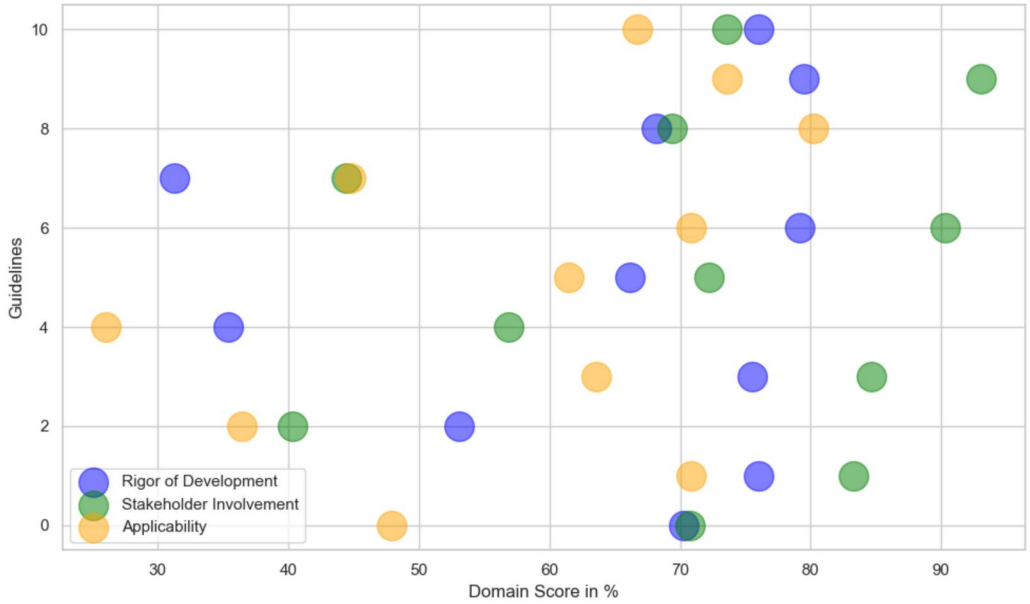

Figure 3. Scatter plot of targeted AGREE II domains across guidelines. Each horizontal grid represents each reporting guideline and each bubble on the respective grid represents the guidelines' score with respect to rigor of development (blue), stakeholder involvement (green), and applicability (yellow).

upkeep of essential resources, such as data, code, or instructions, hinders advancements in research and impedes reproducibility.[47] In addition, the current academic environment encourages researchers to publish prototypes of their AI models rather than ensuring a fully verified system,[46] which also impacts the quality of these models. Figure 4 illustrates the 3 important elements of medical AI research.

Standard frameworks and best practices for AI model reproducibility

Frameworks, guidelines, and best practices should offer guidance to achieve a minimum reproducibility standard to ensure reliable results in future studies.[48]

Heil et al.[45] proposed a reproducibility standard at 3 different levels. According to this work, the level of reproducibility can be given on a time-scale based on the time needed to reproduce the work. The scale starts at "forever" for an irreproducible study and ends at "zero" for an automated and fast reproducible study. On this scale, the 3 degrees "bronze," "silver," and "gold" define which requirements have to be met to achieve reproducibility, with "bronze" symbolizing the bare minimum and "gold" meaning the research team ensured full automation. The checklist for reproducibility focuses on a detailed description (and publication) of all used models and algorithms and the complexity of the analysis.[49] Furthermore, any theoretical claim has to be proven entirely and assumptions explained. Figures, tables, corresponding datasets, and the work flow should be presented in detail.

Discussion

The systematic search resulted in 11 reporting guidelines for AI in medicine. The quality assessment result indicated that

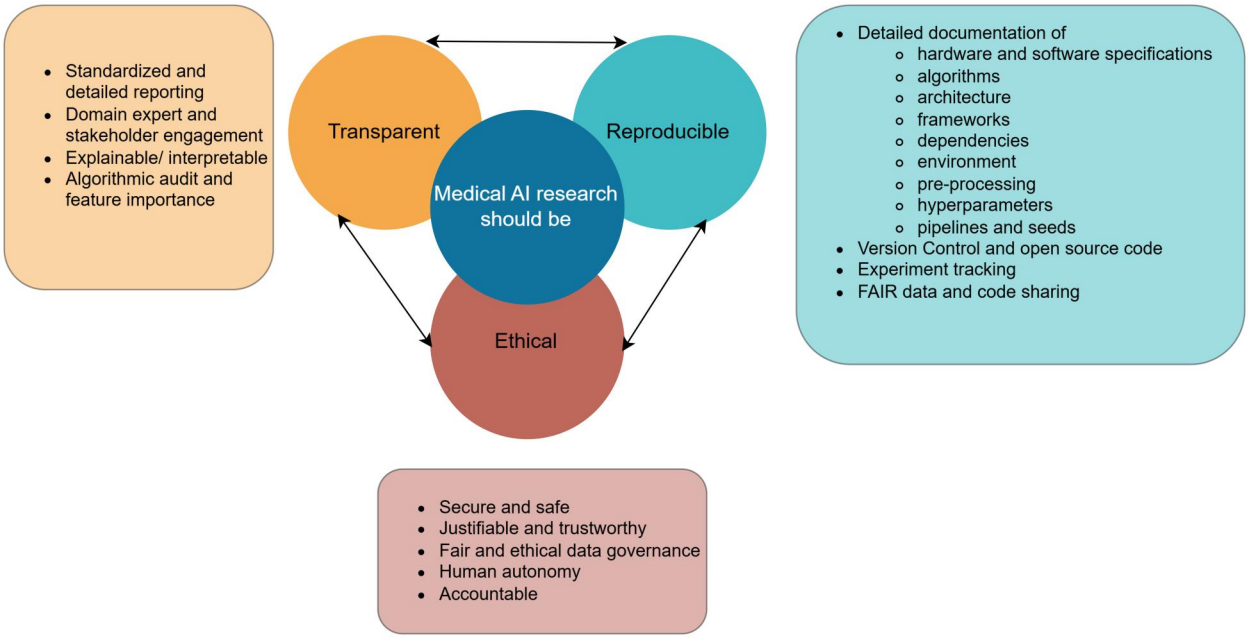

Figure 4. Elements of transparency, reproducibility, and ethics in medical AI research. Abbreviation: AI, artificial intelligence.

the overall quality of available reporting guidelines with respect to describing the scope and purpose (domain 1) and editorial independence is relatively well scored across the guidelines. Greater variability of scores in explaining the applicability (domain 5) and rigor of the guideline development process (domain 3) were observed. With respect to the primary domain of quality evaluation in this study (domain 3), TRIPOD+AI, DECIDE-AI, APPRAISE-AI, SPIRIT-AI, and CONSORT-AI reporting guidelines scored the highest with 79.5%, 79.2%, 76%, 76%, and 75.5%, respectively. The secondary quality criterion, stakeholder involvement (domain 2), was also scored higher by the same guidelines, with score of 93.1%, 90.3%,73.6%, 83.3%, and 84.7%, respectively.

All the identified guidelines present a way of reporting studies as a checklist of important sections such as introduction, methods, results, discussion, conclusion, and additional information sections. The majority of the reporting guidelines were not designed for AI studies per se but were extended to accommodate studies involving AI. The extension was mostly done by adding additional items to the checklists that were already in use for reporting a certain type of research findings. For instance, TRIPOD+AI, SPIRIT-AI, and CONSORT-AI are extensions of TRIPOD, SPIRIT, and CONSORT statements which were originally designed to report prediction models, clinical trial protocols, and clinical trial studies, respectively.[50] There are also other guidelines in development, such as the Quality Assessment of Diagnostic Accuracy Studies Using AI.[51] Originally designed to assess the quality of diagnostic accuracy studies, this framework is currently being adapted to incorporate AI-based studies.

The adaptation of guidelines for AI studies clearly improve the completeness of the report. One step toward reproducibility is the publication of code and related information (see the list of resources for sharing code in Supplementary Material S4.

While these guidelines provide a roadmap to reproducibility, they also highlight the need for a cultural transformation within the medical AI research community. A recent review highlighted that from the total of 63 clinical trials with AI intervention studies conducted since 2021, only 12 (19%) cited the CONSORT-AI reporting guideline.[52] This low uptake illustrates the need not only for high-quality guidelines but also for awareness and enforcement within the research community, which could ensure better adherence and thereby consistent reporting practices. This change should prioritize transparency, quality, and exhaustive documentation over the rush to publish findings. Additionally, journals should take more responsibility and enforce reproducibility for future AI studies.[15] By doing so, they support efforts to establish a standard within the framework of reproducibility and promote sustainable and transparent research.

The "rigor of development" feature assesses whether the following components are clearly stated in the guidelines: a systematic evaluation of evidence synthesis, method of developing the guideline, explicit link between the guideline and the body of evidence, external expert revision of the developed guideline and the procedure to update or modify the guideline is clearly stated in the suggested guidelines.[27] Most of the identified guidelines have not considered a systematic synthesis of previous works. All guidelines have a justified rationale of their purpose and scope, whereas only half of them followed a standardized procedure of guideline development process such as the one suggested by the EQUATOR Network. However, not following a standard procedures for developing guidelines could result in compromised quality of guidelines.[53]

A recent publication[10] reviewed the contents of AI guidelines using translational stage of surveillance domains. The study showed that most guidelines discussed the importance of ethics, reproducibility, and transparency in AI studies but were less likely to engage relevant stakeholders such as patients, end users, and experts during the development process. This result is in line with our findings. To engage relevant stakeholders in the process of developing guidelines helps in converging efforts and maximize the utility and versatility of the guidelines.[54] Specifically, TRIPOD+AI, DECIDE-AI,

CONSORT-AI, and SPIRIT-AI guidelines involved a wide range of stakeholders during their development process, while other guidelines were developed by experts and researchers from different institutions without the engagement of potential stakeholders.

Applicability is another important gap in the identified guidelines. To ensure a guideline's applicability, it is essential to provide a comprehensive description of the factors that facilitate or hinder its application. This can be a detailed presentation of suggested tools and instructions for using the guidelines effectively. It is important to outline the resource implications of applying the guidelines. Furthermore, monitoring or auditing criteria should be explicitly presented to ensure the quality and adherence of a guideline.[27] CLEAR, which is the most recent reporting guideline for radiomics research,[39] followed by TRIPOD+AI were few of the most applicable reporting guidelines in our review. The issue of applicability is not limited to the quality of the guidelines but also limited to study design. For instance, most of the quality guidelines that are widely in use in reporting AI-related studies are focused on clinical trials or specific fields of study. In contrast, most of the studies in medical contexts applying AI methods are observational studies, which consequently creates a reporting gap in these types of studies. Thus, we strongly suggest that a reporting guideline for AI studies in medicine, irrespective of the study design, should be developed to enhance transparent reporting, reproducibility, and reusability, which ultimately contributes to improved health care. The journey toward complete reproducibility in AI research may be lengthy and complex, but it is a worthwhile endeavor. The rewards are not only for individual researchers but for the entire scientific community and society as a whole.

Other important aspects of AI applications in medicine are the moral dimensions such as bias, ethics, and governance, which are still prominent challenges strongly influencing the deployment of AI systems. A solution proposed by researchers is to embed AI ethics in the entire AI model development process.[55]

One aspect of AI that is usually overlooked is its environmental implications.[56] According to the characterizations of the carbon footprint of AI computing considering the lifecycle across large-scale use cases, the carbon emission to train an ML model is considerably high.[57,58] We suggest that future reporting guidelines should include a checklist that encompasses the moral and environmental aspects of AI studies as well.

Limitations

Using AGREE II for nonclinical guidelines has its own limitations. Since it is designed for clinical studies, some of the items in the evaluation metrics may not align perfectly with the nonclinical context. Currently, there is no quality assessment tool for measuring the quality of nonclinical practice guidelines involving AI. Therefore, we believe that the development of a quality assessment measure for nonclinical practice guidelines including reporting guidelines should be considered.

Conclusions

This study systematically assessed the quality of reporting guidelines for AI in medicine using the AGREE II framework. The result underlines the importance of transparent reporting in AI research, particularly in health care.

Although the identified guidelines provide valuable frameworks for reporting AI research, variability in their quality, particularly regarding applicability, stakeholder engagement, and rigor of development, highlights gaps that require attention. Our findings also underscore the critical importance of adherence to standardized development processes in creating robust and reliable reporting guidelines.

We conclude that future reporting guidelines should adopt a multidisciplinary approach, considering diverse study designs, open science practices, environmental impacts, and ethical considerations of medical AI research.

Fostering a culture of rigorous reporting and reproducibility will strengthen trust in AI-driven advancements. However, a significant cultural shift toward transparency is required. This shift should further be supported by journals enforcing the use of standardized reporting guidelines to enhance the reproducibility and reliability of medical AI studies. We acknowledge that the path toward comprehensive reproducibility is complex but essential. Researchers, institutions, and journals should collaboratively ensure that AI research prioritizes transparency, quality, and long-term sustainability, benefiting both science and society.

Acknowledgments

K.B.S. would like to thank the DAAD (German Academic Exchange) and A.A.Z. acknowledge the National Research Data Infrastructure for Personal Health Data (NFDI4Health).

Author contributions

Kirubel Biruk Shiferaw, Dagmar Waltemath, and Atinkut Alamirrew Zeleke contributed substantially to the conception, methodology, and writing of this work. Moritz Roloff, Irina Balaur, and Danielle Welter contributed significantly in methodology, writing, and revising the manuscript. All authors revised this manuscript critically for important intellectual content and approved the version to be published.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

K.B.S. was funded by the DAAD (German Academic Exchange) for supporting the doctoral research study expenses. A.A.Z. was funded by the National Research Data Infrastructure for Personal Health Data (NFDI4Health) DFG-funded project (Project 442326535).

Conflicts of interest

The authors have no competing interest to declare.

Data availability

All data generated or analyzed during this study are included in this article. The python code for generating the plots can be found here: https://github.com/kirubel-Biruk-Shiferaw/Guidelines-and-Standard-Frameworks-for-Artificial-Intelligence-in-Medicine-A-Systematic-Review

References

1. Samoili S, Cobo ML, Gómez E, et al. AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence. Publications Office; 2020. https://doi.org/10.2760/382730
2. Gruetzemacher R, Whittlestone J. The transformative potential of artificial intelligence. *Futures*. 2022;135:102884.
3. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2:230-243.
4. Păvăloaia V-D, Necula S-C. Artificial intelligence as a disruptive technology—a systematic literature review. *Electronics*. 2023;12:1102.
5. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med*. 2022;28:31-38.
6. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25:30-36.
7. Shiferaw KB, Wali P, Waltemath D, et al. Navigating the AI frontiers in cardiovascular research: a bibliometric exploration and topic modeling. *Front Cardiovasc Med*. 2023;10:1308668.
8. Topol EJ. Welcoming new guidelines for AI clinical research. *Nat Med*. 2020;26:1318-1320.
9. WHO. WHO calls for safe and ethical AI for health. Updated 2023. Accessed March 7, 2023. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health.
10. Crossnohere NL, Elsaid M, Paskett J, et al. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. *J Med Internet Res*. 2022;24:e36823.
11. Gama F, Tyskbo D, Nygren J, et al. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. *J Med Internet Res*. 2022;24:e32215.
12. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: applied AI and machine learning in translational medicine. *EBioMedicine*. 2019;47:607-615.
13. Kanbach DK, Heiduk L, Blueher G, et al. The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Rev Manag Sci*. 2024;18:1189-1220.
14. Baker M. Reproducibility crisis. *Nature*. 2016;533:353-366.
15. Haibe-Kains B, Adam GA, Hosny A, et al.; Massive Analysis Quality Control (MAQC) Society Board of Directors. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586:E14-E16.
16. Gundersen OE, Kjensmo S. State of the art: reproducibility in artificial intelligence. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA. 2018.
17. Celi LA, Citi L, Ghassemi M, et al. The PLOS ONE collection on machine learning in health and biomedicine: towards open code and open data. *PLoS One*. 2019;14:e0210232.
18. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022;5:2.
19. Kolbinger FR, Veldhuizen GP, Zhu J, et al. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)*. 2024;4:71.
20. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol*. 2021;49:470-476.
21. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
22. Chue Hong NP, Katz DS, Barker M, et al. FAIR principles for research software (FAIR4RS principles). Zenodo; 2022.
23. Huerta EA, Blaiszik B, Brinson LC, et al. FAIR for AI: an interdisciplinary and international community building perspective. *Sci Data*. 2023;10:487.
24. Harrow I, Balakrishnan R, Küçük McGinty H, et al. Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. *Drug Discov Today*. 2022;27:1441-1447.
25. Barker M, Chue Hong NP, Katz DS, et al. Introducing the FAIR principles for research software. *Sci Data*. 2022;9:622.
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg*. 2021;88:105906.
27. Brouwers MC, Kho ME, Browman GP, et al.; AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ*. 2010;182:E839-E842.
28. Shiferaw KB, Roloff M, Waltemath D, et al. Guidelines and standard frameworks for AI in medicine: protocol for a systematic literature review. *JMIR Res Protoc*. 2023;12:e47105.
29. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210.
30. Collins GS, Reitsma JB, Altman DG, et al.; TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*. 2015;131:211-219.
31. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7:e1000217.
32. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digital Health*. 2020;2:e549-e560.
33. Haneef R, Tijhuis M, Thiébaut R, et al. Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques. *Arch Public Health*. 2022;80:9.
34. Liu X, Cruz Rivera S, Moher D, et al.; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digital Health*. 2020;2:e537-e548.
35. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
36. Bossuyt PM, Reitsma JB, Bruns DE, et al.; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277:826-832.
37. Vasey B, Nagendran M, Campbell B, et al.; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904.
38. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320-1324.
39. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14:75.
40. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
41. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. *JAMA Netw Open*. 2023;6:e2335377.
42. Wang Y, Li N, Chen L, et al. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: systematic review. *J Med Internet Res*. 2023;25:e46089.
43. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155-163.
44. Belbasis L, Panagiotou OA. Reproducibility of prediction models in health services research. *BMC Res Notes*. 2022;15:204-205.
45. Heil BJ, Hoffman MM, Markowetz F, et al. Reproducibility standards for machine learning in the life sciences. *Nat Methods*. 2021;18:1132-1135.
46. Hauschild A-C, Eick L, Wienbeck J, et al. Fostering reproducibility, reusability, and technology transfer in health informatics. *Iscience*. 2021;24:102803.

47. Mangul S, Martin LS, Eskin E, Blekhman R. Improving the usability and archival stability of bioinformatics software. *BioMed Central*. 2019:1-3.
48. Mateen BA, Liley J, Denniston AK, et al. Improving the quality of machine learning in health applications and clinical research. *Nat Mach Intell*. 2020;2:554-556.
49. Pineau J, et al. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *J Mach Learn Res*. 2021;22:7459-7478.
50. Ibrahim H, Liu X, Rivera SC, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*. 2021;22:11-15.
51. Guni A, Sounderajah V, Whiting P, et al. Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): protocol for a qualitative study. *JMIR Res Protoc*. 2024;13:e58202.
52. Han R, Acosta JN, Shakeri Z, et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health*. 2024;6:e367-e373.
53. Simera I, Moher D, Hoey J, et al. The EQUATOR Network and reporting guidelines: helping to achieve high standards in reporting health research studies. *Maturitas*. 2009;63:4-6.
54. Deshpande A, Sharp H. Responsible AI systems: who are the stakeholders? In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, Oxford, United Kingdom. 2022.
55. McLennan S, Fiske A, Tigard D, et al. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Med Ethics*. 2022;23:6.
56. Wu CJ, Raghavendra R, Gupta U, et al. Sustainable AI: environmental implications, challenges and opportunities. *Proc Mach Learn Syst*. 2022;4:795-813.
57. Dhar P. The carbon impact of artificial intelligence. *Nat Mach Intell*. 2020;2:423-425.
58. Shumskaia EI. Artificial intelligence—reducing the carbon footprint? In: Zavyalova, EB, Popkova, EG, eds. *Industry 4.0: Fighting Climate Change in the Economy of the Future*. Springer International Publishing; 2022:359-365.

JAMIA Open, 2025, 8, 1–10
https://doi.org/10.1093/jamiaopen/ooae155
Review