BMJ Open

Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review

Rida Shahzad,[1] Bushra Ayub,[2] M A Rehman Siddiqui [3]

To cite: Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. *BMJ Open* 2022;**12**:e061519. doi:10.1136/bmjopen-2022-061519

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-061519).

Received 30 January 2022
Accepted 17 August 2022

[1]Department of Ophthalmology, Shahzad Eye Hospital, Karachi, Pakistan
[2]Centre for Clinical Best Practices, Aga Khan University Hospital, Karachi, Pakistan
[3]Department of Ophthalmology and Visual Sciences, Aga Khan University Hospital, Karachi, Pakistan

Correspondence to
Dr M A Rehman Siddiqui; rehman.siddiqui@gmail.com

ABSTRACT

Objectives The aim of this study was to evaluate the quality of reporting of randomised controlled trials (RCTs) of artificial intelligence (AI) in healthcare against Consolidated Standards of Reporting Trials—AI (CONSORT-AI) guidelines.

Design Systematic review.

Data sources We searched PubMed and EMBASE databases for studies reported from January 2015 to December 2021.

Eligibility criteria We included RCTs reported in English that used AI as the intervention. Protocols, conference abstracts, studies on robotics and studies related to medical education were excluded.

Data extraction The included studies were graded using the CONSORT-AI checklist, comprising 43 items, by two independent graders. The results were tabulated and descriptive statistics were reported.

Results We screened 1501 potential abstracts, of which 112 full-text articles were reviewed for eligibility. A total of 42 studies were included. The number of participants ranged from 22 to 2352. Only two items of the CONSORT-AI items were fully reported in all studies. Five items were not applicable in more than 85% of the studies. Nineteen per cent (8/42) of the studies did not report more than 50% (21/43) of the CONSORT-AI checklist items.

Conclusions The quality of reporting of RCTs in AI is suboptimal. As reporting is variable in existing RCTs, caution should be exercised in interpreting the findings of some studies.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This systematic review assesses the reporting of randomised trials of artificial intelligence (AI) interventions across medical fields from 2015 to 2021 against Consolidated Standards of Reporting Trials—AI (CONSORT-AI) guidance, establishing a baseline for future studies.
⇒ We did not separately analyse publications from before and after the publication of the CONSORT-AI guidance in September 2020, so were unable to assess whether there was any change in reporting quality following publication of the guidance.
⇒ Only two databases were searched and only English-language publications were eligible for inclusion.

INTRODUCTION

Artificial intelligence (AI) is finding increased utility in the medical realm, with a special emphasis on deep learning. Medical applications of AI range from screening, diagnosis, prognosis and generation of management plans.[1–5] For example, AI has been extensively studied in ophthalmology for various diseases such as diabetic retinopathy,[6] age-related macular degeneration[7] and glaucoma.[8] However, increased hype associated with AI—without sound evidence base—may result in inappropriate clinical decisions, which can potentially be detrimental to healthcare.[9]

Randomised controlled trials (RCTs) are one of the highest quality of evidence used by clinicians in decision making regarding interventions.[10] RCTs may be susceptible to various forms of biases. Adequate reporting of RCTs is vital to allow results and conclusions derived from a study to be assessed critically by readers.[11 12]

The Consolidated Standards of Reporting Trials (CONSORT) statement was introduced in 1996 to establish guidelines to improve the reporting quality of clinical trials. Additionally, the CONSORT statement is a useful guide that helps readers with the critical appraisal of RCTs to ascertain their reliability and clinical applicability.[13] The most recent update of the CONSORT statement was published in 2010, listing 25 minimum reporting requirements.[14] Several extensions to CONSORT also exist, which cater to certain specific study designs.[15–18]

There has been an exponential increase in AI-based healthcare studies in recent years due to rapid advances in computational power. However, the methodological rigour has not kept pace with the development in technology. For example, the design and quality of reporting in these studies have not always been adequate.[19 20] CONSORT-AI was published on 9 September 2020 as an extension of the CONSORT 2010 statement to evaluate RCTs involving AI. Fourteen new items

were added to the checklist—including 11 extensions and 3 elaborations.[21 22] These items mostly relate to the AI intervention in question and are necessary to independently evaluate and replicate the trial.

The aim of this study was to evaluate the quality of reporting of RCTs of AI intervention for medical conditions, published from 2015 to 2021, based on CONSORT-AI guidelines. While CONSORT-AI did not exist for much of this timeline, this study will serve as a baseline measure of reporting quality for comparison with future studies' adherence to CONSORT-AI guidelines.

METHODS

Search strategy

We performed a systematic review of RCTs of AI for medical conditions published from January 2015 to December 2021. The search date range was initially set as an arbitrary period of 5 years from 2015 to 2020; the literature search was later updated to include publications until December 2021. RCTs of AI in healthcare are a nascent field, and we expected very few RCTs of AI in healthcare prior to 2015. We searched PubMed and EMBASE databases for potential studies. The PubMed search was conducted using the MeSH terms: "artificial intelligence", "machine learning" and "deep learning". The terms "artificial intelligence", "deep learning" and "machine learning" were searched in EMBASE. In both the databases, the search was limited to RCTs, publications in the English language, from the year 2015 to 2021 and human subjects (online supplemental appendix 1).

Screening and study selection

The records were screened by two independent investigators (RS and BA) for potential inclusion. The abstracts of RCTs using AI, deep learning and machine learning were further evaluated for possible inclusion. Protocols, conference abstracts, studies on robotics and post hoc analyses of RCTs were excluded.

Full-text articles of all shortlisted abstracts were then screened for eligibility. Publications were included if AI was used as an intervention for a medical condition, if there was a comparator control group in the study and if there was evidence of randomisation. In case of a disagreement, a senior reviewer assessed the full text and the disagreement was resolved with consensus. The exclusion criteria were non-randomised studies, secondary studies, post hoc analyses, or if the intervention investigated was not AI. Additionally, if the target condition was not a medical disease or if the research pertained to medical education, the study was excluded.

Assessment against CONSORT-AI guidance

The CONSORT-AI checklist of 43 items (online supplemental table 1) was used to grade the included studies. Each item was scored fully, partially or not reported. If an item was irrelevant to a particular study, it was labelled as 'not applicable'. Each publication was scored by two

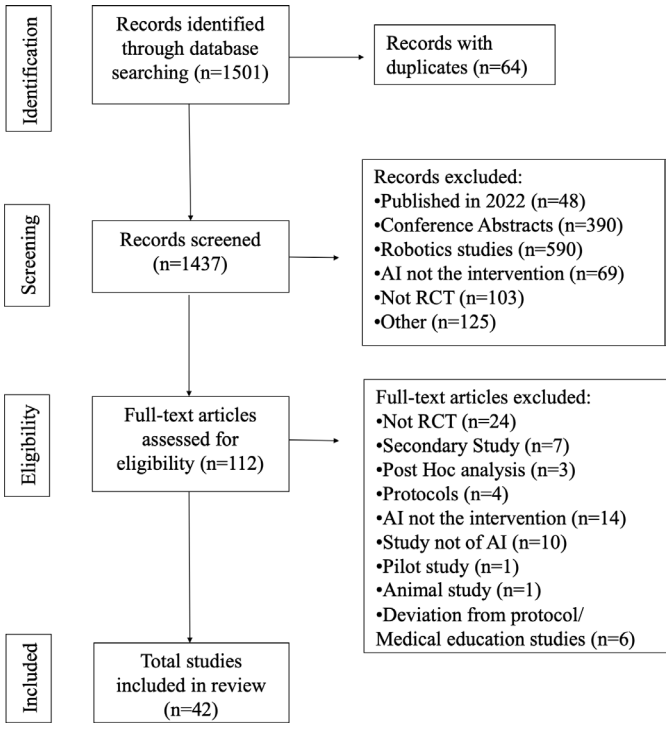

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart. AI, artificial intelligence; RCT, randomised controlled trial.

trained graders (RS and BA) independently. Differences were discussed with the senior reviewer (MARS) to reach a consensus.

The results were tabulated by writing all the reported items as the numerator and the total number of applicable items as the denominator. The descriptive statistics for the study population and clinical characteristics are reported. The only deviation from the initial protocol for the review was the extension of the search until December 2021 to keep this review up-to-date.

Patient and public involvement

None.

RESULTS

Study selection

The initial search identified 1501 potential records. One hundred and twelve articles were considered as potentially eligible after screening of abstracts. Following a review of full-text manuscripts, a total of 42 manuscripts were included in the systematic review (figure 1).

General characteristics

The included studies (online supplemental table 2) were from the years 2016 to 2021(figure 2). The number of participants ranged from 22 to 2352. They pertained to various medical fields, including gastroenterology (n=12), medicine (n=6), cardiology (n=5), psychiatry (n=4), ophthalmology (n=2), endocrinology (n=2), paediatrics (n=2), oncology (n=2), orthopaedics (n=2), surgery (n=1), radiology (n=1), neurology (n=1), pulmonology

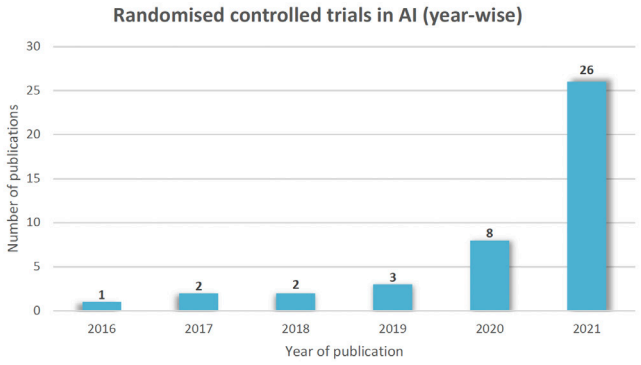

Figure 2 Yearwise distribution of RCTs in AI. AI, artificial intelligence; RCT, randomised controlled trial.

(n=1) and dentistry (n=1). Studies were from different parts of the world, including China (n=16), USA (n=14), Japan (n=3), UK (n=2), Spain (n=2), Netherlands (n=1), Germany (n=1), Korea (n=1), Denmark (n=1) and Israel (n=1). (figure 3)

Adherence to reporting standards

The median number of fully reported CONSORT-AI checklist items in the included studies was 30 (range 7–37) of a possible total of 43. Overall, only 2 (items #1b, and 21) out of possible 43 items were fully reported in all 42 studies. Five items (items #3b, 6b, 7b, 14b and 17b) were deemed not applicable in more than 85% of the included studies. The two least reported items were item #5iii (not reported in 36/42 studies) and item #24 (not reported in 31/42 studies). Nineteen per cent (8/42) of included studies did not report more than 50% (21/43) of the CONSORT-AI checklist items. The reporting of each item is given in table 1.

DISCUSSION

In our review, variable reporting standards of RCTs of AI in healthcare were observed. While some items were reported adequately—for example, those relating to the abstract and introduction of the manuscript—other items particularly in the Methods section, had poor reporting scores.

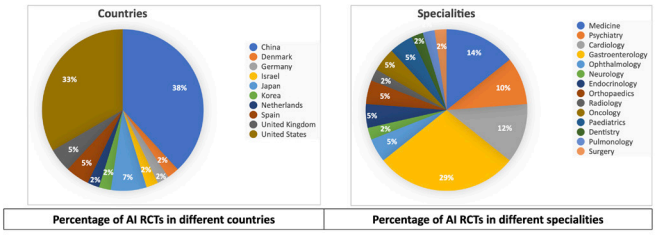

Figure 3 Percentage of AI RCTs in different countries and specialties. AI, artificial intelligence; RCT, randomised controlled trial.

Our results reinforce previously published findings. In a systematic review conducted by Liu *et al*, it was seen that sufficient reporting and external validation were done in less than one-third of the included 82 deep learning studies, thereby limiting their reliability.[23] Similarly, Nagendran *et al* also found deviations from reporting standards, with less than 50% adherence to 12/29 items in the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines, and high levels of bias in AI studies.[20] Bozkurt *et al* reported that demographic specifics of study populations were poorly reported in studies developing Machine learning (ML) models from electronic health records, and external validation was omitted in 88% of the models.[24] In another systematic review of 28 articles regarding machine learning models for medical diagnosis, Yusuf *et al* discovered that all studies in their systematic review failed to follow reporting guidelines.[25] Our study also revealed variable reporting of CONSORT-AI items in RCTs of AI in healthcare, suggesting there is still room in AI studies for further improving the quality of their reporting.

The CONSORT-AI checklist was developed to encourage transparent reporting of RCTs in AI. The extensions and elaborations added to the original CONSORT guideline largely emphasise the peculiarities related to AI intervention itself and its clinical application. These include details of the interventions, such as algorithm version, input and output data, how the intervention was integrated into the trial and whether there was human and AI interaction. This information is crucial for the critical appraisal of a study and facilitates the replication of clinical trials.[23] These items had variable reporting scores in our study (items 4a to 5vi). Twenty-seven out of 42 (64%) studies did not mention the version of the AI algorithm used. This could confuse the reader regarding which version to apply the study findings to because an AI algorithm is likely to undergo multiple updates.[21] Moreover, information regarding input data were largely missed in the majority of included studies; with only 35% (15/42) of the studies identifying the inclusion and exclusion criteria at the level of the input data, and a mere 14% (6/42) of studies reported how poor quality or unavailable input data was handled and assessed. Such details are essential, as the overall performance of any given AI intervention relies on the quality of input data. Additionally, this information allows an evaluator to distinguish AI platforms that may only work in ideal conditions from those which can be applied to real-world settings.[26 27]

On the other hand, items regarding human–AI interaction and required expertise level, as well as AI output were fully reported by majority of studies (37 and 41/42, respectively). Clarity about the human–AI interface is essential to ensure a standard approach and functional safety, as well as to avoid ethical implications.[28 29] For example, it is essential that qualified experts can interpret dynamically complex variables exhibited by AI interfaces which are related to patients as well as the clinical

Table 1 CONSORT-AI scores of included studies

	Item	Fully reported	Partially reported	Not reported	Not applicable
Title and abstract	1a, 1a(i)	41	1	0	0
	1b, 1b(ii)	42	0	0	0
Introduction					
Background and objectives	2a, 2a(i)	41	1	0	0
	2b	38	0	4	0
Methods					
Trial design	3a	26	6	10	0
	3b	6	0	0	36
Participants	4ai	39	0	3	0
	4aii	15	0	27	0
	4b	40	0	2	0
Intervention	5i	15	0	27	0
	5ii	34	0	8	0
	5iii	6	0	36	0
	5iv	37	0	5	0
	5v	41	0	1	0
	5vi	31	0	11	0
Outcomes	6a	39	0	3	0
	6b	2	0	0	40
Sample size	7a	30	0	11	1
	7b	2	0	0	40
Sequence generation	8a	34	0	8	0
	8b	25	0	17	0
Randomisation					
Allocation concealment mechanism	9	24	0	18	0
Implementation	10	18	3	21	0
Blinding	11a	24	0	18	0
	11b	23	0	17	2
Statistical methods	12a	39	0	3	0
	12b	34	0	8	0
Results					
Participant flow	13a	32	2	8	0
	13b	29	1	12	0
Recruitment	14a	38	0	4	0
	14b	1	0	0	41
Baseline data	15	32	0	10	0
Numbers analysed	16	32	1	9	0
Outcomes and estimation	17a	31	3	8	0
	17b	1	0	0	41
Ancillary analyses	18	33	0	9	0
Harms	19	4	11	27	0
Discussion					
Limitations	20	36	0	6	0
Generalisability	21	42	0	0	0

Continued

Table 1 Continued

	Item	Fully reported	Partially reported	Not reported	Not applicable
Interpretation	22	41	0	1	0
Other information					
Registration	23	35	0	7	0
Protocol	24	11	0	31	0
Funding	25	10	20	12	0

CONSORT-AI, Consolidated Standards of Reporting Trials—artificial intelligence.

context—only then it is possible that AI platforms enable an improvement in clinicians' decision-making process.[30] It is encouraging to see most authors report these items clearly.

Interestingly, although missing out on important information regarding the details of AI intervention, 42/42 of the studies were promising generalisability of their findings in the clinical setting. The generalisability of AI systems may be limited, especially when used in the real-world setting outside of the environment they were developed in.[31 32] Therefore, caution must be employed when evaluating such studies.

An important factor to consider about CONSORT-AI, however, is the applicability of each item to clinical trials. Five items of the CONSORT-AI checklist were deemed to be not applicable in the majority of studies evaluated. Three of these items referred to changes made to methods and outcomes after trial commencement, and why the trial was ended (items 3b, 6b and 14b). These items pertain to modifications made in the protocol, which was not the case in most included studies.

Another item not applicable to most of the included studies was an explanation about any interim analysis and stopping guidelines. Since AI is a relatively recent advance in healthcare, harms and adverse events from AI have not been clearly defined yet. Perhaps this is the reason stopping guidelines were not reported in 40 out of 42 included studies. This ties closely to item 19: which requires reporting of adverse events in AI trials and a description of the analysis of performance errors. AI platforms can make errors that can be difficult to predict and go beyond human judgement, but may have harmful effects if employed on a large scale.[31] Only 4/42 studies fully reported this item, even though it is important to report information about error and outline risk mitigation strategies to decide which settings and populations the AI intervention can be safely employed in.[21] These points emphasise that AI clinical trials in healthcare have not integrated the concept of harm related to AI intervention to determine appropriate stopping guidelines.

Certain general observations were made regarding the included RCTs in our review. There was a large range of sample size (22–2352) in the studies. This wide range suggests that a standard approach to sample size calculation is not practised in RCTs of AI. For example, the diagnostic accuracy of healthcare professionals is often set higher than that of AI while employing sample size estimation, which presumes that AI is inferior to humans.[33] It is recommended that sample size calculations are performed using a non-inferior design by setting a more suitable non-inferiority margin, of diagnostic accuracy, for example, 5%.[34] Similarly, the majority of the studies took place in China, and were focused on gastroenterology, making them less representative of other fields and perhaps other parts of the world.

There are some limitations to our review. Potential eligible studies could have been missed in the inclusion process, as only two databases were searched, and only English-language publications were eligible for inclusion. The majority of the included studies were published before the CONSORT-AI checklist was widely available. As such, most study authors would not have been able to use the guidance to inform their reporting. Furthermore, trial reports from before and after the publication of the CONSORT-AI guidance were not analyses separately, so we were not able to assess whether there was any improvement in reporting quality following publication of the guidance.

In conclusion, the standards of reporting in RCTs of AI were variable. We found certain important information regarding the AI intervention was insufficiently reported in many studies. Therefore, caution should be employed by healthcare service providers and policymakers when using these studies to inform decision making.

Twitter M A Rehman Siddiqui @RehmanSiddiqui

Contributors The idea for the study was conceived and planned by MARS. RS and BA carried out the literature review process including screening of abstracts and review of full-text articles, while MARS acted as a senior reviewer. RS and BA independently scored the included studies using the CONSORT-AI checklist and disagreements were resolved following discussions with MARS. The manuscript was prepared by RS and BA and reviewed by MARS. All authors reviewed and approved the final manuscript. MARS is the gaurantor of the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

M A Rehman Siddiqui http://orcid.org/0000-0001-5100-3189

REFERENCES

1 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
2 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
3 Rajpurkar P, Irvin J, Ball RL, *et al*. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686.
4 Tyler NS, Mosquera-Lopez CM, Wilson LM, *et al*. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metab* 2020;2:612–9.
5 Kim H, Goo JM, Lee KH, *et al*. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. *Radiology* 2020;296:216–24.
6 Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA* 2016;316:2366–7.
7 von der Emde L, Pfau M, Dysli C, *et al*. Artificial intelligence for morphology-based function prediction in neovascular age-related macular degeneration. *Sci Rep* 2019;9:1–2.
8 Devalla SK, Liang Z, Pham TH, *et al*. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol* 2020;104:301–11.
9 Andaur Navarro CL, Damen JAA, Takada T, *et al*. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021;375:n2281.
10 Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996;313:570–1.
11 Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323:42–6.
12 Schulz KF, Chalmers I, Hayes RJ, *et al*. Empirical evidence of bias. dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12.
13 Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637–9.
14 Schulz KF, Altman DG, Moher D. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials* 2010;11:1–8.
15 Boutron I, Altman DG, Moher D, *et al*. Consort statement for randomized trials of nonpharmacologic treatments: a 2017 update and a consort extension for nonpharmacologic trial Abstracts. *Ann Intern Med* 2017;167:40–7.
16 Hopewell S, Clarke M, Moher D, *et al*. Consort for reporting randomised trials in Journal and conference Abstracts. *The Lancet* 2008;371:281–3.
17 Gagnier JJ, Boon H, Rochon P, *et al*. Reporting randomized, controlled trials of herbal interventions: an elaborated consort statement. *Ann Intern Med* 2006;144:364–7.
18 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.
19 Gregory J, Welliver S, Chong J. Top 10 reviewer critiques of radiology artificial intelligence (AI) articles: qualitative thematic analysis of reviewer critiques of machine learning/deep learning manuscripts submitted to JMRI. *J Magn Reson Imaging* 2020;52:248–54.
20 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.
21 CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019;25:1467–8.
22 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370:m3164.
23 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.
24 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.
25 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.
26 Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. *Radiol Artif Intell* 2020;2:e190015.
27 Ibrahim H, Liu X, Rivera SC, *et al*. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 2021;22:1–5.
28 Wiens J, Saria S, Sendak M, *et al*. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337–40.
29 Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ* 2020;98:251–6.
30 Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018;320:2199–200.
31 Kelly CJ, Karthikesalingam A, Suleyman M, *et al*. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:1–9.
32 Pooch EH, Ballester P, Barros RC. Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification. In: *International workshop on thoracic image analysis*. Springer, Cham, 2020: 74–83.
33 Lin H, Li R, Liu Z, *et al*. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52–9.
34 Zhou Q, Cao Y-H, Chen Z-H. Optimizing the study design of clinical trials to identify the efficacy of artificial intelligence tools in clinical practices. *EClinicalMedicine* 2019;16:10–11.