nature medicine

Consensus Statement

https://doi.org/10.1038/s41591-025-03953-8

The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence

Received: 21 February 2025

Accepted: 11 August 2025

Published online: 15 September 2025

Check for updates

Viknesh Sounderajah[1,2,41], Ahmad Guni[1,2,41], Xiaoxuan Liu[3,4], Gary S. Collins[5], Alan Karthikesalingam[6], Sheraz R. Markar[2,7], Robert M. Golub[8], Alastair K. Denniston[3,4], Shravya Shetty[9], David Moher[10], Patrick M. Bossuyt[11], Ara Darzi[1,2,42], Hutan Ashrafian[1,2,42] ✉ & STARD-AI Steering Committee*

The Standards for Reporting Diagnostic Accuracy (STARD) 2015 statement facilitates transparent and complete reporting of diagnostic test accuracy studies. However, there are unique considerations associated with artificial intelligence (AI)-centered diagnostic test studies. The STARD-AI statement, which was developed through a multistage, multistakeholder process, provides a minimum set of criteria that allows for comprehensive reporting of AI-centered diagnostic test accuracy studies. The process involved a literature review, a scoping survey of international experts, and a patient and public involvement and engagement initiative, culminating in a modified Delphi consensus process involving over 240 international stakeholders and a consensus meeting. The checklist was subsequently finalized by the Steering Committee and includes 18 new or modified items in addition to the STARD 2015 checklist items. Authors are encouraged to provide descriptions of dataset practices, the AI index test and how it was evaluated, as well as considerations of algorithmic bias and fairness. The STARD-AI statement supports comprehensive and transparent reporting in all AI-centered diagnostic accuracy studies, and it can help key stakeholders to evaluate the biases, applicability and generalizability of study findings.

Diagnosis is fundamental to delivering effective healthcare. Clinical information within electronic health records (EHRs), imaging, laboratory tests and pathology can facilitate the timely and accurate detection of diseases[1–3]. For patients, this can provide an explanation for their health condition and guide clinicians to choose appropriate treatments, potentially improving patient outcomes[4,5]. Public and global health measures are also principally guided by effective diagnostic workflows[6].

Diagnostic research is often at risk of producing biased results, due to flaws in methodological design and lack of transparency[7]. It has also long been a concern that reporting of diagnostic test research is inadequate and inconsistent, leading to substantial research misrepresentation and waste[8–10]. Furthermore, it is often incorrectly assumed that the diagnostic accuracy of a test is a fixed characteristic; it is now well understood that common diagnostic accuracy measures (for example, sensitivity and specificity) can vary across clinical contexts, target populations, disease severity and different definitions of a reference standard[11,12]. Key information about the study design, setting, participants, index tests, reference standards, analysis and outcomes should be reported in all diagnostic test accuracy studies. Missing or unclear information hampers safe translation into clinical practice as key stakeholders, such as healthcare professionals, regulators and policymakers, are unable to evaluate the evidence base of a diagnostic test.

A full list of affiliations appears at the end of the paper. *Lists of authors and their affiliations appear at the end of the paper.
✉e-mail: h.ashrafian@imperial.ac.uk

In response to this, the STARD statement was developed in 2003, and was subsequently updated in 2015 (STARD 2015), to standardize the reporting of diagnostic accuracy research[13,14]. By outlining a list of 30 minimum essential items that should be reported for every diagnostic test accuracy study, STARD can improve the quality of study reporting, help stakeholders judge the risk of bias and applicability of the findings and enhance research reproducibility. The accompanying explanation and elaboration document provides the rationale for each item with examples of good reporting[15]. STARD has since been extended to provide guidance for reporting studies in conference abstracts (STARD for Abstracts)[16]. Evidence suggests that adherence to STARD improves the reporting of key information in diagnostic test accuracy studies[17,18].

The landscape of clinical diagnostics has shifted considerably since the release of STARD 2015. Advances in understanding diseases at both population and molecular levels[19–22], as well as technological breakthroughs such as AI[23,24], could enhance diagnostic capacity and efficacy. As a technology, AI may have the unique potential to both improve the performance of diagnostic systems and streamline workflows to alleviate healthcare resources[25]. Moreover, diagnostics constitutes a substantial proportion of clinical AI focus, with most AI devices achieving regulatory approval thus far belonging to the diagnostic field[26]. However, research in this field has thus far been conducted without a suitable reporting guideline that accounts for the unique properties of AI-driven diagnostic systems and the associated challenges.

For the purposes of this guideline, AI refers to computer systems that can perform tasks that typically require human intelligence, such as classification, prediction or pattern recognition. This includes, but is not limited to, machine learning and deep learning models, natural language processing tools or foundation models that generate or support diagnostic outputs. Systems that include static or manually programmed rules without adaptive learning, such as simple decision trees, were not included in the scope. AI introduces several additional potential sources of bias that are currently not always reported by study authors or accounted for by existing guidelines[27]. These may be related to study design, patient selection, dataset handling, ethical considerations, index test and reference standard conduct, statistical methods, reporting of results and discussion and interpretation of findings. Therefore, an accurate evaluation of the clinical applicability of AI-centered diagnostic systems is not always possible.

To strengthen the reporting of AI-centered diagnostic accuracy studies, the STARD-AI statement was developed. STARD-AI provides a checklist of minimum criteria that should be reported in every diagnostic test accuracy study evaluating an AI system. It joins several complementary EQUATOR Network initiatives that outline reporting guidelines for clinical AI studies, including CONSORT-AI for clinical trials of AI interventions[28], SPIRIT-AI for trial protocols[29], TRIPOD+AI for prediction and prognostic models[30] and CLAIM for medical imaging studies[31]. Relevant reporting guidelines and their scopes can be viewed in Table 1. The aim of STARD-AI is to improve completeness and transparency in study reporting, supporting stakeholders to evaluate the robustness of study methodology, assess the risk of bias and inform applicability and generalizability of study findings. This article outlines STARD-AI and describes the process of its development.

Table 1 | Reporting guidelines for AI-based medical devices and their scope

Reporting guideline	Scope and intended use
STARD-AI	Studies evaluating the diagnostic accuracy of an AI-based test
TRIPOD+AI	Studies developing and validating a prediction model using regression or AI-based methods
TRIPOD-LLM	Studies developing and validating a prediction model using LLMs
CLAIM	Studies developing and validating a medical imaging AI model
DECIDE-AI	Studies reporting the early-stage clinical evaluation of an AI-based decision support system
SPIRIT-AI	Clinical trial protocols for AI-based interventions
CONSORT-AI	Clinical trials evaluating AI-based interventions
CHEERS-AI	Studies evaluating the health economics of AI-based interventions

The STARD-AI statement

The final STARD-AI statement consists of 40 items that are considered essential in reporting of AI-centered diagnostic accuracy studies (Table 2). The development process can be visualized in Fig. 1. A downloadable, user-friendly version of the checklist can be found in Supplementary Table 2. Four items were modified from the STARD 2015 statement (items 1, 3, 7 and 25), and 14 new items were introduced to account for AI-specific considerations (items 6, 11, 12, 13, 14, 15b, 15d, 23, 28, 29, 35, 39, 40a and 40b). In a structure similar to STARD 2015, the checklist contains items relating to the title or abstract (item 1), abstract (item 2), introduction (items 3 and 4), methods (items 5–23), results (items 24–32), discussion (items 33–35) and other important information (items 36–40). Subsections are included within methods and results to make the checklist clearer to follow and interpret. The methods section is subdivided into study design, ethics, participants, dataset, test methods and analysis subsections, and the results section contains subitems relating to the participants, dataset and test results. In line with STARD 2015, a diagram illustrating the flow of participants is expected in reports (item 24); a template diagram is available in the STARD 2015 publication[14]. The rationale for new or modified items is outlined in Supplementary Table 3. For convenience, the STARD for Abstracts checklist is reproduced in Table 3 (ref. 16).

AI poses several considerations in various domains that are often not encountered in traditional diagnostic test accuracy studies. In particular, STARD-AI introduces several items that focus on data handling practices. These include detailing the eligibility criteria at both a dataset level and a participant level (item 7); source of the data and how they have been collected (item 11); dataset annotation (item 12); data capture devices and software versions (item 13); data acquisition protocols and preprocessing (item 14); partitioning of datasets into training, validation and test set purposes (item 15b); characteristics of the test set (item 25); and whether the test set represents the target condition (item 28). These items can substantially affect the diagnostic accuracy outcomes of a study and influence the risk of bias and applicability. As well as aiding evaluation of study findings, sufficient reporting of these items, in addition to clear explanations of the index test and reference standard, may facilitate reproducibility and aid in replicating studies. In line with collaborative open science practices, STARD-AI encourages disclosure of commercial interests (item 39), public availability of datasets and code (item 40a) and the external audit or evaluation of outputs (item 40b).

Use of STARD-AI can aid the comprehensive reporting of research that assesses AI diagnostic accuracy using either single or combined test data and can be applied across a broad range of diagnostic modalities. Examples include imaging, such as X-rays or computed tomography scans[32]; pathology through digital whole-slide images[33]; and clinical information in the form of EHRs[34]. In addition, studies may use other ways besides test accuracy to express diagnostic performance, including incremental accuracy gains within diagnostic pathways or clinical utility measures[35,36]. STARD-AI also supports the evaluation of multimodal diagnostic tools and can be used in studies that assess the diagnostic accuracy of large language models (LLMs), where the output consists of a diagnostic classification of differential diagnosis. By contrast, if the study focuses on the development or evaluation of a multivariable prediction model using regression, machine learning or LLM-based approaches to predict diagnostic or prognostic outcomes,

https://doi.org/10.1038/s41591-025-03953-8

Table 2 | The STARD-AI checklist

Section and topic	No.	STARD-AI item
Title or abstract		
	1†	Identification as a study reporting AI-centered diagnostic accuracy and reporting at least one measure of accuracy within title or abstract
Abstract		
	2	Structured summary of study design, methods, results and conclusions (for specific guidance, see STARD for Abstracts)
Introduction		
	3†	Scientific and clinical background, including the intended use of the index test, whether it is novel or an established index test and its integration into an existing or new workflow, if applicable
	4	Study objectives and hypotheses
Methods		
Study design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)
Ethics	6*	Formal approval from an ethics committee. If not required, justify why.
Participants	7†	Eligibility criteria: listing separate inclusion and exclusion criteria in the order that they are applied at both participant level and data level
	8	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests and inclusion in registry)
	9	Where and when potentially eligible participants were identified (setting, location and dates)
	10	Whether participants formed a consecutive, random or convenience series
Dataset	11*	Source of the data and whether they have been routinely collected, specifically collected for the purpose of the study or acquired from an open-source repository
	12*	Who undertook the annotations for the dataset (including experience levels and background) and how (within the same clinical context or in a post hoc fashion), if applicable
	13*	Devices (manufacturer and model) that were used to capture data; software (with version number) used to engineer the index test, highlighting the intended use
	14*	Data acquisition protocols (for example, contrast protocol or reconstruction method for medical images) and details of data preprocessing, in sufficient detail to allow replication
Test methods	15a	Index test, in sufficient detail to allow replication
	15b*	How the index test was developed, including any training, validation, testing and external evaluation, detailing sample sizes, when applicable
	15c	Definition of and rationale for test positivity cutoffs or result categories of the index test, distinguishing prespecified from exploratory
	15d*	The specified end-user of the index test and the level of expertise required of users
	16a	Reference standard, in sufficient detail to allow replication
	16b	Rationale for choosing the reference standard (if alternatives exist)
	16c	Definition of and rationale for test positivity cutoffs or result categories of the reference standard, distinguishing prespecified from exploratory
	17a	Whether clinical information and reference standard results were available to the performers or readers of the index test
	17b	Whether clinical information and index test results were available to the assessors of the reference standard
Analysis	18	Methods for estimating or comparing measures of diagnostic accuracy
	19	How indeterminate index test or reference standard results were handled
	20	How missing data on the index test and reference standard were handled
	21	Any analyses of variability in diagnostic accuracy, distinguishing prespecified from exploratory
	22	Intended sample size and how it was determined
	23*	Details of any performance error analysis and algorithmic bias and fairness assessments, if undertaken
Results		
Participants and dataset	24	Flow of participants, using a diagram
	25†	Baseline demographic, clinical and technical characteristics of training, validation and test sets, if applicable
	26a	Distribution of severity of disease in those with the target condition
	26b	Distribution of alternative diagnoses in those without the target condition
	27	Time interval and any clinical interventions between index test and reference standard
	28*	Whether the datasets represent the distribution of the target condition that one would expect from the intended use population
	29*	For external evaluation on an independent dataset, an assessment of how this differs from the training, validation and test sets

https://doi.org/10.1038/s41591-025-03953-8

Table 2 (continued) | The STARD-AI checklist

Section and topic	No.	STARD-AI item
Test results	30	Cross-tabulation of the index test results (or their distribution) by the results of the reference standard
	31	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	32	Any adverse events from performing the index test or the reference standard
Discussion		
	33	Study limitations, including sources of potential bias, statistical uncertainty and generalizability
	34	Implications for practice, including the intended use and clinical role of the index test
	35*	Ethical considerations and adherence to ethical standards associated with the use of the index test and issues of fairness
Other information		
	36	Registration number and name of registry
	37	Where the full study protocol can be accessed
	38	Sources of funding and other support; role of funders
	39*	Commercial interests, if applicable
	40a*	Availability of datasets and code, detailing any restrictions on their reuse and repurposing
	40b*	Whether outputs are stored, auditable and available for evaluation, if necessary

* New items † Modified items

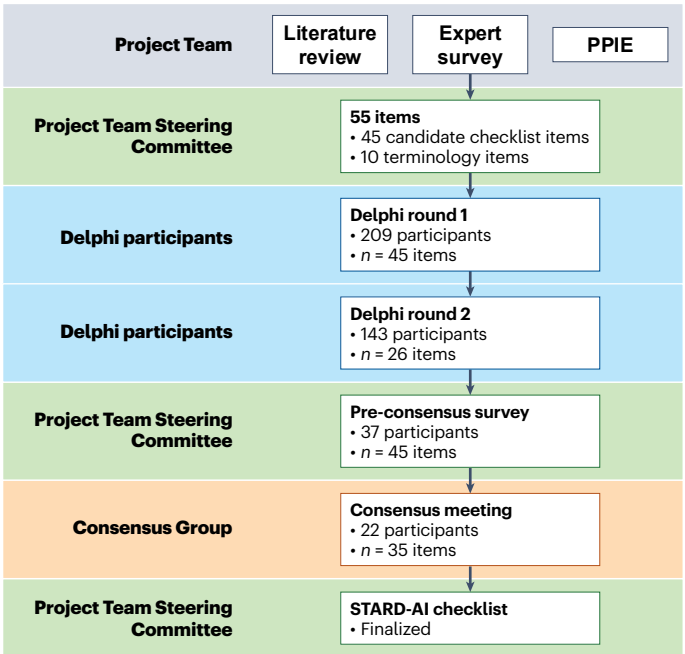

Fig. 1 | STARD-AI checklist development process. The checklist was developed through a multistage process, including a literature review, expert and public input (PPIE), Delphi surveys and a final consensus meeting. The number of participants and items assessed at each stage are shown.

use of TRIPOD+AI or TRIPOD-LLM is more appropriate[30,37]. CLAIM may be considered for the development or validation of a medical imaging AI model[31], whereas STARD is more applicable where diagnostic accuracy of a model is the primary focus. Where relevant, authors can consider referring to multiple checklists but may select the guideline most aligned with the study's primary aim and evaluation framework for pragmatic reasons.

Discussion

STARD-AI is a new reporting guideline that can support the reporting of AI-centered diagnostic test accuracy studies. It was developed through a multistage process consisting of a comprehensive item generation phase followed by an international multistakeholder consensus. STARD-AI addresses considerations unique to AI technology, predominantly related to algorithmic and data practices, that are not accounted by its predecessor, STARD 2015. Although it proposes a set of items that should be reported in every study, many studies may benefit from reporting additional information related to individual study methodology and outcomes. STARD-AI should, therefore, be seen as a minimum set of essential items and not as an exhaustive list.

Research into clinical diagnostics using AI tools has thus far mostly focused on establishing the diagnostic accuracy of models. However, there are many challenges to successfully translating AI models to a clinical setting, including the limited number of well-conducted external evaluation studies to date; the lack of comparative and prospective trials; the use of study metrics that may not reflect clinical efficacy; and difficulties in achieving generalizability to new populations[38]. The deployment of these models into clinical scenarios outside research settings has raised concerns that intrinsic biases could propagate or entrench population health inequalities or even cause patient harm[39]. Therefore, it is crucial for potential users of diagnostic AI tools to focus not only on model performance but also on the robustness of the underlying evidence base, primarily through identifying flaws in study design or conduct that could lead to biases and poor applicability. STARD-AI can help on this front by guiding authors to include the important information needed for readers to evaluate a study.

Specific AI diagnostic elements to consider include transparency in AI models, bias, generalizability, algorithm explainability, clinical pathway integration, data provenance and quality, validation and robustness and ethical and regulatory considerations. As diagnostic tools currently dominate the landscape of regulatory-approved AI devices[26], guidelines such as STARD-AI may help to enhance the quality and transparency of studies reported for these devices. Ultimately, this may aid the development and deployment of AI models that leads to healthcare outcomes that are fair, appropriate, valid, effective and safe[40]. It may also support the deployment of AI models that align with Coalition for Health AI principles for trustworthy AI, namely algorithms that are reliable, testable, usable and beneficial[41,42].

STARD-AI provides many new criteria that outline appropriate dataset and algorithmic practices, stresses the need to identify and mitigate algorithmic biases and requires authors to consider fairness in both the methods (item 23) and the discussion (item 35) sections. In this context, fairness refers to the equitable treatment of individuals or groups across key attributes, including demographic

https://doi.org/10.1038/s41591-025-03953-8

Table 3 | STARD for Abstracts[16]

Section and topic	No.	Item
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values or area under the curve)
Background and objectives	2	Study objectives
Methods	3	Data collection: whether this was a prospective or retrospective study
	4	Eligibility criteria for participants and settings where the data were collected
	5	Whether participants formed a consecutive, random or convenience series
	6	Description of the index test and reference standard
Results	7	Number of participants with and without the target condition included in the analysis
	8	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
Discussion	9	General interpretation of the results
	10	Implications for practice, including the intended use of the index test
Registration	11	Registration number and name of registry

factors or socioeconomic status. This includes the expectation that an AI-based system should not systematically underperform or misclassify subgroups of patients in a manner that may reinforce existing health disparities. Ensuring model fairness is especially imperative in the context of diagnostic AI technology as these may eventually be deployed to assist clinical decision-making in population-wide diagnostic or screening strategies. If fairness is not considered sufficiently, equitable healthcare delivery may be hampered on a population level, and disparities between demographic groups may be exacerbated[43]. Datasets used to train, validate and test should ideally be diverse and represent the intended target population of the index test evaluated. Additional algorithmic practices can further reduce fairness gaps while maintaining performance[39].

The addition of 10 main items, and 14 subitems in total, increases the length of the checklist compared to STARD 2015. Although this may be seen as a barrier to implementation, it was deemed necessary to address AI-specific considerations that may substantially impact the quality of study reporting. Notably, other checklists, such as TRIPOD+AI and CLAIM, contain a similar number of total items and subitems[30,31]. We intend to release an explanation and elaboration document to provide examples and rationale for each new or modified item in STARD-AI, which we briefly outline in Supplementary Table 3. However, many of the items remain unchanged from STARD 2015, reflecting that the general principles of reporting diagnostic accuracy studies are still essential for AI tools. In the meantime, the STARD 2015 explanation and elaboration document provides rationale and examples of appropriate reporting for the unchanged items[15].

STARD-AI is designed to support the reporting of studies that evaluate the diagnostic accuracy of an AI tool. However, the increasing integration of AI system into clinical workflows highlights the growing importance of AI–human collaboration. In many real-world scenarios, AI tools are intended not to replace clinical decision-making but, rather, to inform or enhance it. Therefore, future studies should also assess the impact of AI assistance on end-user performance, in addition to reporting the standalone accuracy of the AI system. This should ideally include a comparison to a baseline in which clinical decisions are made without AI, which will aid in evaluating the clinical utility of AI on decision-making and workflows[44]. The experience and expertise of end users will also be important in determining performance outcomes. Addressing these elements moving forward may require the development of a separate consensus.

Although STARD-AI was developed prior to the wider introduction of generative AI and LLMs, many of these items nevertheless remain applicable to generative AI models that report diagnostic accuracy. Unlike classical AI models, which are typically trained on labeled datasets for specific tasks, LLMs and transformer-based architectures are generally pretrained on large-scale, unstructured datasets and can be subsequently finetuned for specific diagnostic tasks. Although STARD-AI can be applied to studies that investigate generative AI and future advances in AI platforms, it is likely that STARD-AI and other complementary guidelines will need to be regularly updated in response to the rapidly shifting nature of this field. Next-generation generative AI technology may consist of multimodal and generalist models that input medical and biomedical data to improve predictions[45–47]. Further advances in fields such as reinforcement learning[48,49], graph neural networks[50,51] and explainable AI (XAI) solutions[52] may also substantially change the landscape of health AI and require new considerations in the next iteration of reporting guidelines.

The rapid pace of technological advancement may also present inherent limitations to reporting guidelines. Although many of the STARD-AI items remain applicable to newer forms of AI, including foundation models and multiagent systems, the increasing complexity and versatility of these tools may challenge traditional concepts of diagnostic evaluation. Emerging systems may provide differential diagnoses ranked on probability or even interact dynamically with users via natural language and adapt outputs based on population characteristics or user expertise. These capabilities extend beyond conventional frameworks and may not be fully captured by traditional diagnostic accuracy metrics alone. Although STARD-AI offers a strong foundation for transparent reporting, complementary frameworks such as CRAFT-MD may be better suited for evaluating different forms of AI-driven clinical support[53].

We are confident that STARD-AI will prove useful to many stakeholders. STARD-AI provides study authors with a set of minimum criteria to improve the quality of reporting, although it does not aim to provide prescriptive step-by-step instructions to authors. If adopted as a reporting standard before or during manuscript submission to journals, editors and reviewers may be able to more effectively appraise submissions; its use by journals may also help to ensure that all information essential for readers is included in the published article. In the future, it may be possible that AI-based tools, such as LLMs, may assist in prescreening manuscripts for STARD-AI adherence, offering a scalable means to support checklist compliance during peer review and editorial assessment. Beyond the academic field, policymakers, regulators and industry partners are recommended to incorporate STARD-AI where the requirement for transparency of evidence is universally recognized, as well as complementary reporting guidelines within the EQUATOR Network[54], in clinical AI product and policy assessments to better guide downstream decisions and recommendations. End users such as clinicians may be able to more effectively evaluate the clinical utility of AI systems to their patient populations prior to use, and patients may benefit from the eventual outcome of higher-quality research.

Conclusion

Diagnostic pathways stand to benefit substantially from the use of AI. For this to happen, researchers should report their findings in sufficient detail to facilitate transparency and reproducibility. Similarly, readers and other decisionmakers should have the necessary information to judge the risk of bias, diagnostic accuracy test determinants, clinical context and applicability of study findings. STARD-AI is a consensus-based reporting guideline that clarifies these requirements.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-03953-8.

References

1. Reichlin, T. et al. Early diagnosis of myocardial infarction with sensitive cardiac troponin assays. *N. Engl. J. Med.* **361**, 858–867 (2009).
2. Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *BMJ* **364**, l408 (2019).
3. Neal, R. D. et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. *Br. J. Cancer* **112**, S92–S107 (2015).
4. Leifer, B. P. Early diagnosis of Alzheimer's disease: clinical and economic benefits. *J. Am. Geriatr. Soc.* **51**, S281–S288 (2003).
5. Crosby, D. et al. Early detection of cancer. *Science* **375**, eaay9040 (2022).
6. Fleming, K. A. et al. The *Lancet* Commission on diagnostics: transforming access to diagnostics. *Lancet* **398**, 1997–2050 (2021).
7. Whiting, P. F., Rutjes, A. W., Westwood, M. E. & Mallett, S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J. Clin. Epidemiol.* **66**, 1093–1104 (2013).
8. Glasziou, P. et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **383**, 267–276 (2014).
9. Ioannidis, J. P. et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166–175 (2014).
10. Lijmer, J. G. et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* **282**, 1061–1066 (1999).
11. Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C. & Lijmer, J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* **324**, 669–671 (2002).
12. Moons, K. G., van Es, G. A., Deckers, J. W., Habbema, J. D. & Grobbee, D. E. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* **8**, 12–17 (1997).
13. Bossuyt, P. M. et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern. Med.* **138**, W1–W12 (2003).
14. Bossuyt, P. M. et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* **351**, h5527 (2015).
15. Cohen, J. F. et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
16. Cohen, J. F. et al. STARD for Abstracts: essential items for reporting diagnostic accuracy studies in journal or conference abstracts. *BMJ* **358**, j3751 (2017).
17. Korevaar, D. A. et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology* **274**, 781–789 (2015).
18. Korevaar, D. A., van Enst, W. A., Spijker, R., Bossuyt, P. M. & Hooft, L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid. Based Med.* **19**, 47–54 (2014).
19. Miao, Z., Humphreys, B. D., McMahon, A. P. & Kim, J. Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).
20. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. Williamson, E. J. et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
22. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
23. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
24. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
25. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
26. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit. Med.* **3**, 118 (2020).
27. Liu, X. 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* **1**, e271–e297 (2019).
28. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
29. Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K. & Calvert, M. J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* **370**, m3210 (2020).
30. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
31. Tejani, A. S. et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol. Artif. Intell.* **6**, e240300 (2024).
32. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit.Med.* **4**, 65 (2021).
33. McGenity, C. et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *NPJ Digit. Med.* **7**, 114 (2024).
34. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
35. Moons, K. G. M., de Groot, J. A. H., Linnet, K., Reitsma, J. B. & Bossuyt, P. M. M. Quantifying the added value of a diagnostic test or marker. *Clin. Chem.* **58**, 1408–1417 (2012).
36. Bossuyt, P. M. M., Reitsma, J. B., Linnet, K. & Moons, K. G. M. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clin. Chem.* **58**, 1636–1643 (2012).
37. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
38. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
39. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
40. The White House. Delivering on the Promise of AI to Improve Health Outcomes. https://bidenwhitehouse.archives.gov/briefing-room/blog/2023/12/14/delivering-on-the-promise-of-ai-to-improve-health-outcomes/ (2023).
41. Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. https://www.chai.org/workgroup/responsible-ai/blueprint-for-trustworthy-ai (2023).
42. Guni, A., Varma, P., Zhang, J., Fehervari, M. & Ashrafian, H. Artificial intelligence in surgery: the future is now. *Eur. Surg. Res.* https://doi.org/10.1159/000536393 (2024).

43. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
44. Krakowski, I. et al. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. *NPJ Digit. Med.* **7**, 78 (2024).
45. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
46. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* **1**, AIoa2300138 (2024).
47. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
48. Barata, C. et al. A reinforcement learning model for AI-based decision support in skin cancer. *Nat. Med.* **29**, 1941–1946 (2023).
49. Mankowitz, D. J. et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature* **618**, 257–263 (2023).
50. Corso, G., Stark, H., Jegelka, S., Jaakkola, T. & Barzilay, R. Graph neural networks. *Nat. Rev. Methods Primers* **4**, 17 (2024).
51. Li, H. et al. CGMega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection. *Nat. Commun.* **15**, 5997 (2024).
52. Pahud de Mortanges, A. et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *NPJ Digit. Med.* **7**, 195 (2024).
53. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
54. EQUATOR Network. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org/

[1]Institute of Global Health Innovation, Imperial College London, London, UK. [2]Department of Surgery and Cancer, Imperial College London, London, UK. [3]Birmingham Health Partners Centre for Regulatory Science and Innovation, Birmingham, UK. [4]College of Medicine and Health, University of Birmingham, Birmingham, UK. [5]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [6]Google Research, London, UK. [7]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. [8]Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [9]Google Health, Palo Alto, CA, USA. [10]Ottawa Hospital Research Institute, Ottawa, Ontario, Canada. [11]Department of Epidemiology and Data Science, Amsterdam University Medical Centres, Duivendrecht, the Netherlands. [41]These authors contributed equally: Viknesh Sounderajah, Ahmad Guni. [42]These authors jointly supervised this work: Ara Darzi, Hutan Ashrafian.
✉e-mail: h.ashrafian@imperial.ac.uk

STARD-AI Steering Committee

Alan Karthikesalingam[6], Alastair K. Denniston[3,4], Amish Acharya[1,2], Ara Darzi[1,2,42], Bilal A. Mateen[12], Christopher Kelly[6], Daniel Ting[13], Darren Treanor[14,15,16], David Moher[10], Dominic King[17], Felix Greaves[18], Gary S. Collins[5], Hugh Harvey[19], Hutan Ashrafian[1,2,42], Jeffrey De Fauw[20], Jérémie F. Cohen[21,22], Jonathan Godwin[20], Jonathan Pearson-Stuttard[23], Karel Moons[24], Leanne Harling[2], Lena Maier-Hein[25], Lotty Hooft[24], Matthew DF McInnes[26], Nader Rifai[27], Nenad Tomasev[20], Pasha Normahani[2], Patrick M. Bossuyt[11], Penny Whiting[28], Ravi Aggarwal[1,2], Robert M. Golub[8], Sebastian Vollmer[29], Sheraz R. Markar[2], Shravya Shetty[9], Trishan Panch[30], Viknesh Sounderajah[1,2,41] & Xiaoxuan Liu[3,4]

STARD-AI Consensus Group

Alan Karthikesalingam[6], Alastair K. Denniston[3,4], Ben Glocker[31], Darren Treanor[14,15,16], David Taylor[32], David Moher[9], Diana Samuel[33], Gary S. Collins[5], Hugh Harvey[19], Hutan Ashrafian[1,2,42], Jeffrey De Fauw[20], Johan Ordish[34], Karandeep Singh[35], Lena Maier-Hein[25], Leo Celi[36,37,38], Matthew DF McInnes[26], Patrick Bossuyt[11], Robert M. Golub[8], Sherri Rose[39], Shravya Shetty[9], Suchi Saria[40] & Xiaoxuan Liu[3,4]

[12]PATH, London, UK. [13]Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore. [14]Leeds Teaching Hospitals NHS Trust, Leeds, UK. [15]University of Leeds, Leeds, UK. [16]Linkoping University, Linkoping, Sweden. [17]Optum, London, UK. [18]Department of Primary Care and Public Health, Imperial College London, London, UK. [19]Hardian Health, Haywards Heath, UK. [20]DeepMind Technologies, London, UK. [21]INSERM UMR1153 (Centre for Research in Epidemiology and Statistics, CRESS), Université Paris Cité, Paris, France. [22]Department of General Pediatrics and Pediatric Infectious Diseases, Necker-Enfants Malades Hospital, APHP, Université Paris Cité, Paris, France. [23]School of Public Health, Imperial College London, London, UK. [24]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, the Netherlands. [25]Department of Intelligent Medical Systems, German Cancer Research Centre, Heidelberg, Germany. [26]Departments of Radiology and Epidemiology, University of Ottawa, The Ottawa Hospital Research Institute MIR Program, Ottawa, Ontario, Canada. [27]Harvard Medical School, Boston, MA, USA. [28]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [29]The Alan Turing Institute, London, UK. [30]Division of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [31]Department of Computing, Imperial College London, London, UK. [32]Digital Health Council, Royal Society of Medicine, London, UK. [33]The Lancet Digital Health, London, UK. [34]Medicines and Healthcare Products Regulatory Agency, London, UK. [35]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA. [36]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA. [37]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [38]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [39]Department of Health Policy and Center for Health Policy, Stanford University, Stanford, CA, USA. [40]Johns Hopkins University, Baltimore, MD, USA.

Methods

STARD-AI is an international initiative that seeks to provide a multistakeholder consensus on a reporting guideline for AI-centered diagnostic test accuracy studies. A Project Team comprising experts in this field (V.S., X.L., G.S.C., A.K., S.R.M., R.M.G., A.K.D., S. Shetty, D.M., P.M.B., A.D. and H.A.) coordinated the development process, made key methodological decisions and managed day-to-day operations. In addition, a Steering Committee was selected by the Project Team to oversee the guideline development process and provide strategic oversight, consisting of a diverse panel of international stakeholders with expertise in healthcare, computer science, academia, journal editing, epidemiology, statistics, industry, medical regulation and health policymaking. The Consensus Group, distinct from the Project Team and Steering Committee, included invited stakeholders who participated in the Delphi process and consensus meeting. Additional Delphi participants, who were not part of the Consensus Group or committees, contributed to the online survey rounds. The development process is visualized in Fig. 1. A full list of members of the Steering Committee and Consensus Group is provided in a footnote at the end of the article.

STARD-AI was announced in 2020 after the publication of a correspondence highlighting the need for an AI-specific guideline in this field[55]. The initiative to develop the reporting guideline was registered with the EQUATOR Network in June 2020, and its development adhered to the EQUATOR Network toolkit for reporting guidelines[54]. A protocol that outlined the process for developing STARD-AI was subsequently published[56].

Ethical approval was granted by the Imperial College London Joint Research Compliance Office (SETREC reference number: 19IC5679). Written informed consent was obtained from all participants in the online scoping survey, the patient focus group and the Delphi consensus study.

Candidate item generation

A three-stage approach was employed to generate candidate items, consisting of a systematic review, an online survey of experts and a patient and public involvement and engagement (PPIE) exercise. Details of this stage can be found in the study protocol[56]. First, a systematic review was conducted to identify relevant articles. A member of the project team (V.S.) performed a systematic search of MEDLINE and Embase databases through the Ovid platform, as well as a non-systematic exploration of Google Scholar, social networking platforms and articles personally recommended by Project Team members. Two authors (V.S. and H.A.) independently screened abstracts and full texts to identify eligible studies, with any disagreements mediated by discussion. This review built upon the findings of a prior systematic review conducted by members of the STARD-AI team, which evaluated the diagnostic accuracy of deep learning in medical imaging and highlighted widespread variability in study design, methodological quality and reporting practices[32]. Themes and material extracted from included articles were used to establish considerations unique to AI-based diagnostic accuracy studies and to highlight possible additions, removals or amendments to STARD 2015 items. These considerations were subsequently framed as potential candidate items.

Second, an online survey of 80 international experts was carried out. This generated over 2,500 responses, relating to existing STARD 2015 items and potential new items or considerations. Experts were selected to reflect the full diagnostic AI continuum, including those with expertise in conventional diagnostic modalities, AI development and statistical methods, for diagnostic accuracy. This breadth of expertise was intended to ensure that candidate items reflected both the technical and clinical aspects of AI-centered diagnostic evaluation. Responses were grouped thematically to generate candidate items. Patients and members of the public were then invited to an online focus group through Zoom (Zoom Video Communications) in order to provide input as part of a PPIE exercise. This provided a patient perspective on issues that were not uncovered during the literature review or expert survey. Although no new domains were introduced from the PPIE exercise, participants placed increased emphasis on the importance of ethics and fairness, particularly in relation to how AI may impact different patient subgroups or exacerbate existing health disparities. As these elements were not a major focus of the original STARD guideline, their prioritization during the consensus process helped to refine the framing and inclusion of items in the final checklist. A list of 55 items, including 10 terminology-related items and 45 candidate checklist items, was finalized by the Project Team and Steering Committee and entered the modified Delphi consensus process.

Modified Delphi consensus process

Experts were invited to join the STARD-AI Consensus Group and participate in the online Delphi surveys as well as the consensus meeting. The Project Team and Steering Committee identified participants on the basis of being a key stakeholder, ensuring to account for a diversity in geographics and demographics to maintain a representative panel. All invited participants were provided with written information about the study and given 3 weeks to respond to the initial invitation. The Delphi process included more than 240 international participants, including healthcare professionals, clinician scientists, academics, computer scientists, machine learning engineers, statisticians, epidemiologists, journal editors, industry leaders, health regulators, funders, patients, ethicists and health policymakers.

The first two rounds of the Delphi process were online surveys conducted on DelphiManager software (version 4.0), which is maintained by the Core Outcome Measures in Effectiveness Trials (COMET) initiative. Participants were asked to rate each item on a five-point Likert scale (1, very important; 2, important; 3, moderately important; 4, slightly important; 5, not at all important). Items receiving 75% or higher ratings of 'very important' or 'important' were immediately put forward for discussion in the final round. Items achieving 75% or more responses of 'slightly important' or 'not at all important' were excluded. Items that did not achieve either threshold were entered into the next round of the Delphi process. The 75% threshold was pre-set before the beginning of the process. Participants were also given the opportunity to provide free-text comments on any of the items considered or to suggest new items. These were used by the Project Team to rephrase, merge or generate new items for subsequent rounds. The stakeholder groups represented in the Delphi rounds are outlined in Supplementary Table 1. A full list of participants in the online survey and Delphi rounds is provided in the Supplementary Note.

The first round was conducted between 6 January and 20 February 2021. Invitations were extended to 528 participants in total, of whom 240 responded (response rate of 45%). Of the participants who responded, 209 fully completed the survey (completion rate of 87%). Forty-five candidate checklist items were rated after the multistage evidence generation process. Free-text comments were collected for these items and also for the 10 terminology items. Twenty-three candidate items achieved consensus for 'very important' or 'important' and were formally moved into the consensus meeting. Fifteen items were removed or replaced by an amended item based on participant feedback. Seven items did not achieve consensus, and 19 additional items were constructed after feedback from participants, resulting in 26 total items put forward to the second round. The second round was conducted between 21 April and 4 June 2021. Invitations were sent to 235 participants, of whom 203 responded (response rate of 86%), and 143 completed the survey (completion rate of 70%). Users were again asked to rate each item and add free-text comments. A majority consensus was achieved for 22 items.

Forty-five items reached consensus over the first two rounds. As this was deemed too many to include in an instrument, a pre-consensus survey consisting of 37 members of the Project Team, Steering Committee and other key external stakeholders was conducted to agree on

a final list of items for discussion in the consensus meeting, receiving a 100% response rate. Participants were asked to rate whether each item should be included in the instrument as a standalone item, included in the accompanying explanation and elaboration document or excluded from the process. Twenty-two items received a majority consensus for inclusion in the final checklist; 13 items did not reach the 75% predefined threshold; and 10 items were excluded from the process. In total, 35 items were finalized for discussion at the consensus meeting.

The virtual consensus meeting took place on 1 November 2021 and was chaired by D.M. An information sheet was pre-circulated to all participants, and individual consent was obtained. In total, 22 delegates representing all of the key stakeholder groups attended the meeting. Items were discussed in turn to gain insight into content that warrants inclusion in the checklist, particularly focusing on the 13 items that did not reach consensus from the Delphi process. Voting on each item was anonymized using the Mentimeter software platform. After this, a meeting among key members of the Steering Committee finalized the checklist based on the outcome of the consensus meeting.

References

55. Sounderajah, V. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat. Med.* **26**, 807–808 (2020).
56. Sounderajah, V. et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* **11**, e047709 (2021).

Acknowledgements

We thank the participants who were involved in the online survey and the Delphi study (Supplementary Note). Infrastructure support for this research was provided by the National Institute for Health and Care Research Imperial Biomedical Research Centre. G.S.C. was supported by Cancer Research UK (program grant: C49297/A27294).

Author contributions

V.S., X.L., G.S.C., S. Shetty, R.M.G., A.K.D., D.M., P.M.B., A.D. and H.A. were involved in the conception and design of the study. The Steering Committee (A.K., A.K.D., A.D., A.A., B.A.M., C.K., D. Ting, D. Treanor, D.M., D.K., F.G., G.S.C., H.H., H.A., J.D.F., J.F.C., J.G., J.P.-S., K.M., L. Harling, L.M.-H., L. Hooft, M.M., N.R., N.T., P.N., P.M.B., P.W., R.A., R.M.G., S.V., S.R.M., S. Shetty, T.P., V.S. and X.L.) oversaw the development of the guideline and the direction of the study at all stages. The Consensus Group (A.K., A.K.D., B.G., D. Treanor, D. Taylor, D.S., G.S.C., H.H., H.A., J.D.F,. J.O., K.S., L.M.-H., L.C., M.M., P.M.B., R.M.G., S.R., S. Shetty, S. Saria and X.L.) decided the final content of the checklist during the consensus meeting, which was chaired by D.M. A.G. drafted the manuscript, with edits and comments from all authors. All authors approved the final manuscript.

Competing interests

V.S., A.K., S. Shetty and C.K. are employees of Alphabet. H.A. is the Chief Scientific Officer of Preemptive Health and Medicine, Flagship Pioneering. A.D. is the Executive Chair for Preemptive Health and Medicine, Flagship Pioneering. G.S.C. is a National Institute for Health and Care Research (NIHR) Senior Investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. D.S. is employed as Acting Deputy Editor of *The Lancet Digital Health*.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-03953-8.

Correspondence and requests for materials should be addressed to Hutan Ashrafian.

Peer review information *Nature Medicine* thanks Harald Kittler, Danielle Bitterman and Tien Yin Wong for their contribution to the peer review of this work. Primary Handling Editor: Karen O'Leary, in collaboration with the *Nature Medicine* team.