



Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare

Susan Cheng Shelmerdine ¹, Owen J Arthurs,¹ Alastair Denniston,²
Neil J Sebire ³

To cite: Shelmerdine SC, Arthurs OJ, Denniston A, *et al*. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health Care Inform* 2021;**28**:e100385. doi:10.1136/bmjhci-2021-100385

Received 22 April 2021
Accepted 09 August 2021

ABSTRACT

High-quality research is essential in guiding evidence-based care, and should be reported in a way that is reproducible, transparent and where appropriate, provide sufficient detail for inclusion in future meta-analyses. Reporting guidelines for various study designs have been widely used for clinical (and preclinical) studies, consisting of checklists with a minimum set of points for inclusion. With the recent rise in volume of research using artificial intelligence (AI), additional factors need to be evaluated, which do not neatly conform to traditional reporting guidelines (eg, details relating to technical algorithm development). In this review, reporting guidelines are highlighted to promote awareness of essential content required for studies evaluating AI interventions in healthcare. These include published and in progress extensions to well-known reporting guidelines such as Standard Protocol Items: Recommendations for Interventional Trials-AI (study protocols), Consolidated Standards of Reporting Trials-AI (randomised controlled trials), Standards for Reporting of Diagnostic Accuracy Studies-AI (diagnostic accuracy studies) and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis-AI (prediction model studies). Additionally there are a number of guidelines that consider AI for health interventions more generally (eg, Checklist for Artificial Intelligence in Medical Imaging (CLAIM), minimum information (MI)-CLAIM, MI for Medical AI Reporting) or address a specific element such as the 'learning curve' (Developmental and Exploratory Clinical Investigation of Decision-AI). Economic evaluation of AI health interventions is not currently addressed, and may benefit from extension to an existing guideline. In the face of a rapid influx of studies of AI health interventions, reporting guidelines help ensure that investigators and those appraising studies consider both the well-recognised elements of good study design and reporting, while also adequately addressing new challenges posed by AI-specific elements.

INTRODUCTION

Recent, rapid developments in computational technologies and increased volumes of digital data for analysis have resulted in an unprecedented growth in research activities relating to artificial intelligence (AI), particularly within healthcare. This volume

of work has even led to several high impact journals launching their own subjournals within the 'AI healthcare' field (eg, *Nature Machine Intelligence*,¹ *Lancet Digital Health*,² *Radiology: Artificial Intelligence*).³ High-quality research should be accompanied by transparency, reproducibility and validity of techniques for adequate evaluation and translation into clinical practice. Standardised reporting guidelines help researchers define key components of their study, ensuring that relevant information is provided in the final publication.⁴ Studies pertaining to algorithm development and clinical application of AI however, have brought unique challenges and added complexities in how such studies are reported, assessed and compared in relation to elements that are not conventionally prespecified in traditional reporting guidelines. This could lead to missing information and high risk of hidden bias. If these actual or potential limitations are not identified, then it may lead to tacit approval through publication which in turn may support premature adoption of new technologies.^{5 6} Conversely well-designed, well-delivered studies that are poorly reported may be judged unfavourably due to being adjudged to have a high risk of bias, simply due to a lack of information.

Inadequacies of reporting of AI clinical studies are increasingly well-recognised. In 2019, a systematic review by Liu *et al*⁷ reviewed over 20 500 articles, but found that fewer than 1% of these were sufficiently robust in their design and reporting allowing independent reviewers to have confidence in their claims. Similarly Nagendran *et al*⁸ identified high levels of bias in the field. In another study,⁹ it was reported that only 6% of over 500 eligible radiological-AI research publications performed any external validation of their models, and none used multicentre or prospective data collection. Similarly most studies using machine learning (ML) models



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Radiology, Great Ormond Street Hospital NHS Foundation Trust, London, UK

²Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK

³Digital Research, Informatics and Virtual Environments Unit (DRIVE), London, UK

Correspondence to

Dr Susan Cheng Shelmerdine; susie.shelmerdine@gmail.com

for medical diagnosis¹⁰ did not have adequate detail on how these were evaluated nor sufficient detail for these to be reproduced. Inconsistencies in how ML models from electronic health records have also been reported, with details regarding race and ethnicity of participants omitted in 64% of studies, and only 12% of models being externally validated.¹¹

In order to address these concerns, adapted research reporting guidelines based on the well-established EQUATOR Network (Enhancing the QUALity and Transparency Of health Research)^{12 13} and de novo recommendations by individual societies have been published, with a greater relevance for AI research. In this review, we highlight those that will cover the majority of healthcare focused AI-related studies, and explain how they differ to the well-known guidance for non-AI related clinical work. Our intention is to raise awareness of how such studies should be structured, thereby improving the quality of future submissions and providing a helpful aid for researchers, peer reviewers and editors.

In compiling a detailed, yet relevant list of study guidelines, we reviewed the EQUATOR network¹³ website for those containing the terms AI, ML or deep learning. A separate search was also conducted using Medline, Scopus and Google Scholar databases for publications using the same search terms with the addition of 'reporting guideline', 'checklist' or 'template'. Opinion pieces were excluded. Articles were included where the description of the recommendations were provided, and published at time of the search (March 2021).

TYPES OF RESEARCH REPORTING GUIDELINES

An ideal reporting guideline should be a clear, structured tool with a minimum list of key information to include within a published scientific manuscript. The EQUATOR Network¹³ is the international 'standard bearer' for reporting guidelines, committed to improving 'the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines'. Since the landmark publication of Consolidated Standards of Reporting Trials (CONSORT),¹⁴ the network has overseen the development and publication of a number of guidelines that address other types of study design (eg, diagnostic accuracy studies). The EQUATOR guidelines are centrally registered (available via a core library) which ensures adherence to robust methodology of development and avoids redundancy of parallel initiatives to address the same issue. Importantly these guidelines are not medical specialty specific but are focused on the type of study, which helps ensure that there is a consistent approach and quality for addressing the same study design. It is recognised that certain specific scenarios may require specific extensions to these guidelines. For example, the increasing recognition of the importance of patient-reported outcomes (PROs) has led to the development of Standard Protocol Items: Recommendations for Interventional Trials

(SPIRIT-PRO)¹⁵ and CONSORT-PRO.¹⁶ In a similar way, the specific attributes of AI as an intervention, has led to a number of AI extensions, both published and in process, which build on the robust methodology of the original EQUATOR guidelines, while ensuring AI-specific elements are also addressed.

In parallel to the work of the EQUATOR network, a number of experts and institutions have developed their own recommendations for good practice and reporting. In contrast, these start with the intervention (ie, AI) rather than the study type (ie, randomised controlled trial (RCT)), and therefore, cover essentially the same territory. They vary in depth, and there can be differences in nuance depending on their primary purpose. For example some have originated from the need to support reviewers and editorial staff ('is this complete and is it good enough?'), whereas others are addressing at building a shared understanding of appropriate design and delivery ('this is what good looks like').

Given the number of different reporting guidelines in this area, there is value in setting them in context to help support users in understanding which is most appropriate for a particular setting (table 1). Ultimately the most important elements of a high-quality study are contained within the methodology of the study design itself and not within the intervention. It is these elements that help minimise the major biases that all studies must address. In line with leading journals, we would, therefore, recommend starting with the guideline that addresses that particular study design (eg, CONSORT¹⁴ for an RCT). If an AI extension is already in existence for that study type then these are clearly appropriate for that study (eg, CONSORT-AI).¹⁷⁻¹⁹ If no such -AI extension exists then we recommend using the appropriate EQUATOR guideline (eg, Standards for Reporting of Diagnostic Accuracy Studies (STARD)²⁰ for diagnostic accuracy studies), but supplementing with AI-specific elements recommended in other guidelines (eg, SPIRIT-AI,²¹⁻²³ CONSORT-AI¹⁷⁻¹⁹ or the non-EQUATOR guidelines described below). Indeed all the guidelines considered here contain valuable insights into the specific challenges of AI studies, and are recommended reading into good practice for design and reporting.

EQUATOR NETWORK GUIDELINES

Clinical trials protocols

The quality of a study and the trustworthiness of its findings, starts at the design phase. The study protocol should contain all elements of the study design, sufficient for independent groups to carry out the study and expect replicability. Prepublication of the study protocol, helps avoid biases such as post-hoc assignment of the primary outcome in which the trialist can 'cherry pick' one of a number of outcomes that point in the desired direction.

Guidance for recommended items to include in a trial protocol are provided by the SPIRIT Statement (latest version published in 2013),²⁴ which has been recently

Table 1 Summary of reporting guidelines for common study types used in radiological research, and their corresponding guideline extensions where these involve artificial intelligence

| Study design | Reporting guideline | Latest version | AI-related extension | Date of AI-extension published |
|--|---------------------|----------------|--------------------------------|--------------------------------|
| Clinical Trial Protocol | SPIRIT | 2013 | SPIRIT-AI | September 2020 |
| Diagnostic Accuracy Studies | STARD | 2015 | STARD-AI | Expected 2021 |
| | | | CLAIM | March 2020 |
| | | | MINIMAR | June 2020 |
| Prediction models for diagnostic or prognostication purposes | TRIPOD | 2015 | TRIPOD –AI/ML | Expected 2021 |
| | PROBAST | 2019 | PROBAST-ML | Expected 2021 |
| Randomised Controlled Trials (Interventional Study Design) | CONSORT | 2010 | CONSORT-AI | September 2020 |
| Systematic reviews and meta-analyses | PRISMA | 2009 | None planned or announced | |
| | PRISMA-DTA | 2018 | | |
| Critical appraisal and data extraction of publications relating to prediction models | CHARMS | 2014 | Applicable to machine learning | |
| Evaluation of human factors in early algorithm deployment | Not applicable | | DECIDE-AI | Expected 2021/2022 |

AI, artificial intelligence; CHARMS, Checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CONSORT, Consolidated Standards of Reporting Trials; DECIDE-AI, Developmental and Exploratory Clinical Investigation of Decision-support systems driven by Artificial Intelligence; DTA, Diagnostic Trials of Accuracy; MINIMAR, Minimum Information for Medical AI Reporting; ML, machine learning; PRISMA, Preferred Reporting Items for Systematic Review and Meta-analysis; PROBAST, Prediction model Risk Of Bias Assessment Tool; SPIRIT, Standard Protocol Items: Recommendations for Interventional Trials; STARD, Standards for Reporting of Diagnostic Accuracy Studies; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

adapted for trials with an AI-related focus, termed the ‘SPIRIT-AI’ guideline.^{21–23} This adaptation includes an additional 15 items (12 extensions, 3 elaborations) to the existing 33-item SPIRIT 2013 guideline. The key differences are outlined in table 2, mostly focused on the methodology of the trial, (accounting for eight extensions, one elaboration) with emphasis on inclusion/exclusion of data and participants, dealing with poor quality data and how the AI intervention will be applied to and benefit clinical practice.

Clinical trials reports

While most AI studies are currently at early-phase validation stages, those evaluating the use of ‘AI-interventions’ in real world setting are fast emerging, and will become of increasing importance, since these are required for real-world clinical benefit demonstration. RCTs are the exemplar study design in providing a robust evidence basis for efficacy and safety of a given intervention, with the CONSORT statement, 2010 version¹⁴ providing a 25-item checklist for the minimum reporting content in such studies. An adapted version, entitled the ‘CONSORT-AI’ extension^{17–19} was published in September 2020 for ‘AI intervention’ studies. This includes an additional 14 items (11 extensions, 3 elaborations) to the existing CONSORT 2010 statement, the majority of which (8 extensions, 1 elaboration) relate to the study participants and details of the ‘AI intervention’ being evaluated, which are similar to those additions already described in the SPIRIT-AI extension. Specific key differences in the new guideline

are outlined in table 3. Although not specific for AI interventions, some aspects of the checklist Template for Intervention Description and Replication, 2014²⁵ may be a helpful addition when reporting details of the interventional elements of a study (ie, as an extension of item 5 of the CONSORT 2010 statement or as item 11 of the SPIRIT 2013 statement). These include details regarding any modifications of the intervention during a study, including how and why certain aspects were personalised or adapted. There are currently no publicly proposed plans to publish an ‘AI’ extension to this guideline to the best of our knowledge.

Diagnostic accuracy studies

The STARD statement, 2015 version²⁰ is the most widely accepted reporting standard for diagnostic accuracy studies. A steering group has been established to devise an AI-specific extension to the latest version of the 30-item STARD statement (called the STARD-AI extension.²⁶ At the time of writing this is undergoing an international consensus survey among leaders in the AI field for suggested adaptations and pending publication.

Prediction models

Extensions to reporting guidelines describing prediction models that use ML have been announced, and are anticipated for publication soon. These include adapted versions of the ‘Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis’ (TRIPOD), 2015 version,²⁷ which will be entitled

Table 2 Additional items proposed for studies relating to AI intervention clinical protocols within the SPIRIT-AI statement (in addition to the SPIRIT 2013 statement)

| Section | Item no | SPIRIT 2013 item | Amendment | SPIRIT-AI item |
|---|---------|--|-------------|--|
| Administrative information | | | | |
| Title | 1 | Descriptive title identifying the study design, population, interventions and if applicable, trial acronym | Elaboration | Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model. |
| | | | Elaboration | Specify the intended use of the AI intervention. |
| Introduction | | | | |
| Background and rationale | 6a | Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention | Extension | Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (eg, healthcare professionals, patients, public). |
| | | | Extension | Describe any pre-existing evidence for the AI intervention. |
| Methods: Participants, interventions and outcomes | | | | |
| Study Setting | 9 | Description of study settings (eg, community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained | Extension | Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting. |
| Eligibility criteria | 10 | Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centres and individuals who will perform the interventions (eg, surgeons, psychotherapists) | Elaboration | State the inclusion and exclusion criteria at the level of participants. |
| | | | Extension | State the inclusion and exclusion criteria at the level of the input data. |
| Interventions | 11a | Interventions for each group with sufficient detail to allow replication, including how and when they will be administered | Extension | State which version of the AI algorithm will be used. |
| | | | Extension | Specify the procedure for acquiring and selecting the input data for the AI intervention. |
| | | | Extension | Specify the procedure for assessing and handling poor quality or unavailable input data. |
| | | | Extension | Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users. |
| | | | Extension | Specify the output of the AI intervention. |
| | | | Extension | Explain the procedure for how the AI intervention's output will contribute to decision making or other elements of clinical practice. |
| Methods: Monitoring | | | | |
| Harms | 22 | Plans for collecting, assessing, reporting and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct | Extension | Specify any plans to identify and analyse performance errors. If there are no plans for this, justify why not. |
| Access to data | 29 | Statement of who will have access to the final trial dataset and disclosure of contractual agreements that limit such access for investigators | Extension | State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or reuse. |

Table adapted from Cruz Rivera *et al.*^{21–23} Items within the SPIRIT 2013 statement that have not changed for the SPIRIT-AI statement have been omitted.

AI, artificial intelligence; SPIRIT, Standard Protocol Items: Recommendations for Interventional Trials.

Table 3 Additional criteria to be included for studies relating to AI interventions within the CONSORT-AI statement (in addition to the CONSORT 2010 statement)

| Section | Item no | CONSORT 2010 item | Amendment | CONSORT-AI item |
|---------------------------|---------|---|-------------|--|
| Title and abstract | | | | |
| Title and abstract | 1a | Identification as a randomised trial in the title | Elaboration | Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model. |
| | 1b | Structured summary of trial design, methods, results and conclusions | Elaboration | State the intended use of the AI intervention within the trial in the title and/or abstract. |
| Introduction | | | | |
| Background and objectives | 2a | Scientific background and explanation of rationale | Extension | Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (eg, healthcare professionals, patients, public). |
| Methods | | | | |
| Participants | 4a | Eligibility criteria for participants | Elaboration | State the inclusion and exclusion criteria at the level of participants. |
| | | | Extension | State the inclusion and exclusion criteria at the level of the input data. |
| | 4b | Settings and locations where the data were collected | Extension | Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements. |
| Interventions | 5 | The interventions for each group with sufficient details to allow replication, including how and when they were actually administered | Extension | State which version of the AI algorithm was used. |
| | | | Extension | Describe how the input data were acquired and selected for the AI intervention. |
| | | | Extension | Describe how poor quality or unavailable input data were assessed and handled |
| | | | Extension | Specify whether there was human–AI interaction in the handling of the input data, and what level of expertise was required of users. |
| | | | Extension | Specify the output of the AI intervention. |
| | | | Extension | Explain how the AI intervention’s outputs contributed to decision making or other elements of clinical practice. |
| Results | | | | |
| Harms | 19 | | Extension | Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not. |
| Discussion | | | | |
| Funding | 25 | | Extension | State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use. |

Table adapted from Liu *et al.*^{17–19} Items within the CONSORT 2010 statement that have not been changed for the CONSORT-AI statement have been omitted.

AI, artificial intelligence; CONSORT, Consolidated Standards of Reporting Trials.

‘TRIPOD-AI’,^{28 29} and supported by the ‘Prediction model Risk Of Bias Assessment Tool’ (PROBAST, 2019 version)³⁰ which is proposed to be entitled PROBAST-ML.^{28 29}

Human factors

Another upcoming guideline, focused on the evaluation of the ‘human factors’ in algorithm implementation, has been announced: the checklist (Developmental and Exploratory Clinical Investigation of Decision-support systems driven by AI).³¹ This checklist is intended for use in early small-scale clinical trials that evaluate and provide information on how algorithms may be used in practice, bridging the gap between the algorithm development/

validation stage (which would follow TRIPOD-AI, STARD-AI or Checklist for Artificial Intelligence in Medical Imaging (CLAIM)), but before large-scale clinical trials of AI interventions (where the CONSORT-AI would be used). Publication is anticipated to be late 2021 or early 2022.

Systematic reviews

Given the increasing volume of radiological AI-related research for a growing variety of conditions and clinical settings, it is also likely that we will encounter more systematic reviews and meta-analyses that aim to aggregate the evidence from studies in this field (eg, recent

publications have already emerged that summarise research regarding the role of AI in COVID-19.^{32–34} At present, the ‘Preferred Reporting Items for Systematic Reviews and Meta-analyses’ (PRISMA), 2009³⁵ guidelines are the most established for systematic reviews and meta-analyses, with a modified version specifically tailored for meta-analyses relating to diagnostic test accuracies (ie, the PRISMA-Diagnostic Trials of Accuracy (DTA), 2018).³⁶ Currently, there have not been any announcements for an update to these guidelines for AI-related systematic reviews or meta-analyses, and therefore, it is suggested that the PRISMA 2009³⁵ or PRISMA-DTA 2018³⁶ guidance should be followed.

In the planning stages for conducting systematic reviews of prediction models, the ‘Checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies’ (CHARMS, 2014)³⁷ was developed by the Cochrane Prognosis Methods Group. This was not intentionally created for publications relating to AI *per se*, but applicable to a wide range of studies, which also happen to include the evaluation of ML models. The developers provide the checklist to help authors frame their review question, design and extract relevant items from published reports of prediction models and guide assessment of risk of bias (rather than in the analysis of these). This checklist will, therefore, be useful to those who wish to plan a review of AI tools that provide a ‘risk score’ or ‘probability of diagnosis’. A tutorial on how to carry out a ‘CHARMS analysis’ for prognostic multivariate models with real-life worked examples has been published³⁸ and may be a helpful resource for readers wishing to carry out similar work. It is worth noting that the authors of CHARMS still recommend reference to the PRISMA 2009³⁵ and PRISMA-DTA 2018³⁶ statements for the reporting and analysis of trial results, in conjunction with their own checklist for planning of the review design.

OTHER (NON-EQUATOR NETWORK) GUIDELINES

Alternative guidelines have been published by expert interest groups and endorsed by different specialty societies. A few are described here to supplement further reading and interest.

The Radiological Society of North America recently published the ‘CLAIM’³⁹ in 2020, containing elements of the STARD 2015 guideline and applicable for trials addressing a wide spectrum of AI applications using medical images (eg, classification, reconstruction, text analysis, workflow optimisation). This checklist comprises of 42 items, of which 6 are new (pertaining to model design and training), 8 are extensions of pre-existing STARD 2015 items, 14 items are elaborations (mostly relating to methods and results) and 14 items remain the same. Particular emphasis is given to data, the reference standard of ‘ground truth’ and the precise development and methodology of the AI algorithm being tested. These are listed in further detail in [table 4](#), where differences to the STARD 2015 are highlighted. Care should be taken

to avoid any confusion with another similarly named checklist entitled ‘minimum information about clinical AI modelling’ (MI-CLAIM),⁴⁰ which is less of a reporting guideline but a document outlining required shared understanding in the development and evaluation of AI models aimed to serve clinical and data scientists), repository managers and model users.

It is also worth noting that the American Medical Informatics Association produced a set of guidelines in 2020 termed the ‘MI for Medical AI Reporting’ (MINIMAR),⁴¹ specific to studies reporting the use of AI solutions in healthcare. Rather than a list of items for manuscript writing, this guidance provides suggestions for details pertaining to data sources used in algorithm development and their intended usage, spread across four key subject areas (ie, study population and setting, patient demographics, model architecture and model evaluation). There are many similarities with the aforementioned CLAIM checklist, although the key differences include the granularity by which the MINIMAR suggests researchers should explicitly state participant demographics (eg, ethnicity and socioeconomic status, rather than just age and sex) and how code and data can be shared with the wider community.

FURTHER READING

There is an increasing need to build a cadre of researchers and reviewers with sufficient domain knowledge of technical aspects (including limitations and risk) and of the principles of good trial methodology (including areas of potential bias, analysis issues, etc). There is also a need for ML experts and clinical trial communities to increasingly learn each other’s language, to ensure accurate and precise communication of concepts, and enable comparison between studies. A number of reviews are highlighted here for further reading^{42–46} along with work⁴⁷ explaining different evaluation metrics used in AI and ML studies. It is also worth bearing in mind the wider clinical and ethical context of how any AI tool would fit into our existing clinical pathways and healthcare systems.⁴⁸

CONCLUSION

In conclusion, this article has provided readers an overview of changes to standard clinical reporting guidelines specific for AI-related studies. The fundamental basics of describing the trial setup, inclusion and exclusion criteria, detailing the study methodology and standards used, together with details on algorithm development, should create transparency and address reproducibility. Those which are most relevant for a particular healthcare specialty will depend on the type of research being conducted in that particular field (eg, guidelines for AI-related diagnostic accuracy trials may be more relevant for radiological or pathological specialties, whereas those addressing patient outcomes with the aid of an

Table 4 Criteria for the CLAIM checklist for diagnostic accuracy studies using AI

| Section | Item no | STARD 2015 item | Amendment | CLAIM item |
|---------------------------|---------|--|--|---|
| Title and abstract | | | | |
| Title | 1 | Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values or AUC). | Elaboration | Identification as a study of AI methodology, specifying the category of technology used (eg, deep learning). |
| Abstract | 2 | Structured summary of study design, methods, results and conclusions. | Same | |
| Introduction | | | | |
| Background | 3 | Scientific and clinical background, including the intended use and clinical role of the index test. | Elaboration | Scientific and clinical background, including the intended use and clinical role of the AI approach. |
| Objectives | 4 | Study objectives and hypotheses. | Same | |
| Methods | | | | |
| Study design | 5 | Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study). | Same Extension | Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial. |
| Participants | 6 | Eligibility criteria (inclusion/exclusion). | Extension | State data sources. |
| | 7 | On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry). | Same | |
| | 8 | Where and when potentially eligible participants were identified (setting, location and dates). | | |
| | 9 | Whether participants formed a consecutive, random or convenience series. | Extension Extension Extension Extension | Data preprocessing steps. Selection of data subsets, if applicable. Definitions of data elements, with references to common data elements. Deidentification methods. |
| Test methods | 10b | Reference standard, in sufficient detail to allow replication. | Elaboration | Definition of 'ground truth' (ie, reference standard), in sufficient detail to allow replication. |
| | | | Elaboration | Source of ground truth annotations; qualifications and preparation of annotators. |
| | | | Elaboration | Annotation tools. |
| | 11 | Rationale for choosing the reference standard (if alternatives exist). | Same | |
| | 12b | Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing prespecified from exploratory. | Elaboration | Measurement of inter-rater and intrarater variability; methods to mitigate variability and/or resolve discrepancies for ground truth. |
| Model | | | New New New | Detailed description of model, including inputs, outputs, all intermediate layers and connections. Software libraries, frameworks, and packages. Initialisation of model parameters (eg, randomisation, transfer learning). |

Continued

Table 4 Continued

| Section | Item no | STARD 2015 item | Amendment | CLAIM item |
|-------------------|---------|--|-------------|---|
| Training | | | New | Details of training approach, including data augmentation, hyperparameters, number of models trained. |
| | | | New | Method of selecting the final model. |
| | | | New | Ensembling techniques, if applicable |
| Analysis | 14 | Methods for estimating or comparing measures of diagnostic accuracy. | Elaboration | Metrics of model performance. |
| | 16 | How missing data on the index test and reference standard were handled. | Same | |
| | 17 | Any analyses of variability in diagnostic accuracy, distinguishing prespecified from exploratory. | Elaboration | Statistical measures of significance and uncertainty (eg, CIs). |
| | | | Elaboration | Robustness or sensitivity analysis. |
| | | | Elaboration | Methods for explainability or interpretability (eg, saliency maps) and how they were validated. |
| | 18 | Intended sample size and how it was determined. | Elaboration | Validation or testing on external data. |
| | | | Same | |
| | | | Extension | How data were assigned to partitions; specify proportions. |
| | | | Extension | Level at which partitions are disjoint (eg, image, study, patient, institution). |
| Results | | | | |
| Participants | 19 | Flow of participants, using a diagram. | Same | |
| | 20 | Baseline demographic and clinical characteristics of participants. | Elaboration | Demographic and clinical characteristics of cases in each partition. |
| Test results | 23 | Cross tabulation of the index test results (or their distribution) by the results of the reference standard. | Elaboration | Performance metrics for optimal model(s) on all data partitions. |
| | 24 | Estimates of diagnostic accuracy and their precision (such as 95% CIs). | Same | |
| | 25 | Any adverse events from performing the index test or the reference standard. | Elaboration | Failure analysis of incorrectly classified cases. |
| Discussion | | | | |
| Limitations | 26 | Study limitations, including sources of potential bias, statistical uncertainty and generalisability. | Same | |
| Implications | 27 | Implications for practice, including the intended use and clinical role of the index test. | Same | |
| Other Information | | | | |
| Registration | 28 | Registration no and name of registry. | Same | |
| Protocol | 29 | Where the full study protocol can be accessed. | Same | |
| Funding | 30 | Sources of funding and other support; role of funders. | Same | |

This is based on the STARD 2015 guidelines,²⁰ demonstrating which aspects are new, the same or elaborated on. Items not included in the CLAIM checklist (which were previously present in the STARD guideline) have been removed. Table adapted from Bossuyt *et al*²⁰ and Mongan *et al*.³⁹

AI, artificial intelligence; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; STARD, Standards for Reporting of Diagnostic Accuracy Studies.

AI algorithm may be more relevant for oncological or surgical specialties).

Although the reporting guidelines outlined may seem comprehensive, there remain areas that will need to be addressed, such as for economic health evaluation of AI-tools and algorithms (many are currently developed for 'pharmacoeconomic evaluations').⁴⁹ It is likely that future guidelines may take the form of an extension to the widely used CHEERS guidance (Consolidated Health Economic Evaluation Reporting Standards^{50 51} available via the EQUATOR network.¹³ Nevertheless, a wide variation in opinion regarding the most appropriate economic evaluation guideline already exists for non-AI related tools, and this may be reflected in future iterations of such guidelines depending on how the algorithms are funded in different healthcare systems.⁵²

The current guidelines outlined here will likely continue to be updated in the light of new understanding of the specific challenges of AI as an intervention and, how traditional study designs and reports need to be adapted.

Funding OJA is funded by a National Institute for Health Research (NIHR) Career Development Fellowship (NIHR-CDF-2017-10-037). SS, OJA and NJS receive funding from the Great Ormond Street Children's Charity and the Great Ormond Street Hospital NIHR Biomedical Research Centre. AD receives funding from Health Data Research UK. An initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations and leading medical research charities.

Disclaimer The funding source(s) did not have any direct involvement in the methodology, design or write-up of this review article.

Competing interests None declared.

Patient and public involvement statement Not required

Patient consent for publication Not required.

Ethics approval Not required

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Susan Cheng Shelmerdine <http://orcid.org/0000-0001-6642-9967>

Neil J Sebire <http://orcid.org/0000-0001-5348-9063>

REFERENCES

- More than machines. *Nat Mach Intell* 2019;1.
- The Lancet Digital Health. A digital (r)evolution: introducing The Lancet Digital Health. *Lancet Digit Health* 2019;1:e1.
- Kahn CE, Charles E, Kahn J. Artificial intelligence, real radiology. *Radiol Artif Intell* 2019;1:e184001.
- Moher D. Reporting guidelines: doing better for readers. *BMC Med* 2018;16:233.
- Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the radiology editorial board. *Radiology* 2020;294:487-9.
- CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019;25:1467-8.
- Liu X, Faes L, Kale AU, 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:e271-97.
- Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting Standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
- Kim DW, Jang HY, Kim KW, et al. 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;20:405-10.
- Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* 2020;10:e034568.
- Bozkurt S, Cahan EM, Seneviratne MG, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J Am Med Inform Assoc* 2020;27:1878-84.
- Altman DG, Simera I, Hoey J, et al. EQUATOR: reporting guidelines for health research. *Lancet* 2008;371:1149-50.
- The EQUATOR Network. Enhancing the quality and transparency of health research. Available: <https://www.equator-network.org> [Accessed 22 Mar 2021].
- Moher D, Hopewell S, Schulz KF, et al. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
- Calvert M, Kyte D, Mercieca-Bebber R, et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO extension. *JAMA* 2018;319:483-94.
- Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT pro extension. *JAMA* 2013;309:814-22.
- Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;2:e537-48.
- Liu X, Rivera SC, Moher D. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370:m3164.
- Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364-74.
- Bossuyt PM, Reitsma JB, Bruns DE, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology* 2015;277:826-32.
- Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2:e549-60.
- Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26:1351-63.
- Rivera SC, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210.
- Chan A-W, Tetzlaff JM, Gøtzsche PC, et al. Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013;346:e7586.
- Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687.
- Sunderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. *Nat Med* 2020;26:807-8.
- Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
- Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577-9.
- Andaur Navarro CL, Damen JAAG, Takada T, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open* 2020;10:e038832.
- Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1.
- DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186-7.

- 32 Albahri OS, Zaidan AA, Albahri AS, *et al.* Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects. *J Infect Public Health* 2020;13:1381–96.
- 33 Li WT, Ma J, Shende N, *et al.* Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med Inform Decis Mak* 2020;20:247.
- 34 Syeda HB, Syed M, Sexton KW, *et al.* Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. *JMIR Med Inform* 2021;9:e23811.
- 35 Liberati A, Altman DG, Tetzlaff J, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
- 36 McInnes MDF, Moher D, Thombs BD, *et al.* Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;319:388–96.
- 37 Moons KGM, de Groot JAH, Bouwmeester W, *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Med* 2014;11:e1001744.
- 38 Palazón-Bru A, Martín-Pérez F, Mares-García E, *et al.* A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models. *Stat Med* 2020;39:3207–25.
- 39 Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. *Radiol Artif Intell* 2020;2:e200029.
- 40 Norgeot B, Quer G, Beaulieu-Jones BK, *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020;26:1320–4.
- 41 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, *et al.* MINIMAR (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 2020;27:2011–5.
- 42 Luo W, Phung D, Tran T, *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18:e323.
- 43 Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med* 2020;3:126.
- 44 Kocak B, Kus EA, Kilickesmez O. How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. *Eur Radiol* 2021;31:1819–30.
- 45 Faes L, Liu X, Wagner SK, *et al.* A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. *Transl Vis Sci Technol* 2020;9:7.
- 46 Do S, Song KD, Chung JW. Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean J Radiol* 2020;21:33–41.
- 47 Handelman GS, Kok HK, Chandra RV, *et al.* Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol* 2019;212:38–43.
- 48 McCradden MD, Joshi S, Mazwi M, *et al.* Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health* 2020;2:e221–3.
- 49 Sullivan SM, Wells G, Coyle D. What guidance are economists given on how to present economic evaluations for policymakers? A systematic review. *Value Health* 2015;18:915–24.
- 50 Husereau D, Drummond M, Petrou S, *et al.* Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value Health* 2013;16:231–50.
- 51 Husereau D, Drummond M, Petrou S, *et al.* Consolidated health economic evaluation reporting standards (cheers) statement. *Value Health* 2013;16:e1–5.
- 52 Sharma D, Aggarwal AK, Downey LE, *et al.* National healthcare economic evaluation guidelines: a Cross-Country comparison. *Pharmacoecon Open* 2021;5:349–64.