Clinical *Standard*

CLINICAL STANDARD · VOL. 1 NO. 1 · 2026
DOI 10.68913/CS.2026.0006 · CC BY 4.0

CLINICAL REVIEW · DIAGNOSTICS & IMAGING

How to read a study of a diagnostic artificial intelligence tool: a narrative review and checklist for clinicians

Abdoulrahman Elsayed Youssef Mahmoud Elnaemy[1*], Ahmed Eldesoki[2]

[1] Department of Orthopaedic Surgery, Faculty of Medicine, Suez Canal University Hospitals, Suez Canal University, Ismailia, Egypt
[2] Department of Trauma and Orthopaedics, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

[*] Correspondence: Abdoulrahman Elsayed Youssef Mahmoud Elnaemy, abdelrahman_elsayed@med.suez.edu.eg

ABSTRACT

Diagnostic tools based on artificial intelligence (AI) now read mammograms, colonoscopy video, chest radiographs, electrocardiograms and retinal photographs in routine care, yet most published evaluations are retrospective, single-centre comparisons on curated data; in one systematic review, only 2 of 83 published studies comparing deep learning with clinicians were randomised trials. This narrative review, based on a PubMed search to September 2026, gives clinicians a checklist for appraising such studies. Six questions expose most weaknesses: whether the patients resemble the intended population; whether the reference standard was sound and training and test data were kept apart; whether the tool was validated externally; whether discrimination is reported with calibration and subgroup performance; whether clinicians using the tool were compared with clinicians without it in real workflow; and whether a patient outcome was measured. Worked examples show the range of evidence. In the Mammography Screening with Artificial Intelligence (MASAI) trial of 105,915 women, AI-supported reading detected 29% more cancers, reduced screen-reading workload by 44% and gave a non-inferior interval cancer rate (proportion ratio 0.88, 95% confidence interval 0.65 to 1.18). By contrast, a widely deployed sepsis prediction model had an area under the receiver operating characteristic curve of 0.63 (95% confidence interval 0.62 to 0.64) on external validation. Computer-aided polyp detection increases adenoma detection in trials, but guideline panels disagree because no reduction in cancers has been shown. Clinicians should treat a diagnostic AI tool without external, preferably prospective, validation and evidence of patient benefit as experimental.

Keywords: Artificial Intelligence; Machine Learning; Diagnosis · Computer-Assisted; Sensitivity and Specificity · Evidence-Based Medicine · Randomized Controlled Trials as Topic.

Key points

- Ask who the patients were: accuracy measured on enriched or curated image sets overstates performance, so look for a consecutive series from the setting in which you would use the tool.
- Do not accept the area under the curve alone; look for calibration, the operating threshold, confidence intervals and results by sex, age, ethnicity and equipment.
- Look for external validation at more than one site, ideally prospective, and expect performance to fall when the population, equipment or workflow changes.
- Prefer studies that compare clinicians with and without the tool in real workflow and that measure patient outcomes, and watch for automation bias and deskilling.

- Treat a tool without external validation or prospective evaluation as experimental, and expect evidence of monitoring for drift once it is in use.

Introduction

Diagnostic tools built on artificial intelligence (AI), mostly machine-learning models, now read mammograms, colonoscopy video, chest radiographs, electrocardiograms and retinal photographs in routine care, and hundreds have regulatory authorisation. Clinicians are increasingly asked to use, buy or trust these tools, usually on the strength of a published study that they have little time to appraise.

The literature on diagnostic AI grew faster than its quality. A 2020 systematic review of studies comparing deep learning with clinicians found only two published randomised trials and 81 non-randomised studies, of which only nine were prospective, most were at high risk of bias, and most made claims in the abstract that the design could not support [1]. A meta-analysis published in 2019 found that few studies tested the algorithm and the clinicians on the same external data set [2], and a living review of 232 prediction models for COVID-19 rated almost all at high risk of bias [3]. Many of these are the old problems of diagnostic accuracy research, described by Lijmer and by Whiting two decades earlier, in which case–control sampling, differential verification and partial reference standards inflate accuracy [4,5]; others arise from the way models are trained.

Since then the field has matured. Reporting guidelines specific to AI have been published for diagnostic accuracy studies, prediction models, trial protocols and reports, and early clinical evaluation, and risk-of-bias tools have been extended [6–13]. Randomised trials with patient-relevant outcomes have reported in breast screening, colonoscopy and retinal screening. What remains unclear for most clinicians is how to apply this body of guidance quickly to a single paper.

This narrative review addresses one question: which features of a published study of a diagnostic AI tool should a non-specialist clinician check to decide whether the tool is ready for their patients? It distils the evidence into six questions and a ten-item checklist that can be applied in about ten minutes (Table 1).

Methods

This is a narrative review. PubMed (MEDLINE) was searched through the National Center for Biotechnology Information E-utilities interface from inception to September 2026. Searches combined terms for artificial intelligence (artificial intelligence, machine learning, deep learning) with terms for each topic: AI-specific reporting and risk-of-bias guidelines; systematic reviews of the design and quality of diagnostic AI studies; randomised and prospective trials of AI tools in screening mammography, colonoscopy, diabetic retinopathy screening, electrocardiography, chest radiography, stroke imaging and sepsis prediction; and studies of dataset shift, data leakage, calibration, fairness, automation bias and post-deployment monitoring. English-language articles were eligible. Sources were selected by the authors for relevance to clinical appraisal, with priority given to reporting guidelines, systematic reviews and randomised trials, and the reference lists of key articles were checked. Each cited reference was verified against its PubMed record and digital object identifier.

Selection was not systematic, and no formal risk-of-bias assessment or quantitative synthesis was performed. Estimates are reported as published in the source studies, with confidence intervals where these were reported. Because no new data were analysed, no statistical software was used and missing data did not arise. A self-assessment against the Scale for the Assessment of Narrative Review Articles (SANRA) [14] is provided in Supplementary Material S1. Ethical approval was not required.

Results

Question 1: who were the patients?

Ask how the images or records were selected. Public image sets assembled for algorithm development often have labels extracted

automatically from radiology reports, uncertain diagnoses and little clinical context [15], and enriched or case–control samples inflate sensitivity and specificity compared with a consecutive series of the patients who would actually be tested [4,16]. Models can also learn shortcuts: a deep learning model for detecting pneumonia on chest radiographs learned to recognise the hospital and the type of radiograph, such as portable films taken on the sickest wards, and performed worse on radiographs from other hospitals [17,18]. Spectrum, prevalence and case mix should be described as they would be for any diagnostic test [6].

Question 2: what was the reference standard, and how were the data split?

The label the model learns is only as good as the reference standard, and labels derived from radiology reports or discharge codes carry the errors of those documents. Then ask whether the test set was truly separate from the training data. Leakage, in which information from the test set influences training through duplicated patients, shared pre-processing or feature selection on the whole data set, is a widespread cause of over-optimistic results that cannot be reproduced [19,20]. Splits at the level of the patient, and temporal splits in which the model is tested on later patients, are more honest than random splits of images.

Question 3: was there external validation?

External validation remains uncommon. Of 516 studies of AI for diagnostic analysis of medical images published in 2018, only 6% validated the algorithm on external data [16]. In a systematic review of 86 deep learning algorithms for radiological diagnosis that had been tested externally, most performed worse on external data, about a quarter substantially so [21]. The fall reflects dataset shift: different scanners, protocols, populations and disease prevalence [22]. Of 130 AI devices cleared by the United States Food and Drug Administration (FDA) up to 2020, 126 had been evaluated only retrospectively and 93 did not report evaluation at more than one site [23], and other analyses of cleared devices found similar gaps, including little subgroup reporting [24,25]. External validation at more than one site, ideally prospective, is the minimum for a tool intended for wide use.

Question 4: is discrimination reported with calibration and subgroups?

The area under the receiver operating characteristic curve (AUC), the usual measure of discrimination, is not enough. Calibration, the agreement between predicted and observed risk, determines whether a threshold set in one place produces the same number of alerts in another, and it is often not reported [26]. Subgroup performance matters because models trained on unbalanced data underperform in the groups least represented: chest radiograph classifiers underdiagnosed disease more often in female, Black, Hispanic, younger and Medicaid-insured patients (Medicaid is a public insurance programme for people on low incomes in the United States) [27,28], and dermatology models performed worse on images of darker skin tones [29]. The Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence (STARD-AI) guideline and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis guideline updated for AI (TRIPOD+AI) require both to be reported [6,7].

Question 5: was the comparison realistic?

Many studies compare an algorithm with clinicians reading images without clinical information, outside their usual workflow and sometimes under time pressure that the algorithm does not share [1]. The relevant comparison is the clinician with the tool against the clinician without it, in the setting where it will be used. That comparison brings human factors into view. When radiologists were shown incorrect AI suggestions of the Breast Imaging Reporting and Data System (BI-RADS) category, their accuracy fell, and inexperienced readers were the most affected [30]. The benefit of AI assistance varies widely between radiologists and is not reliably predicted by their experience or unaided accuracy [31]. In a multicentre observational study, endoscopists who routinely used computer-aided detection had a lower

adenoma detection rate in colonoscopies performed without it (22.4% versus 28.4% before exposure), one of the first reports of deskilling [32]. Multi-reader multi-case studies, prospective reader studies reported according to the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) guideline, and randomised trials answer this question; retrospective bench comparisons do not [10].

Question 6: was the outcome a patient outcome?

Detecting more findings is not the same as benefit. In colonoscopy, computer-aided detection (CADe) increases the adenoma detection rate in randomised trials [33,34], but the additional lesions are mostly small, the effect in pragmatic community trials and non-randomised studies has been smaller or absent [35–38], and no trial has yet shown fewer interval cancers. Guideline panels reading the same evidence therefore disagree: the BMJ living guideline makes a weak recommendation against routine use of CADe, and the American Gastroenterological Association makes no recommendation for or against it [39,40].

In sepsis, a proprietary prediction model in wide use (Epic Sepsis Model; Epic Systems Corporation, Verona, WI, USA) had, on external validation, an AUC of 0.63 (95% confidence interval [CI] 0.62 to 0.64); it missed two thirds of patients with sepsis while generating alerts for 18% of all hospitalised patients [41], and a second independent validation also found modest performance [42]. An updated version has since been validated prospectively across several centres, with better but still modest performance [43], and alert fatigue is the measurable cost of such tools [44]. By contrast, in a prospective study at five hospitals of the Targeted Real-time Early Warning System (TREWS), a machine-learning sepsis alert, patients whose alert was confirmed by a clinician within 3 hours had lower in-hospital mortality than those whose alert was not, although this design shows association rather than causation [45]. An earlier small randomised trial in two intensive care units reported lower mortality and shorter hospital stay with a machine-learning sepsis alert [46], and a cluster-randomised trial of decision support for septic shock in paediatric emergency departments reported clinical outcomes rather than alert performance [47]. Reviews find that randomised trials of AI in practice remain few, are concentrated in gastroenterology and radiology, and often under-report how clinicians interacted with the tool [48,49].

Worked example: mammography screening

The Swedish Mammography Screening with Artificial Intelligence (MASAI) trial randomised 105,915 women to AI-supported reading or to standard double reading by two radiologists. In the intervention arm, the AI system (Transpara; ScreenPoint Medical, Nijmegen, the Netherlands) triaged examinations to single or double reading and highlighted suspicious findings. AI-supported reading reduced the screen-reading workload by 44% [50], detected 29% more cancers [51], and gave a non-inferior interval cancer rate (1.55 versus 1.76 per 1000 women; proportion ratio 0.88, 95% CI 0.65 to 1.18), with higher sensitivity (80.5% versus 73.8%) and identical specificity (98.5%) [52]. A prospective paired-reader study in Stockholm [53], an analysis of how radiologists acted on examinations flagged by AI in that study [54], and a nationwide real-world implementation in Germany including more than 460,000 women [55] support these findings. Together these studies answer questions 1 to 6: consecutive screening populations, histologically confirmed cancer as the reference standard, prospective evaluation in practice, radiologists using the tool as intended, and interval cancer as the outcome. An earlier retrospective evaluation of a different system, which reported performance exceeding that of radiologists on curated data sets, could not answer them [56].

Worked example: diabetic retinopathy

Two autonomous systems have been evaluated in pivotal prospective trials in primary care against a reading-centre reference standard. The first (IDx-DR, now LumineticsCore; Digital Diagnostics, Coralville, IA, USA) had a sensitivity of 87.2% (95% CI 81.8 to 91.2) and a specificity of 90.7% (95% CI 88.3 to 92.7) and in 2018 became the

first autonomous AI diagnostic system authorised by the FDA [57]; a second (EyeArt; Eyenuk, Woodland Hills, CA, USA) performed similarly in its own pivotal study [58]. A subsequent randomised trial in young people with diabetes showed that autonomous screening at the point of care increased the proportion who completed screening and follow-up [59], an outcome that matters more than the accuracy figure.

Worked example: electrocardiography

An AI algorithm that identifies low left ventricular ejection fraction from a standard 12-lead electrocardiogram increased the diagnosis of low ejection fraction in a pragmatic cluster-randomised trial in primary care in the United States [60], and increased the detection of cardiomyopathy in a pragmatic randomised trial in pregnant and postpartum women in Nigeria [61]. Both trials used the tool in clinical workflow and measured diagnoses made, not the AUC.

Worked example: chest radiography and stroke

In a randomised trial in a health screening population, AI assistance increased the detection of actionable lung nodules on chest radiographs [62], and a mixed-methods study is evaluating whether software from Qure.ai (Mumbai, India) that prioritises chest radiographs for reporting speeds the diagnosis of lung cancer [63]. Automated detection of large-vessel occlusion shortened transfer and treatment times in some stroke networks but not in others, because the gain depends on how the care team responds to the alert [64–66].

Appraisal instruments and regulation

Table 2 lists the instruments. For reporting, STARD-AI covers diagnostic accuracy studies [6], TRIPOD+AI covers prediction models [7], the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI) extension and the Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence (SPIRIT-AI) extension cover trial reports and protocols [8,9], DECIDE-AI covers early clinical evaluation and human factors [10], the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) covers imaging studies [13,67], the Minimum Information for Medical AI Reporting (MINIMAR) standard sets the minimum information for any medical AI report [68], and the Chatbot Assessment Reporting Tool (CHART) covers studies of chatbot health advice [69]. For risk of bias, PROBAST+AI extends the Prediction model Risk Of Bias ASsessment Tool (PROBAST) to machine-learning prediction models [11], and the Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence (QUADAS-AI) tool extends QUADAS-2 to AI diagnostic accuracy studies, with a revised version in development [12,70].

Regulators are moving in the same direction. The European Union Artificial Intelligence Act classifies AI systems that are medical devices requiring third-party conformity assessment as high risk, with obligations for post-market monitoring [71]; the European Society of Radiology has issued consensus recommendations for surveillance after deployment [72,73]; and methods for detecting harmful data shift after deployment are being tested [74]. A tool that was accurate at approval can drift, so evidence of monitoring belongs in the appraisal of any tool that has been in use for more than a year.

Discussion

Principal findings

Most of the weaknesses in published evaluations of diagnostic AI can be exposed with six questions about patients, reference standard and data separation, external validation, calibration and subgroups, the realism of the comparison, and the outcome measured. The tools with the strongest evidence, AI-supported mammography reading, autonomous diabetic retinopathy screening and AI-enabled electrocardiography, are those evaluated prospectively, in workflow, with outcomes beyond accuracy. Tools assessed mainly on retrospective data, such as the original sepsis model, have performed worse when tested elsewhere.

Comparison with other guidance

The checklist builds on the users' guide to reading articles that use machine learning [75] and on the

AI extensions of reporting and risk-of-bias instruments [6–13]. It differs in being written for the non-specialist clinician who must decide about a tool, and in placing more weight on the realism of the comparison, calibration in the local population and patient outcomes than on the technical details of model development.

Strengths and limitations

The checklist is grounded in published reporting and risk-of-bias guidelines and illustrated with randomised and prospective studies, and every reference was checked against its source record. The review has the limitations of its design. Only one database was searched, selection was not systematic and may favour well-known trials, and the field is changing quickly. Most examples come from imaging and from high-income countries, and the checklist has not itself been validated as a tool for appraisal. AI tools assisted the literature search and drafting, as declared below; the authors checked all content.

Implications for clinicians and policy

Clinicians should ask vendors and colleagues for external, ideally local, validation that reports calibration and subgroup results before relying on a tool, and should remain alert to automation bias. Health services should require prospective evaluation before purchase and continuous monitoring after deployment, and journals should require adherence to the AI reporting guidelines.

Unanswered questions

It is not yet known whether CADe reduces post-colonoscopy cancers, whether AI-supported mammography reduces breast cancer mortality, how large and how reversible deskilling is, how to monitor deployed tools for drift efficiently, and whether performance gaps between population groups can be closed. Tools based on large language models raise these questions again in a less mature evidence base [69].

Conclusion

The question for a clinician is not whether an algorithm can match an expert on a test set but whether patients seen with the tool do better than patients seen without it. Only a few tools have answered that question, most convincingly in breast screening; many in use have not. The six questions and the ten-item checklist let a reader tell the two apart from the paper alone, and the new reporting and appraisal instruments make it fair to expect authors to answer them.

DECLARATIONS

Author contributions

Abdoulrahman Elsayed Youssef Mahmoud Elnaemy: conceptualisation, methodology, investigation, writing – original draft, writing – review and editing, project administration. Ahmed Eldesoki: validation, supervision, writing – review and editing. Both authors approved the final version and are accountable for all aspects of the work.

Acknowledgements

None.

Funding

No specific funding.

Competing interests

None declared. Both authors have completed the International Committee of Medical Journal Editors (ICMJE) disclosure form and declare no financial or non-financial interests in the past 36 months.

Ethical approval

Not applicable. This is a review of published literature and involved no human participants, patient data or animals.

Data availability

No new data were generated or analysed. All sources are cited in the reference list.

Registration

Not applicable. Narrative reviews are not eligible for registration in PROSPERO. Use of AI Grammarly (Grammarly Inc., San Francisco, CA, USA) was used for language editing. No AI tool was used to create or alter data or images. The authors checked every reference against its source record, reviewed and edited all text, and take full responsibility for the content.

REFERENCES

1. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. doi:10.1136/bmj.m689
2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2
3. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi:10.1136/bmj.m1328
4. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061-6. doi:10.1001/jama.282.11.1061
5. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189-202. doi:10.7326/0003-4819-140-3-200402030-00010
6. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-9. doi:10.1038/s41591-025-03953-8
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi:10.1038/s41591-020-1034-x
9. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ. 2020;370:m3210. doi:10.1136/bmj.m3210
10. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
11. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505
12. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021;27(10):1663-5. doi:10.1038/s41591-021-01517-0
13. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. doi:10.1148/ryai.240300
14. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4:5. doi:10.1186/s41073-019-0064-8
15. Oakden-Rayner L. Exploring large-scale public medical image datasets. Acad Radiol. 2020;27(1):106-12. doi:10.1016/j.acra.2019.10.006
16. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20(3):405-10. doi:10.3348/kjr.2019.0025
17. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683
18. Banerjee I, Bhattacharjee K, Burns JL, Trivedi H, Purkayastha S, Seyyed-Kalantari L, et al. “Shortcuts” causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. J Am Coll Radiol. 2023;20(9):842-51. doi:10.1016/j.jacr.2023.06.025
19. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804

20. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: part II—generalization and overfitting. Acta Neurochir Suppl. 2022;134:15-21. doi:10.1007/978-3-030-85292-4_3
21. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. 2022;4(3):e210064. doi:10.1148/ryai.210064
22. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-6. doi:10.1056/NEJMc2104626
23. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582-4. doi:10.1038/s41591-021-01312-x
24. Ebrahimian S, Kalra MK, Agarwal S, Bizzo BC, Elkholy M, Wald C, et al. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. Acad Radiol. 2022;29(4):559-66. doi:10.1016/j.acra.2021.09.002
25. Potnis KC, Ross JS, Aneja S, Gross CP, Richman IB. Artificial intelligence in breast cancer screening: evaluation of FDA device regulation and future recommendations. JAMA Intern Med. 2022;182(12):1306-12. doi:10.1001/jamainternmed.2022.4969
26. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
27. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82. doi:10.1038/s41591-021-01595-0
28. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput. 2021;26:232-43.
29. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv. 2022;8(32):eabq6147. doi:10.1126/sciadv.abq6147
30. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4):e222176. doi:10.1148/radiol.222176
31. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med. 2024;30(3):837-49. doi:10.1038/s41591-024-02850-w
32. Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol. 2025;10(10):896-903. doi:10.1016/S2468-1253(25)00133-5
33. Hassan C, Spadaccini M, Mori Y, Foroutan F, Facciorusso A, Gkolfakis P, et al. Real-time computer-aided detection of colorectal neoplasia during colonoscopy: a systematic review and meta-analysis. Ann Intern Med. 2023;176(9):1209-20. doi:10.7326/M22-3678
34. Soleymanjahi S, Huebner J, Elmansy L, Rajashekar N, Lüdtke N, Paracha R, et al. Artificial intelligence-assisted colonoscopy for polyp detection: a systematic review and meta-analysis. Ann Intern Med. 2024;177(12):1652-63. doi:10.7326/ANNALS-24-00981
35. Thiruvengadam NR, Solaimani P, Shrestha M, Buller S, Carson R, Reyes-Garcia B, et al. The efficacy of real-time computer-aided detection of colonic neoplasia in community practice: a pragmatic randomized controlled trial. Clin Gastroenterol Hepatol. 2024;22(11):2221-30.e15. doi:10.1016/j.cgh.2024.02.021
36. Patel HK, Mori Y, Hassan C, Rizkala T, Radadiya DK, Nathani P, et al. Lack of effectiveness of computer aided detection for colorectal neoplasia: a systematic review and meta-analysis of nonrandomized studies. Clin Gastroenterol Hepatol. 2024;22(5):971-80.e15. doi:10.1016/j.cgh.2023.11.029
37. Spadaccini M, Hassan C, Mori Y, Massimi D, Correale L, Facciorusso A, et al. Variability in computer-aided detection effect on adenoma detection rate in randomized controlled trials: a meta-regression analysis. Dig Liver Dis. 2025;57(6):1141-8. doi:10.1016/j.dld.2025.01.192
38. Wei MT, Fay S, Yung D, Ladabaum U, Kopylov U. Artificial intelligence-assisted colonoscopy in real-world clinical practice: a systematic review and meta-analysis. Clin Transl Gastroenterol. 2024;15(3):e00671. doi:10.14309/ctg.0000000000000671
39. Foroutan F, Vandvik PO, Helsingen LM, Kalager M, Rutter M, Selby K, et al. Computer aided detection and diagnosis of polyps in adult patients undergoing colonoscopy: a living clinical practice guideline. BMJ. 2025;388:e082656. doi:10.1136/bmj-2024-082656
40. Sultan S, Shung DL, Kolb JM, Foroutan F, Hassan C, Kahi CJ, et al. AGA living clinical practice guideline on computer-aided detection-assisted colonoscopy. Gastroenterology. 2025;168(4):691-700. doi:10.1053/j.gastro.2025.01.002
41. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70. doi:10.1001/jamainternmed.2021.2626

42. Cull J, Brevetta R, Gerac J, Kothari S, Blackhurst D. Epic sepsis model inpatient predictive analytic tool: a validation study. Crit Care Explor. 2023;5(7):e0941. doi:10.1097/CCE.0000000000000941
43. Wong A, Currey D, Schwinne M, Park-Egan B, Meyer S, Gutting A, et al. Multicenter prospective validation of an updated proprietary sepsis prediction model. JAMA Netw Open. 2026;9(2):e260181. doi:10.1001/jamanetworkopen.2026.0181
44. Ray CE, Wilson GM, Hughes AM, Cunningham Goedken C, Liu EP, Fitzpatrick MA, et al. Alert fatigue measurement in clinical decision support: a systematic review. J Am Med Inform Assoc. 2026;33(8):1523-31. doi:10.1093/jamia/ocag064
45. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med. 2022;28(7):1455-60. doi:10.1038/s41591-022-01894-0
46. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res. 2017;4(1):e000234. doi:10.1136/bmjresp-2017-000234
47. Scott HF, Sevick CJ, Colborn KL, Deakyne Davies SJ, Greer CH, Dafoe A, et al. Clinical decision support for septic shock in the emergency department: a cluster randomized trial. Pediatrics. 2025;156(1):e2024069478. doi:10.1542/peds.2024-069478
48. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health. 2024;6(5):e367-73. doi:10.1016/S2589-7500(24)00047-5
49. Ong AQC, Ang CS, Bojic I, Johnson CL, Aggour H, Ng FS, et al. Artificial intelligence in cardiovascular care: a systematic review and meta-analysis of randomised controlled trials. EClinicalMedicine. 2026;98:104123. doi:10.1016/j.eclinm.2026.104123
50. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936-44. doi:10.1016/S1470-2045(23)00298-X
51. Hernström V, Josefsson V, Sartor H, Schmidt D, Larsson AM, Hofvind S, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. Lancet Digit Health. 2025;7(3):e175-83. doi:10.1016/S2589-7500(24)00267-X
52. Gommers J, Hernström V, Josefsson V, Sartor H, Schmidt D, Hjelmgren A, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet. 2026;407(10527):505-14. doi:10.1016/S0140-6736(25)02464-X
53. Dembrower K, Crippa A, Colón E, Eklund M, Strand F; ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health. 2023;5(10):e703-11. doi:10.1016/S2589-7500(23)00153-X
54. Dembrower KE, Crippa A, Eklund M, Strand F. Human-AI interaction in the ScreenTrustCAD trial: recall proportion and positive predictive value related to screening mammograms flagged by AI CAD versus a human reader. Radiology. 2025;314(3):e242566. doi:10.1148/radiol.242566
55. Eisemann N, Bunk S, Mukama T, Baltus H, Elsner SA, Gomille T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;31(3):917-24. doi:10.1038/s41591-024-03408-6
56. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
57. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. doi:10.1038/s41746-018-0040-6
58. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open. 2021;4(11):e2134254. doi:10.1001/jamanetworkopen.2021.34254
59. Wolf RM, Channa R, Liu TYA, Zehra A, Bromberger L, Patel D, et al. Autonomous artificial intelligence increases screening and follow-up for diabetic retinopathy in youth: the ACCESS randomized control trial. Nat Commun. 2024;15(1):421. doi:10.1038/s41467-023-44676-z
60. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med. 2021;27(5):815-9. doi:10.1038/s41591-021-01335-4

61. Adedinsewo DA, Morales-Lara AC, Afolabi BB, Kushimo OA, Mbakwem AC, Ibiyemi KF, et al. Artificial intelligence guided screening for cardiomyopathies in an obstetric population: a pragmatic randomized clinical trial. Nat Med. 2024;30(10):2897-906. doi:10.1038/s41591-024-03243-9
62. Nam JG, Hwang EJ, Kim J, Park N, Lee EH, Kim HJ, et al. AI improves nodule detection on chest radiographs in a health screening population: a randomized controlled trial. Radiology. 2023;307(2):e221894. doi:10.1148/radiol.221894
63. Duncan SF, McConnachie A, Blackwood J, Stobo DB, Maclay JD, Wu O, et al. Radiograph accelerated detection and identification of cancer in the lung (RADICAL): a mixed methods study to assess the clinical effectiveness and acceptability of Qure.ai artificial intelligence software to prioritise chest X-ray (CXR) interpretation. BMJ Open. 2024;14(9):e081062. doi:10.1136/bmjopen-2023-081062
64. Le NM, Iyyangar AS, Kim Y, Chaudhry MR, Salazar-Marioni S, Abdelkhaleq R, et al. Machine learning-enabled automated large vessel occlusion detection improves transfer times at primary stroke centers. Stroke Vasc Interv Neurol. 2024;4(3):e001119. doi:10.1161/SVIN.123.001119
65. Hassan AE, Ringheanu VM, Preston L, Tekle WG. Artificial intelligence-parallel stroke workflow tool improves reperfusion rates and door-in to puncture interval. Stroke Vasc Interv Neurol. 2022;2(5):e000224. doi:10.1161/SVIN.121.000224
66. Ebirim EC, Le NM, Samaha JN, Azeem H, Iyyangar A, Ballekere AN, et al. Workflow improvements from automated large vessel occlusion detection algorithms are dependent on care team engagement. J Neurointerv Surg. 2026;18(2):385-9. doi:10.1136/jnis-2024-022896
67. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029. doi:10.1148/ryai.2020200029
68. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011-5. doi:10.1093/jamia/ocaa088
69. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med. 2025;4(1):e001632. doi:10.1136/bmjmed-2025-001632
70. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. 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. doi:10.2196/58202
71. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): implications for healthcare. Health Policy. 2024;149:105152. doi:10.1016/j.healthpol.2024.105152
72. Cuocolo R, Bernardini D, Pinto Dos Santos D, Klontzas ME, Akinci D'Antonoli T, Semedo LC, et al. AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. Insights Imaging. 2025;16(1):275. doi:10.1186/s13244-025-02146-8
73. Lampignano JAP, Kale A, Kumar S, Shah D, Reddy B, Lowe DJ. Beyond validation: operationalising post-deployment surveillance of AI medical devices in clinical practice. J Med Syst. 2026;50(1). doi:10.1007/s10916-026-02426-w
74. Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. JAMA Netw Open. 2025;8(6):e2513685. doi:10.1001/jamanetworkopen.2025.13685
75. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA. 2019;322(18):1806-16. doi:10.1001/jama.2019.16489
76. Ten-item checklist for appraising a study of a diagnostic artificial intelligence tool
77. AUC, area under the receiver operating characteristic curve; CONSORT-AI, Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI, Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence; STARD-AI, Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence.
78. Instruments for reporting and appraising studies of diagnostic and predictive artificial intelligence
79. AI, artificial intelligence; CHART, Chatbot Assessment Reporting Tool; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CONSORT-AI, Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI, Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence; MINIMAR, Minimum Information for Medical AI Reporting; PROBAST+AI, Prediction model Risk Of Bias ASsessment Tool updated for artificial intelligence; QUADAS-AI, Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; STARD-AI, Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence.

Received 02 October 2026 **Accepted** 07 October 2026 **Published** 07 October 2026
Citation Abdoulrahman Elsayed Youssef Mahmoud Elnaemy, Ahmed Eldesoki. How to read a study of a diagnostic artificial intelligence tool: a narrative review and checklist for clinicians. Clinical Standard. 2026;1(1). doi:10.68913/cs.2026.0006
Licence CC BY 4.0. This is an open-access article distributed under the terms of the Creative Commons Attribution licence.