Original Paper

Data Availability for Critical Appraisal Tools for Evaluating Artificial Intelligence in Clinical Studies: Dataset for a Scoping Review

Juan Bautista Cabello López[1], MD, PhD; Vicente Ruiz García[2], MD, PhD; Miguel Torralba[3], MD, PhD; Miguel Maldonado Fernandez[4], MSc, MPH, MD, PhD; Marimar Úbeda-Carrillo[5], MSc; Eukene Ansuategi[6], MSc; Luis Ramos-Ruperto[7], MSc, MD; José Ignacio Emparanza[8], MD, PhD; Iratxe Urreta-Barallobre[8,9], MD, PhD; María-Teresa Iglesias Gaspar[8,10], PhD; José Ignacio Pijoan Zubizarreta[11], MD, PhD; Amanda J Burls[12], MBBS, MSc

[1]CASP España & Grupo Medicine AI, Alicante, Spain
[2]Servicio de Hospitalización a Domicilio, Hospital Universitari i Politècnic La Fe, Valencia, Spain
[3]Servicio de Medicina Interna (Infecciosas), Hospital Universitario de Guadalajara, Guadalajara, Spain
[4]Hospital Vital Álvarez-Buylla, Mieres, Spain
[5]Library Service, Donostia University Hospital, Osakidetza Basque Health Service, San Sebastián, Spain
[6]Biblioteca virtual de salud de Euskadi, Vitoria, Spain
[7]Hospital Universitario La Paz, Madrid, Spain
[8]Clinical Epidemiology Unit, Donostia University Hospital, Osakidetza Basque Health Service, San Sebastián, Spain
[9]CIBER de Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain
[10]Faculty of Health Sciences, University of Deusto, San Sebastián, Spain
[11]Hospital de Cruces, Bilbao, Spain
[12]City St George's, University of London, London, United Kingdom

Corresponding Author:

Miguel Maldonado Fernandez, MSc, MPH, MD, PhD
Hospital Vital Álvarez-Buylla
Calle Vistalegre, 2
Mieres 33611
Spain
Phone: 34 652835133
Email: maldonado2000@gmail.com

Abstract

Background: AI is increasingly used in clinical research, generating a growing need for robust critical appraisal tools to evaluate methodological quality, reporting standards, and potential biases. While traditional instruments exist for conventional clinical studies, specific tools designed for AI-based research are still emerging.

Objective: This dataset accompanies a scoping review that aimed to identify and describe existing critical appraisal tools, reporting frameworks, and bias classification systems applicable to clinical studies using AI, including chatbot-based interventions.

Methods: We systematically searched MEDLINE, Embase, CINAHL, PsycINFO, and IEEE Xplore from inception to April 2024. Eligible studies included those proposing or using tools for critical appraisal, reporting, quality assessment, or risk of bias in AI-related clinical research. Screening and extraction followed JBI and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) recommendations. Data extraction combined human review with a supervised GPT-4o–based retrieval-augmented generation (RAG) process to enhance transparency and reproducibility. All AI-assisted outputs were verified independently by two reviewers.

Results: Seventy records were included: 46 reporting guidelines (comprising 26 guides for reporting AI studies, 16 critical appraisal tools, 2 quality assessment instruments, and 2 risk-of-bias tools), 9 bias classification or bias mitigation studies, and 15 chatbot evaluation studies. All datasets, extraction templates, and RAG prompts are publicly available to facilitate validation and reuse.

Conclusions: This dataset provides a comprehensive overview of critical appraisal and reporting tools for AI-based clinical research. It may support the development of standardized evaluation frameworks and promote transparency in future AI-assisted health studies.

Trial Registration: OSF Registries 10.17605/OSF.IO/ETYDS; https://osf.io/etyds/overview

JMIR Data 2026;7:e85688; doi: 10.2196/85688

Keywords: artificial intelligence; critical appraisal; bias; data sharing; scoping review; reporting guidelines; chatbot

Introduction

The use of predictive or generative AI in health research is rapidly growing [1-6]. As in traditional clinical studies, the methods used in AI-assisted studies can introduce systematic errors. The translation of AI-assisted evidence into clinical practice and research requires critical appraisal tools for clinical decision-makers and researchers [7-10]. We carried out a scoping review [11] to identify existing tools for the critical appraisal of clinical studies that use AI and to examine the concepts and domains these tools explore. Our research question is framed using the PCC (population, concept, and context) framework [12,13]: the population includes clinical studies using AI; the concept refers to tools for critical appraisal and associated constructs such as quality, reporting, validity, risk of bias, and applicability; and the context is clinical practice. In addition, bias classification and chatbot assessment studies were included.

A total of 70 records were included in the review, comprising a heterogeneous group. Of these, 46 were tools (26 guides for reporting AI studies, 16 tools for critical appraisal, 2 tools for study quality, and 2 tools for risk of bias), 9 were papers focused on bias classification or mitigation, and 15 were chatbot studies (6 chatbot assessment studies and 9 systematic reviews of chatbot studies).

Just as traditional health research may include systematic errors that lead to biased or nongeneralizable results, so AI methods can introduce their own systematic errors at the design, data collection, training, or evaluation stages, which threaten the validity and reliability of AI models' data analyses, findings, and conclusions. Such errors can arise from a number of different sources, including, but not limited to, flawed data, biased algorithms, and incorrect training. Health care decision-makers therefore need to be able to critically appraise AI studies to detect these problems so they can assess the certainty and relevance of the presented evidence.

Access to this dataset will help systematize the available tools and allow other researchers to compare and select appropriate ones. The dataset will also be of use for teaching purposes and risk-of-bias evaluation. In addition, we believe that it is good scientific practice to share data.

Our objective is to make the primary data used in our scoping review publicly available.

Methods

Overview

We searched medical and engineering databases (MEDLINE, Embase, CINAHL, PsycINFO, and IEEE) from inception to April 2024. We included primary clinical research that used tools for critical appraisal. Classic reviews and systematic reviews were included in the first phase of screening and used to identify new tools by forward snowballing. They were excluded in the second phase. We excluded nonhuman, computer, and mathematical research and letters, opinion papers, and editorials. We used Rayyan for screening [14].

The protocol was previously registered with OSF [15]. We adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Checklist 1) [16] and the PRISMA-S (Preferred Reporting Items for Systematic Reviews Literature Search Extension) for reporting literature in systematic reviews [17].

Data extraction for tools and bias was performed according to JBI recommendations [18], and each study was rated by two observers. Discrepancies were resolved by discussion and consensus in an iterative process. (Data extraction tables and guidance are shown in Multimedia Appendix 1.)

Data extraction for chatbot studies used a hybrid approach, combining the active involvement of a researcher with a retrieval-augmented generation (RAG) approach using custom ChatGPT instances based on the GPT-4o model, accessed directly through the ChatGPT web interface, with default model parameters (the RAG prompt template in Spanish is available in Spanish in Multimedia Appendix 2 and in English in Multimedia Appendix 3). Each PDF was uploaded and analyzed as an independent prompt, and all extracted information was subsequently reviewed by the primary investigator and independently validated by a second investigator to minimize hallucinations and ensure that the large language model was used solely as a facilitative extraction tool. No sensitive data were exposed. To promote transparency and reproducibility, the exact prompts used in the RAG process are incorporated into this dataset.

As an example to explain the process, the systematic review approach with RAG (in Spanish) was used to extract data from Oh et al [19]. A PDF was uploaded to ChatGPT (GPT-4o), and the output was provided in a PDF file (Multimedia Appendix 4). The result was incorporated as a column in the rev_sis extraction XLS file, which can be found in column H in the dataset [11]. This procedure was repeated

for all systematic review articles and for all primary research articles (rev_sis extraction and primary studies extraction, respectively). Both XLS files were transposed from columns to rows (matrix transposition) to build the final XLS files (S reviews table draft and primary studies table draft, respectively). The results were translated from Spanish to English by the reviewers. Both table drafts were reviewed and their results compared with the actual papers by LR-R and afterward verified by JBCL. The final tables were synthesized as shown in the scoping review, where they appear as Tables 4 and 5 [11].

We identified 4392 records in the selected databases and registries. After eliminating 470 duplicates, 3922 records were screened by title and abstract, and 3803 were excluded. The remaining 119 underwent full-text screening, and 59 were excluded. The reasons for exclusion were as follows: 50 studies were systematic reviews, 7 met the exclusion criteria, and 2 did not meet the inclusion criteria. Full details are available in Multimedia Appendix 5. Of the 50 systematic reviews, 42 used specific AI tools to assess the quality of the reviewed studies, and the tools retrieved were incorporated into the "records identified via other methods" category.

Twelve studies were identified in the EQUATOR Network library, and 4 additional studies were obtained from experts and organizations; therefore, 58 records were identified by other methods. Forty-eight of these were already captured in the 60 included studies from the search of electronic databases, leaving 10 additional studies to be included.

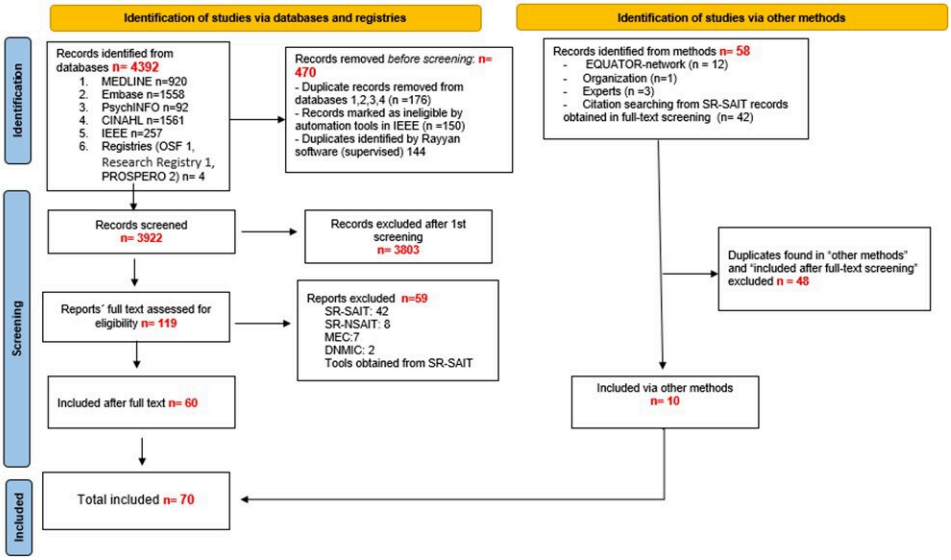
Ethical Considerations

Institutional review board approval was not applicable to this study and dataset. The scoping review built on this dataset has been accepted for publication in the *Journal of Medical Internet Research*.

Results

We identified 4392 records in databases and registries. After the selection process and inclusion of additional studies described in the Methods section, a total of 70 studies were included in the review (see scoping review [11]; Figure 1). Forty-six records reported tools (26 guidelines, 16 tools for critical appraisal, 2 tools for study quality, and 2 tools for risk of bias), and 9 were focused on bias classification or mitigation (Multimedia Appendix 6). In addition, 15 records were chatbot studies, comprising 6 chatbot assessment studies and 9 systematic reviews of chatbot studies (Multimedia Appendix 7).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. DNMIC: does not meet inclusion criteria; MEC: methodological/editorial/commentary; SR-NSAIT: systematic review—nonstudies assessing AI tools; SR-SAIT: systematic review—studies assessing AI tools.



Discussion

Overview

As explained in our main paper [11], we conducted a comprehensive scoping review and identified 70 papers corresponding to the 3 proposed areas of research: tools for critical appraisal, bias and bias mitigation, and chatbot assessment studies. Although critical appraisal tools are the main focus of the review, types of AI bias were also included in the review because the validity (or absence of bias) is an important component of critical appraisal. Chatbot studies were included in the review because they represent an important recent, disruptive technology. The three areas together map the current landscape of evidence in the critical appraisal of clinical AI studies.

We selected critical appraisal as the main domain for the review because it is a wider and more inclusive concept than risk of bias, quality, or reporting, and it is more related to clinical practice. This decision required a change to the published protocol and was made after discussion.

Reporting guidelines are essential for authors in writing studies and for editors in maintaining consistency across

publications. Critical appraisal tools are more focused on making judgments about the validity and applicability of evidence, and they are mainly intended for dissemination and teaching purposes. A paper may be of no use in a clinical setting, even if it is perfectly reported and its data are valid. Finally, both quality and risk of bias are precise concepts, and their tools are complex and designed as far as possible to avoid inconsistencies. Such tools are more suitable for use in research syntheses. Nevertheless, reporting, critical appraisal, risk of bias, and quality form a cluster of closely related constructs with overlapping areas.

Adequate reporting varies by the structure and type of study and is not only an editorial requirement but a part of study quality. Obviously, good reporting is a precondition to assess study quality, but there is also empirical evidence that some reporting flaws (and some nonreporting flaws) are associated with bias in the effect estimation [20,21]. Therefore, exploring reporting is essential to judge the validity of any study, as it facilitates study replication, risk-of-bias or quality assessments, interpretation of the results, and judgment of the value and applicability of the results in real clinical settings for individualized or collective decisions. It is also needed for the inclusion and assessment of studies in systematic reviews and for the evaluation of systematic reviews. Therefore, it is part of the critical appraisal process [8,22].

The overlap of reporting and critical appraisal was a source of inconsistency between raters when classifying papers in this scoping review. Iterative discussions were necessary to reach consensus. The most important criteria we used to classify papers within the critical appraisal category were the relevance of the question in the clinical context and a clear intent to help with applicability.

On the other hand, chatbot assessment studies are heterogeneous and inconsistent in their design, analysis, and reporting, so we used ChatGPT (GPT-4o) for data extraction; however, all outputs were independently reviewed by two authors against the original articles, and no major corrections to the extracted information were required. Therefore, we believe this was a valid and consistent procedure for data extraction. Nevertheless, we include the RAG and instructions for matrix transposition to enable other researchers to replicate the process.

A recent systematic review [23] synthesized reporting guidelines as well as tools for basic and laboratory research. However, the search was conducted only through 2022. The available reporting guidelines should be harmonized, and the review would benefit from being updated or reformulated from a clinical standpoint.

There are some limitations of this scoping review. First, we used a general search that included all of our study's questions; it was not specifically designed to search for bias and bias mitigation or for chatbot assessment. However, the absence of MeSH for chatbot studies and the heterogeneity of objectives, research questions, study design, devices, and analyses make searches for this type of study difficult. In addition, a potential limitation lies in the methods used to organize data extraction, as the application of large language models in evidence synthesis is novel, and formal standards for their integration are still under development. Finally, this field is evolving very quickly, so many conclusions drawn from existing evidence have a limited period of validity.

Critical appraisal tools are enormously varied, with different nuances and approaches, so selecting one can be very challenging. We believe that this topic deserves a qualitative synthesis to clarify the key elements for choosing the appropriate tool.

New risk-of-bias tools for AI in prognosis and diagnosis (such as QUADAS-AI [Quality Assessment of Diagnostic Accuracy Studies Using AI] and PROBAST+AI [Prediction Model Risk of Bias Assessment Tool]) and the PRISMA-AI (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Artificial Intelligence Extension) for systematic reviews are expected to be published, as is CHART (Chatbot Assessment Reporting Tool), a tool for reporting chatbot assessment studies. The AI extensions of other classic tools, such as the Cochrane risk-of-bias tool and ROBINS-I (Risk of Bias in Non-Randomized Studies of Interventions), among others, should be considered. On the other hand, the development of standards for the design, reporting, and assessment of chatbot assessment studies and chatbot health-advising studies is a clear gap in our toolbox and needs to be addressed.

In clinical practice, it is important to clarify the appropriate selection of tools for critical appraisal; furthermore, it is essential to develop teaching strategies to promote skills for the critical appraisal of AI-assisted studies, including understanding the types of bias to be tackled.

Conclusions

This dataset may be useful for other researchers who want to corroborate or further develop our results about critical appraisal tools for clinical studies that use AI.

Acknowledgments

Although we used ChatGPT (GPT-4o) for data extraction, we did not use generative AI for manuscript production.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data extraction template.
[DOCX File (Microsoft Word File), 23 KB-Multimedia Appendix 1]

Multimedia Appendix 2
ChatGPT retrieval-augmented generation prompt template (Spanish).
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 2]

Multimedia Appendix 3
ChatGPT retrieval-augmented generation prompt template (English).
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 3]

Multimedia Appendix 4
Example of ChatGPT output.
[PDF File (Adobe File), 278 KB-Multimedia Appendix 4]

Multimedia Appendix 5
Exclusions after screening.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 5]

Multimedia Appendix 6
Included studies reporting critical appraisal tools and bias mitigation.
[XLSX File (Microsoft Excel File), 160 KB-Multimedia Appendix 6]

Multimedia Appendix 7
Included chatbot studies.
[XLSX File (Microsoft Excel File), 31 KB-Multimedia Appendix 7]

Checklist 1 PRISMA-ScR checklist
[DOCX File (Microsoft Word File), 111 KB-Checklist 1 PRISMA-ScR checklist]

References

1. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. Oct 2020;92(4):807-812. [doi: 10.1016/j.gie.2020.06.040] [Medline: 32565184]
2. Kohane IS. Injecting artificial intelligence into medicine. NEJM AI. Jan 2024;1(1). [doi: 10.1056/AIe2300197]
3. Jayakumar S, Sounderajah V, Normahani P, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digit Med. Jan 27, 2022;5(1):11. [doi: 10.1038/s41746-021-00544-y] [Medline: 35087178]
4. Quirk J, Mac Donnchadha C, Vaantaja J, et al. Future implications of artificial intelligence in lung cancer screening: a systematic review. BJR Open. Oct 15, 2024;6(1):tzae035. [doi: 10.1093/bjro/tzae035] [Medline: 39444460]
5. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. Mar 2020;46(3):383-400. [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
6. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell. 2023;6:1237704. [doi: 10.3389/frai.2023.1237704] [Medline: 38028668]
7. Barker TH, Stone JC, Sears K, et al. Revising the JBI quantitative critical appraisal tools to improve their applicability: an overview of methods and the development process. JBI Evid Synth. Mar 1, 2023;21(3):478-493. [doi: 10.11124/JBIES-22-00125] [Medline: 36121230]
8. Moher D. Reporting guidelines: doing better for readers. BMC Med. Dec 14, 2018;16(1):233. [doi: 10.1186/s12916-018-1226-0] [Medline: 30545364]
9. 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. Jan 6, 2021;22(1):11. [doi: 10.1186/s13063-020-04951-6] [Medline: 33407780]
10. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res. Aug 25, 2022;24(8):e36823. [doi: 10.2196/36823] [Medline: 36006692]

11. Cabello JB, Ruiz Garcia V, Torralba M, et al. Critical appraisal tools for evaluating artificial intelligence in clinical studies: scoping review. J Med Internet Res. Dec 8, 2025;27:e77110. [doi: 10.2196/77110] [Medline: 41359958]
12. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. Sep 20, 2010;5(1):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
13. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. JBI; 2024. [doi: 10.46658/JBIMES-24-01]
14. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. Dec 5, 2016;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
15. Critical appraisal tool for artificial intelligence clinical studies. a scoping review. Open Science Framework. Apr 18, 2024. URL: https://osf.io/etyds/overview [Accessed 2026-09-07]
16. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
17. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. Jan 26, 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
18. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. Mar 1, 2023;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
19. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act. Dec 11, 2021;18(1):160. [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
20. Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. Sep 2016;77:24-37. [doi: 10.1016/j.jclinepi.2016.04.005] [Medline: 27140444]
21. Dwan K, Altman DG, Clarke M, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Med. Jun 2014;11(6):e1001666. [doi: 10.1371/journal.pmed.1001666] [Medline: 24959719]
22. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. Apr 26, 2010;8(1):24. [doi: 10.1186/1741-7015-8-24] [Medline: 20420659]
23. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med (Lond). Apr 11, 2024;4(1):71. [doi: 10.1038/s43856-024-00492-0] [Medline: 38605106]

Abbreviations

CHART: Chatbot Assessment Reporting Tool
PCC: population, concept, and context
PRISMA-AI: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Artificial Intelligence Extension
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies Using AI
RAG: retrieval-augmented generation
ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions

Edited by Amaryllis Mavragani; peer-reviewed by Andrea Conti, Seongsoon Kim; submitted 13.Oct.2025; final revised version received 27.Apr.2026; accepted 03.Jun.2026; published 28.Aug.2026

Please cite as:
Cabello López JB, Ruiz García V, Torralba M, Maldonado Fernandez M, Úbeda-Carrillo M, Ansuategi E, Ramos-Ruperto L, Emparanza JI, Urreta-Barallobre I, Iglesias Gaspar MAT, Pijoan Zubizarreta JI, Burls AJ
Data Availability for Critical Appraisal Tools for Evaluating Artificial Intelligence in Clinical Studies: Dataset for a Scoping Review
JMIR Data 2026;7:e85688
URL: https://data.jmir.org/2026/1/e85688
doi: 10.2196/85688