Protocol

Guidelines and Standard Frameworks for AI in Medicine: Protocol for a Systematic Literature Review

Kirubel Biruk Shiferaw[*], MPH; Moritz Roloff[*], BSc; Dagmar Waltemath[*], PhD; Atinkut Alamirrew Zeleke[*], PhD

Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

[*]all authors contributed equally

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
Phone: 49 1728989478
Email: s-kishif@uni-greifswald.de

Abstract

Background: Applications of artificial intelligence (AI) are pervasive in modern biomedical science. In fact, research results suggesting algorithms and AI models for different target diseases and conditions are continuously increasing. While this situation undoubtedly improves the outcome of AI models, health care providers are increasingly unsure which AI model to use due to multiple alternatives for a specific target and the "black box" nature of AI. Moreover, the fact that studies rarely use guidelines in developing and reporting AI models poses additional challenges in trusting and adapting models for practical implementation.

Objective: This review protocol describes the planned steps and methods for a review of the synthesized evidence regarding the quality of available guidelines and frameworks to facilitate AI applications in medicine.

Methods: We will commence a systematic literature search using medical subject headings terms for medicine, guidelines, and machine learning (ML). All available guidelines, standard frameworks, best practices, checklists, and recommendations will be included, irrespective of the study design. The search will be conducted on web-based repositories such as PubMed, Web of Science, and the EQUATOR (Enhancing the Quality and Transparency of Health Research) network. After removing duplicate results, a preliminary scan for titles will be done by 2 reviewers. After the first scan, the reviewers will rescan the selected literature for abstract review, and any incongruities about whether to include the article for full-text review or not will be resolved by the third and fourth reviewer based on the predefined criteria. A Google Scholar (Google LLC) search will also be performed to identify gray literature. The quality of identified guidelines will be evaluated using the Appraisal of Guidelines, Research, and Evaluation (AGREE II) tool. A descriptive summary and narrative synthesis will be carried out, and the details of critical appraisal and subgroup synthesis findings will be presented.

Results: The results will be reported using the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) reporting guidelines. Data analysis is currently underway, and we anticipate finalizing the review by November 2023.

Conclusions: Guidelines and recommended frameworks for developing, reporting, and implementing AI studies have been developed by different experts to facilitate the reliable assessment of validity and consistent interpretation of ML models for medical applications. We postulate that a guideline supports the assessment of an ML model only if the quality and reliability of the guideline are high. Assessing the quality and aspects of available guidelines, recommendations, checklists, and frameworks—as will be done in the proposed review—will provide comprehensive insights into current gaps and help to formulate future research directions.

International Registered Report Identifier (IRRID): DERR1-10.2196/47105

(JMIR Res Protoc 2023;12:e47105) doi: 10.2196/47105

KEYWORDS

artificial intelligence; biomedical; guidelines; machine learning; medicine

Introduction

Rationale

Artificial intelligence (AI) has become a hot topic in biomedical and clinical routines in the past decade [1,2]. In fact, an increasing number of scientific publications suggest algorithms and AI models for target diseases and conditions [3]. The application of AI in health care ranges from medical image analysis to text mining, targeting prognosis, diagnosis, event and outcome prediction, and treatment [1]. The number of scientific contributions in peer-reviewed journals focusing on AI applications in medicine is also highly increasing [4,5].

The comparison of algorithm performance for a target condition has become a common way of presenting and evaluating machine learning (ML) models. The diversity of ML models leads to competition and fast progress. However, the current situation leaves health care providers unsure about which ML model to use given multiple alternative models for a specific target [6] and the “black box” nature of ML [7]. Moreover, the fact that studies do not use guidelines consistently in developing and reporting AI models poses another challenge in trusting and adapting models for practical implementation [8-10].

In fact, guidelines and recommended frameworks for developing and reporting AI studies have been published by different experts and work groups to facilitate reliable assessment of model validity and consistent interpretation [11-14]. In addition, the authors have proposed recommendations for the evaluation of AI studies [15-18]. This review protocol describes the planned steps and methods of a future review that aims to synthesize evidence regarding the quality of available guidelines and frameworks developed to facilitate AI applications in medicine and summarize their content.

Objective

The systematic review will address the following scientific questions:

- What are the available guidelines and frameworks with regard to predictive AI model development and reporting in medicine?
- What are the main aspects (target or purpose) of the available guidelines and frameworks?
- What is the quality of the available guidelines with respect to the Appraisal of Guidelines, Research and Evaluation (AGREE II) domains?
- What are the reported implementation challenges in the available guidelines?

Methods

Eligibility Criteria

All guidelines, standard frameworks, best practices, checklists, and recommendations will be included, irrespective of the study design. Studies will be limited to English and to the time period between database inception and June 2023 (the time of data collection).

Search Strategy and Information Sources

This protocol adheres to the PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) 2015 [19]. A systematic literature search will be commenced using Medical Subject Headings (MeSH) terms and keywords for medicine, guidelines, and ML (Table S1 in Multimedia Appendix 1). The initial search will be conducted on 3 web-based repositories: PubMed, Web of Science, and EQUATOR (Enhancing the Quality and Transparency of Health Research). The EQUATOR network is a global initiative working toward improving the value of research by promoting robust reporting guidelines [20]. A Google Scholar search will also be performed to identify gray literature. References in selected literature will be scanned, and relevant papers will be included after discussion among the reviewers.

After the comprehensive search, all bibliographic information will be uploaded to a web-based systematic review tool (Rayyan) and afterward processed with CADIMA [21] for further screening and preliminary analysis.

Study Selection

After removing duplicate results, a preliminary scan for titles will be done by 2 reviewers. After the initial scan, the reviewers will rescan the selected literature for abstract review, and any incongruities about whether to include the article for full-text review will be resolved by the third reviewer based on the predefined criteria.

Data Extraction, Collection, and Management

Once the selection process has been finalized, relevant information from the selected literature will be extracted by the 2 reviewers independently using a predefined information extraction sheet. At this stage, discrepancies will be resolved by the third and fourth reviewers.

The information extraction sheet will be designed to collect relevant information such as study characteristics (authors, year of publication, study type, aspect, specific disease or condition focused, and standard followed) from selected literature for further synthesis. “Study type” in this context refers to whether the study is a guideline, a framework, a suggestion, a checklist, a best practice, or a recommendation. The term “aspect” entails the purpose for which the guideline or the framework is primarily designed. Guidelines or frameworks could be designed for different purposes. For instance, if a guideline is developed to elaborate on or declare how ML studies should report their findings, the aspect or purpose will be a “reporting aspect.” If, however, the guideline emphasizes the ethical dimension of ML in medicine, the aspect will be referred to as “ethical aspect.” A study can have multiple aspects (sometimes a guideline or a framework could be about both the development and reporting procedures and sometimes only the reporting or the development procedure), and it will also be extracted under the “aspect” category. The column “specific disease or condition focused” refers to whether the guideline or framework is subject to or validated for a specific disease or condition. Sometimes, researchers suggest customized and focused guidelines for a specific condition. Based on the predefined data extraction Excel (Microsoft Corp) sheet, 2 reviewers will extract the data independently, and any sort of discrepancy will be discussed with the 3rd and 4th reviewers after every 10 extractions. The data extraction template is presented in Table 1.

Table 1. Data extraction template.

Sections	Description
Section 1: study characteristics	
First author's name	For example, John Smith
Year of publication	For example, YYYY
Title	The title of the study
Journal	The name of the journal. For example, scientific reports
Country of first author	For example, United States of America
Section 2: research questions	
Type of outcome	Provide a description about the study whether it is a guideline, standard framework, recommendation, checklist, best practice, or expert opinion.
Aspect	Provide the primary purpose for which the guideline or the framework is designed. For example, reporting, development, ethics, and governance.
Standard followed	Provide a description of the standard followed during the development of the proposed guideline or framework. For example, the EQUATOR[a] network.
Guidelines and frameworks with regard to predictive artificial intelligence model development and reporting in medicine	Provide a description of the content and domain of the guideline or framework.
The quality of the available guidelines	Provide a quantified quality assessment of the identified guidelines based on AGREE II[b] assessment.
Reported gaps in the available guidelines	Describe the reported gaps and implementation challenges.
Specific target domain	Provide information if the guideline or framework is designed for a specific disease domain.

[a] EQUATOR: Enhancing the Quality and Transparency of Health Research.

[b] AGREE II: Appraisal of Guidelines, Research, and Evaluation.

Quality and Risk of Bias Assessment

The quality of identified guidelines will be evaluated using the AGREE II tool [22]. AGREE II assesses the quality of guidelines in terms of the methodological rigorousness and transparency of the guideline development process [22]. It consists of 23 key items organized within 6 domains (scope and purpose, stakeholder involvement, rigor of development, clarity of presentation, applicability, and editorial independence) and 2 global rating items for overall assessment. As recommended by the AGREE II user manual, 4 appraisers will individually appraise the selected guidelines and frameworks, and the percentage of quality will be calculated based on the ratings of each quality domain [22]. Other frameworks and best practices proposed by researchers will be systematically summarized.

Analysis

A descriptive summary and narrative synthesis will be carried out, and the details of critical appraisal and subgroup synthesis findings will be presented. This systematic review will present the list of available guidelines, frameworks, suggestions, checklists, best practices, or recommendations for developing and reporting AI studies in medicine and assess the quality of the guidelines and discuss the aspects.

Results

We anticipate finishing the systematic literature review by November 2023. The expected result from this review will explore and highlight the quality of the available guidelines and point out the existing gaps. Furthermore, the synthesis from the checklists and recommendations will provide a comprehensive set of standardized tools.

Discussion

Guidelines are important sets of instructions that facilitate transparent and reproducible scientific processes [23]. The potential benefits of guidelines are, however, only as good as the quality of the guidelines themselves. This protocol describes the setup of a systematic review to identify available guidelines, frameworks, best practices, and recommendations for developing and reporting AI studies in medicine. The review will assess the quality of guidelines, summarize the gap in guidelines, and identify critical considerations in the application of AI in medicine. As the search is limited to English due to resource limitations, there might be language bias, indicating that some important studies conducted in different languages might be missed. Thus, conclusions should be drawn carefully.

Acknowledgments

The authors would like to acknowledge the German Academic Exchange Service (DAAD) for funding the research expenses. The research expense of coauthor AAZ is funded by NFDI4Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Keywords.
[DOCX File , 13 KB-Multimedia Appendix 1]

References

1. Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: a review. : IEEE; 2018 Presented at: 2018 Second International Conference of Electronics, Communication and Aerospace Technology (ICECA); 29-31 March 2018; Coimbatore, India [doi: 10.1109/iceca.2018.8474918]
2. Verma VK, Verma S. Machine learning applications in healthcare sector: an overview. Mater Today: Proc 2022;57:2144-2147 [doi: 10.1016/j.matpr.2021.12.101]
3. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. J Med Internet Res 2020;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
4. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
5. Shiferaw KB, Waltemath D, Zeleke A. Disparities in regional publication trends on the topic of artificial intelligence in biomedical science over the last five years: a bibliometric analysis. Stud Health Technol Inform 2022;294:609-613 [doi: 10.3233/SHTI220541] [Medline: 35612161]
6. Gilbank P, Johnson-Cover K, Truong T. Designing for physician trust: toward a machine learning decision aid for radiation toxicity risk. Ergon Des 2019;28(3):27-35 [doi: 10.1177/1064804619896172]
7. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886 [doi: 10.1136/bmj.l886] [Medline: 30862612]
8. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med 2020;10(2):21 [FREE Full text] [doi: 10.3390/jpm10020021] [Medline: 32244292]
9. Barrett L, Hu J, Howell P. Systematic review of machine learning approaches for detecting developmental stuttering. IEEE/ACM Trans Audio Speech Lang Process 2022;30:1160-1172 [doi: 10.1109/taslp.2022.3155295]
10. Shiferaw KB, Zeleke A, Waltemath D. Assessing the FAIRness of deep learning models in cardiovascular disease using computed tomography images: data and code perspective. Stud Health Technol Inform 2023;302:63-67 [doi: 10.3233/SHTI230065] [Medline: 37203610]
11. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
12. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
13. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340 [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55-63 [FREE Full text] [doi: 10.7326/M14-0697] [Medline: 25560714]
15. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020;13(10):e006556 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.120.006556] [Medline: 33079589]
16. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. JACC Cardiovasc Imaging 2020;13(9):2017-2035 [FREE Full text] [doi: 10.1016/j.jcmg.2020.07.015] [Medline: 32912474]
17. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent 2021;107:103610 [doi: 10.1016/j.jdent.2021.103610] [Medline: 33631303]

18. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
19. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015;350:g7647 [FREE Full text] [doi: 10.1136/bmj.g7647] [Medline: 25555855]
20. Enhancing the quality and transparency of health research. EQUATOR network. URL: https://www.equator-network.org/ [accessed 2023-09-25]
21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
22. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010;182(18):E839-E842 [FREE Full text] [doi: 10.1503/cmaj.090449] [Medline: 20603348]
23. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7(2):e1000217 [FREE Full text] [doi: 10.1371/journal.pmed.1000217] [Medline: 20169112]

Abbreviations

AGREE II: Appraisal of Guidelines, Research and Evaluation
AI: artificial intelligence
EQUATOR: Enhancing the QUAlity and Transparency Of health Research
MeSH: medical subject headings
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols

Edited by A Mavragani; submitted 08.03.23; peer-reviewed by D Kohen, T Hou; comments to author 26.05.23; revised version received 09.06.23; accepted 20.06.23; published 25.10.23

Please cite as:
Shiferaw KB, Roloff M, Waltemath D, Zeleke AA
Guidelines and Standard Frameworks for AI in Medicine: Protocol for a Systematic Literature Review
JMIR Res Protoc 2023;12:e47105
URL: https://www.researchprotocols.org/2023/1/e47105
doi: 10.2196/47105
PMID: 37878365