BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-061519
Article Type:	Original research
Date Submitted by the Author:	30-Jan-2022
Complete List of Authors:	Shahzad, Rida; Shahzad Eye Hospital Ayub, Bushra; Aga Khan University Hospital Siddiqui, M. A. Rehman; Aga Khan University Hospital, Department of Ophthalmology and Visual Sciences
Keywords:	STATISTICS & RESEARCH METHODS, Clinical trials < THERAPEUTICS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title: Quality of reporting of randomized controlled trials in artificial intelligence: a systematic review

Authors:

Rida Shahzad[1]

Bushra Ayub[2]

M. A. Rehman Siddiqui[1,3,4]

Authors Affiliations:

1. Shahzad Eye Hospital, Karachi, Pakistan
2. Aga Khan University Hospital, Karachi, Pakistan
3. Section of Ophthalmology, Department of Surgery, Aga Khan University Hospital, Karachi, Pakistan
4. The Eye Centre, South City Hospital, Karachi, Pakistan

Corresponding author:

M. A. Rehman Siddiqui

Section of Ophthalmology, Department of Surgery,

Aga Khan University Hospital,

National Stadium Road, P.O. Box 3500,

Karachi, Pakistan

Email: rehman.siddiqui@gmail.com

Funding: None

Keywords:

CONSORT, CONSORT-AI, artificial intelligence, deep learning, machine learning, randomized controlled trials.

SYNOPSIS

Quality of reporting of randomised controlled trials of artificial intelligence in health care is suboptimal.

ABSTRACT

Objectives: The aim of this study was to evaluate quality of reporting of randomised controlled trials (RCTs) of artificial intelligence (AI) in health care from 2015 to 2020 against Consolidated Standards of Reporting Trials – Artificial Intelligence (CONSORT-AI) guidelines.

Methods: We searched PubMed and EMBASE databases to identify eligible studies from 2015 to 2020. The included studies were graded using the CONSORT-AI checklist comprising of 43 items by two independent graders. The results were tabulated and descriptive statistics were reported.

Results: We screened 939 potential abstracts, of which 73 full-text articles were reviewed for eligibility. A total of 15 studies were included. Number of participants ranged from 28 to 1058. Studies pertained to medical fields including medicine (n = 2), psychiatry (n = 3), gastroenterology (n = 5), cardiology (n = 2), ophthalmology (n = 1), endocrinology (n = 1), and neurology (n =1). Studies were from countries including China (n = 6), United States (n = 6), United Kingdom (n = 1), Netherlands (n = 1), and Israel (n = 1). Only 3 items of the CONSORT-AI items were fully reported in all studies. Five items were not applicable in more than 85% of the studies. Twenty per cent of the studies did not report more than 50% of the CONSORT-AI checklist items.

Conclusions: Quality of reporting of RCTs in AI in suboptimal. There is a high risk of bias in existing RCTs, therefore caution must be taken when generalizing the findings of these RCTs in real-world settings.

ARTICLE SUMMARY

Strengths and Limitations of the Study:

- This systematic review is the first study to evaluate quality of reporting of RCTs in AI of all medical fields using the CONSORT-AI guidelines.
- It helps to assess applicability of AI algorithms in real world clinical settings.
- A limitation of this study is the utilization of two databases, covering a span of 5 years only.

INTRODUCTION

Artificial intelligence (AI) is a field of immense potential in healthcare. With an increase in computational power and portability, AI is finding increased utility in the medical realm, with a special emphasis on deep learning and neural networks. 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 - and subsequent failure to deliver, may result in disappointment.[9]

Randomized controlled trials (RCTs) are 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, so that the results and conclusions derived from a study may be assessed critically by the readers.[11,12]

The CONSORT (Consolidated Standards of Reporting Trials) statement was introduced in 1996 to establish guidelines to improve the reporting quality of clinical trials. Additionally, 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]

RCTs of AI have unique characteristics. Issues have been voiced regarding the design and reporting of such studies.[19,20] CONSORT-AI was recently published 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]

The aim of this study was to evaluate the quality of reporting of RCTs of AI intervention for medical conditions, published from 2015 to 2020, based on CONSORT-AI guidelines.

METHODS

We performed a systematic review of RCTs of AI for medical conditions published between January 2015 and December 2020. 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 term “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 2020, and human subjects (Appendix 1). The records were screened by two independent investigators for potential inclusion. The abstracts of RCTs using artificial intelligence, deep

learning, and machine learning were further evaluated for possible inclusion. Protocols, studies on robotics, and post-hoc analyses of randomized controlled trials 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 there was evidence of randomization. In case of a disagreement, a senior reviewer assessed the full text and the disagreement was resolved with consensus. The exclusion criteria were non-randomized 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.

The CONSORT-AI checklist of 43 items (Supplementary 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 labeled as “not applicable”. Each publication was scored by two trained graders independently. Differences were discussed with the senior reviewer 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 protocol of this study was submitted to PROSPERO in February 2021.

RESULTS

Study selection

The initial search identified 939 potential records. Seventy-three articles were considered as potentially eligible after screening of abstracts. Following a review of full-text manuscripts, a total of 15 manuscripts were included in the systematic review (Figure 1).

General characteristics

The included studies (Supplementary table 2) were from the years 2016 to 2020. The number of participants ranged from 28 to 1058. They pertained to various medical fields, including medicine (n = 2), psychiatry (n = 3), gastroenterology (n = 5), cardiology (n = 2), ophthalmology (n = 1), endocrinology (n = 1), and neurology (n =1). Studies were from different parts of the world, including China (n = 6), United States (n = 6), United Kingdom (n = 1), Netherlands (n = 1), and Israel (n = 1). There were no multicenter trials.

Adherence to reporting standards

The median number of fully reported CONSORT-AI checklist items in the included studies was 26 (range 7-36) of a possible total of 43. Overall, only 3 (items # 1b, 4b, and 21) out of possible 43 items were fully reported in all 15 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 14/15 studies) and item #24 (not reported in 12/15 studies). Twenty per cent (3/15) of included studies did not report more than 50% of the CONSORT-AI checklist items. The reporting of each item is given in Table 1.

Table 1: Summary of CONSORT-AI Items:

	Item	Fully Reported	Partially Reported	Not Reported	Not Applicable	Applicable Items Total Score	Applicable Items Total (%)
Title and Abstract	1a, 1a(i)	14	1	0	0	15	100
	1b, 1b(ii)	15	0	0	0	15	100
Introduction							
Background and objectives	2a, 2a(i)	14	1	0	0	15	100
	2b	13	0	2	0	13	87
Methods							
Trial Design	3a	9	2	4	0	11	73
	3b	2	0	0	13	2	100
Participants	4ai	13	0	2	0	13	87
	4aii	4	0	11	0	4	27
	4b	15	0	0	0	15	100
Intervention	5i	4	0	11	0	4	27
	5ii	10	0	5	0	10	67
	5iii	1	0	14	0	1	7
	5iv	14	0	1	0	14	93
	5v	14	0	1	0	14	93
	5vi	11	0	4	0	11	73
Outcomes	6a	13	0	2	0	13	87
	6b	0	0	0	15	0	0
Sample size	7a	10	0	5	0	10	67
	7b	1	0	0	14	1	100

Sequence generation	8a	12	0	3	0	12	80
	8b	9	0	6	0	9	60
Randomisation							
Allocation concealment mechanism	9	6	0	9	0	6	40
Implementation	10	7	3	5	0	10	67
Blinding	11a	9	0	6	0	9	60
	11b	7	0	6	2	7	47
Statistical methods	12a	13	0	2	0	13	87
	12b	11	0	4	0	11	73
Results							
Participant flow	13a	12	2	1	0	14	93
	13b	11	1	3	0	12	80
Recruitment	14a	13	0	2	0	13	87
	14b	0	0	0	15	0	0
Baseline data	15	13	0	2	0	13	87
Numbers analyzed	16	11	1	3	0	12	80
Outcomes and estimation	17a	9	3	3	0	12	80
	17b	1	0	0	14	1	7
Ancillary analyses	18	10	0	5	0	10	67
Harms	19	0	5	10	0	5	33
Discussion							
Limitations	20	13	0	2	0	13	87
Generalizability	21	15	0	0	0	15	100
Interpretation	22	14	0	1	0	14	93

Other information							
Registration	23	14	0	1	0	14	93
Protocol	24	3	0	12	0	3	20
Funding	25	4	7	4	0	11	73

The most frequently reported items were: intended use of the AI intervention within the trial in the title and/or abstract [item# 1b], settings and locations where the data were collected [item# 4b], and generalizability of the findings [item# 21].

Patient and Public Involvement

No patient or public was involved in this study.

Ethics approval statement

This study is a systematic review which did not involve human subjects. Therefore, no ethics approval was required.

DISCUSSION

In our Review, the quality of reporting of RCTs varied widely. For example, the item regarding title and abstract (item #1a), authors of all studies successfully identified them as RCTs of AI. Similarly, all studies fully reported the intended use of the AI intervention within the trial (item#1b). The importance of clearly reporting these items is for easy identification of trials during literature search and future systematic review. A recent review suggested that good scientific manuscript titles which convey the main role of their respective studies is positively related to their clarity and impact.[23]

Likewise, items regarding the introduction of the article and description of AI intervention in the clinical pathway had a 100% reporting score, with 93% being fully and 7% being partially reported (item#2a). However, two out of fifteen studies failed to report their specific objectives or hypothesis (item #2b). Adequate reporting of these factors reflects clarity on the role of AI in their respective fields. A well-formulated introduction would inform the reader of relevant background knowledge regarding the goal of the intervention, in addition to giving them essential knowledge on how it was applied and what results it was expected to yield.[24] It is important to note, however, that these items can be easily reported within the manuscript by performing a good background literature review. The main purpose of ensuring appropriate reporting of these items is to ensure that the intended use and context of the AI intervention within the study is well understood by the reader.

When looking at the adherence to reporting standards in the methods section, a large range was observed, with 0%-100% of studies fully reporting each applicable item. Informing the readers regarding the AI intervention in question is essential to aid in understanding its role and in recognizing some of the difficulties faced with its implementation. A large number of studies were suboptimal in reporting certain items related to the AI intervention itself. Eleven out of fifteen studies did not mention the version of the AI algorithm used (item#5i). This could potentially confuse the reader regarding which version to apply the findings of the study to because an AI algorithm is likely to undergo multiple updates.[21]

Especially poor reporting standards were observed for items related to input data. Information about the input data helps readers understand the kind of data provided to train the algorithm to yield a given output. The overall performance of any given AI intervention is reliant on the

quality of input data.[25] However, only 4/15 studies identified the inclusion and exclusion criteria at the level of the input data. 33% of the studies did not report how the input data was acquired for the intervention. Moreover, a mere 7% of studies reported how poor quality or unavailable input data was handled and assessed (item#5iii). Such details are important, because they allow an evaluator to distinguish the AI platforms that are only work in ideal conditions from those which can be applied to real-world settings.[26]

Some items regarding the AI intervention [items 5iv, 5v] were reported adequately in most studies. Fourteen out of 15 studies successfully reported both the output of the AI intervention, along with indicating human-AI interaction and the level of expertise required by users. Clarity on the human-AI interface is essential to ensure a standard approach, as well as to avoid ethical implications, such as being functionally safe.[27,28] 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 context – only then it is possible that they enable an improvement in clinicians' decision-making process.[29] Finally, the type of outputs generated by an AI intervention directly impacts its clinical usability and any subsequent incidents encountered in the process.

Transparency in reporting the methods of a trial is also vital to inform the reader of potential biases in a study and evaluate the applicability of a particular intervention. Missing information may potentially result in readers making incorrect inferences from the findings of a study. While most authors successfully reported items like the location of data collection (100% of the studies), a large percentage of studies missed reporting vital components of the methodology. An essential component of RCTs is the process of randomization itself, to avoid selection bias.[30] However, 40% of the studies did not report the type of randomization

performed (item#8b), and 60% did not report the mechanism employed for random allocation concealment(item#9). Additionally, 40% of the studies did not report information regarding blinding following intervention assignment to study and control groups(item#11a). Only 47% of the studies stated similarities between the interventions assigned to the study and control groups(item#11b), which is important to educate the reader of how the AI intervention differs from standard management protocol. A systematic review of glioma treatment revealed that CONSORT items regarding randomization and blinding were reported in less than 25% of the studies, suggesting that these items lack optimal reporting.[31] Incomplete reporting of these items leave large gaps in reporting of RCTs, and results in a lack of information needed to critically appraise the trial itself.

Presentation and reporting of results (items# 13-19) were made adequately in most of the studies, including participant flow, baseline data, as well as primary and secondary outcomes. However, only 5 out of 15 studies reported the presence or absence of important harms and unintended effects in each group (item# 19). This is an important item to report for the readers to assess potential risks of the intervention. Moreover, ten studies did not describe the results of any analysis of performance errors and how errors were identified. This is an important item to discuss especially as AI platforms can make errors that may be difficult to predict but may have disastrous effects if employed on a large scale.[32] It is therefore important to report cases of error and outline risk mitigation strategies to conclude which settings and populations the AI intervention can be safely executed in.[21]

All studies sufficiently reported the CONSORT-AI items about the discussion section (items# 20-22) Interestingly, 15/15 of the studies were quick to offer promising generalizability of their findings in the clinical setting. Generalizability of AI systems can be limited, especially

when used in the real-world setting outside of the environment they were developed in.[32,33] Therefore it is imperative that the generalizability of such studies be evaluated with caution. Two studies failed to list their limitations (item#20).

Observations regarding the applicability of the CONSORT-AI checklist itself were also made. Five of the items of the checklist were not applicable in more than 85% of the included studies. These included items referring to changes made to methods and outcomes after trial commencement, explanation of any interim analyses and stopping guidelines, why the trial was ended, and presentation of binary outcomes. As can be seen, most of these items pertain to changes made in the trial, which is not the case in most trials. Caution should be employed when scoring the study using CONSORT-AI with regards to these items, as they would result in the underestimation of the overall score of any particular study. A consideration that can be made in the future is to update these items in order to make them more applicable to RCTs in AI.

Another observation was that AI extensions of a few items were unrelated to the CONSORT item itself. For example, the item regarding harms was extended with the inquiry regarding the results of analysis of performance errors. Similarly, the item about funding of the study had the AI extension of accessibility of the AI intervention and/or its code, again making the two unrelated and irrelevant to each other. Full reporting of one of these two components would render the entire item as partially reported, as was seen in 7/15 of the studies in our review. A future update to the guidelines could consider making these items independent to each other, or by explicitly outlining their relevance to one another.

Certain general observations were also made regarding the included RCTs in our review. The sample size in the studies ranged from 28 to 1058. Sample size calculations were reported in only 10/15 of the studies. This wide range suggests that a standard approach to sample size calculation is not practiced in RCTs of AI. Diagnostic accuracy of human professionals is often set higher than that of AI while employing sample size estimation, which means that AI is inferior to human controls.[34] 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, of for example 5%.[35] 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 of our review. Potential eligible studies could have been missed in the inclusion process, as only two databases were searched. Furthermore, this review is a screenshot in time; additional AI RCTs could have been performed following our search final search date which are not included.

In conclusion, the standard of reporting in AI RCTs was suboptimal. We found that many trials were at high risk of bias, and data information was insufficient in most studies. Therefore, caution must be employed while generalizing the findings and applicability of such studies in the real world. Some factors related to poor reporting were related to the CONSORT-AI checklist itself. Improvement of reporting standards may be achieved in the future by addressing these variables.

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 senior reviewer. RS and BA independently scored the included studies using the CONSORT-AI checklist and disagreements were resolved following a combined discussion with MARS. The manuscript was prepared by RS and BA and edited by MARS. All authors reviewed and approved the final manuscript.

Competing interests

None declared

Funding

None

Data sharing statement

All data are freely available.

REFERENCES:

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89-94.

2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-8.

3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 2018;15(11):e1002686.

4. Tyler NS, Mosquera-Lopez CM, Wilson LM, Dodier RH, Branigan DL, Gabo VB, et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metab* 2020;2(7):612-9.

5. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. *Radiol* 2020;296(1):216-24.

6. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA* 2016 Dec 13;316(22):2366-7.

7. von der Emde L, Pfau M, Dysli C, Thiele S, Möller PT, Lindner M, Schmid M, Fleckenstein M, Holz FG, Schmitz-Valckenberg S. Artificial intelligence for morphology-based function prediction in neovascular age-related macular degeneration. *Sci Rep* 2019 Jul 31;9(1):1-2.

8. Devalla SK, Liang Z, Pham TH, Boote C, Strouthidis NG, Thiery AH, Girard MJ. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol* 2020 Mar 1;104(3):301-11.

9. Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *Br Med J* 2021 Oct 20;375.

10. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *British Medical Journal Publishing Group*; 1996.

11. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *Br Med J* 2001;323(7303):42-6.

12. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.

13. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276(8):637-9.

14. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials* 2010;11(1):1-8.

15. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Ann Int Med* 2017;167(1):40-7.

16. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *The Lancet* 2008;371(9609):281-3.

17. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. *Ann Int Med* 2006;144(5):364-7.

18. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013;309(8):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(1):248-54.

20. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br Med J* 2020;368.

21. Liu X, Rivera SC, Faes L, Di Ruffano LF, Yau C, Keane10 PA, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019;25:1467-8.

22. Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. *The Lancet* 2019;394(10205):1225.

23. Rossi MJ, Brand JC. Journal Article Titles Impact Their Citation Rates. *Arthroscopy* 2020;36(7):2025-9. 1.

24. Armağan A. How to write an introduction section of a scientific article? *Turk J Urol* 2013;39(Suppl 1):8-9.

25. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. *Radiol Artif Intell* 2020 Jan 22;2(1):e190015.

26. Ibrahim H, Liu X, Rivera SC, Moher D, Chan AW, Sydes MR, Calvert MJ, Denniston AK. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 2021 Dec;22(1):1-5.

27. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN. Do no harm: a roadmap for responsible machine learning for health care. *Nat Me* 2019 Sep;25(9):1337-40.

28. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ* 2020 Apr 1;98(4):251.

29. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018 Dec 4;320(21):2199-200.

30. Lim CY, In J. Randomization in clinical studies. *Korean J Anesthesiol* 2019 Jun;72(3):221.

31. Tardy MP, Gal J, Chamorey E, Almairac F, Vandenbos F, Bondiau PY, Saada-Bouzid E. Quality of Randomized Controlled Trials Reporting in the Treatment of Adult Patients with High-Grade Gliomas. *Oncologist* 2018 Mar;23(3):337.

32. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Dec;17(1):1-9.

33. 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 2020 Oct 8 (pp. 74-83). Springer, Cham.

34. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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.

35. Zhou Q, Cao YH, Chen ZH. Optimizing the study design of clinical trials to identify the efficacy of artificial intelligence tools in clinical practices. *EClinicalMedicine* 2019 Nov 1;16:10-1.

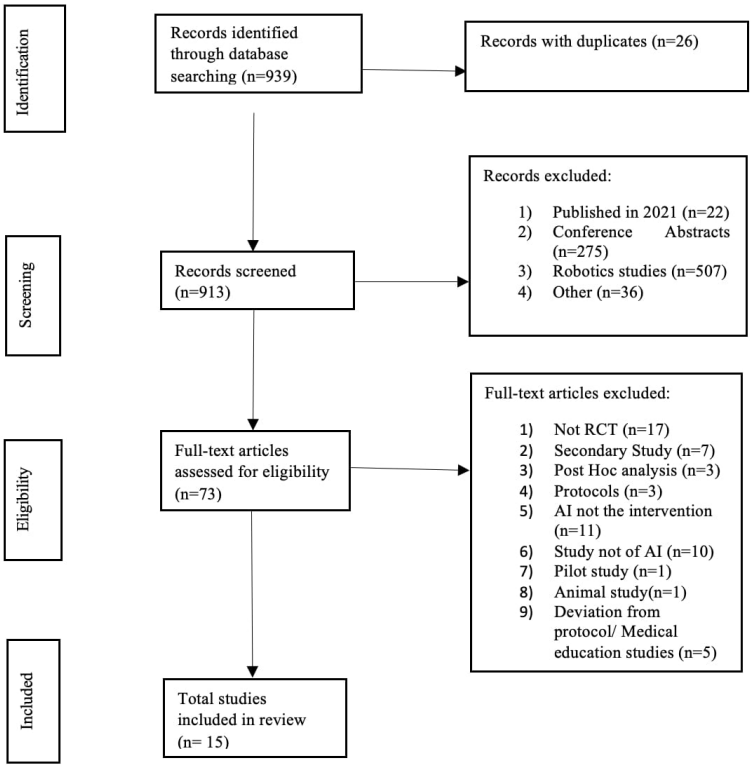


Figure 1: PRISMA flow chart

503x442mm (72 x 72 DPI)

Supplementary table 1: CONSORT-AI checklist**

Section	Item	CONSORT 2010 item	CONSORT-AI item		Addressed on page no.*
Title and abstract					
Title and abstract	1a	Identification as a randomized trial in the title	CONSORT-AI a,b Elaboration	(i) 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 (for specific guidance see CONSORT for abstracts)		(ii) 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	CONSORT-AI a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare	

				professionals, patients, public).	
	2b	Specific objectives or hypotheses			
Methods					
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio			
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons			
Participants	4a	Eligibility criteria for participants	CONSORT-AI a (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.	
			CONSORT-AI a (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.	
	4b	Settings and locations where the data were collected	CONSORT-AI b Extension	Describe how the AI intervention was integrated into the trial setting, including any onsite or on site requirements.	

Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	CONSORT-AI (i) Extension	State which version of the AI algorithm was used.	
			CONSORT-AI (ii) Extension	Describe how the input data were acquired and selected for the AI intervention.	
			CONSORT-AI (iii) Extension	Describe how poor quality or unavailable input data were assessed and handled.	
			CONSORT-AI (iv) Extension.	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.	
			CONSORT-AI (v) Extension	Specify the output of the AI intervention	
			CONSORT-AI (vi) Extension	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome			

		measures, including how and when they were assessed			
	6b	Any changes to trial outcomes after the trial commenced, with reasons			
Sample size	7a	How sample size was determined			
	7b	When applicable, explanation of any interim analyses and stopping guidelines			
Sequence generation	8a	Method used to generate the random allocation sequence			
	8b	Type of randomization; details of any restriction (such as blocking and block size)			
Randomization					
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the			

		sequence until interventions were assigned			
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions			
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how			
	11b	If relevant, description of the similarity of interventions			
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes			
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses			
Results					

Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome			
	13b	For each group, losses and exclusions after randomization, together with reasons			
Recruitment	14a	Dates defining the periods of recruitment and follow-up			
	14b	Why the trial ended or was stopped			
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group			
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups			
Outcomes and estimation	17a	For each primary and secondary outcome,			

		results for each group, and the estimated effect size and its precision (such as % confidence interval)			
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended			
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory			
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	CONSORT-AI 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					
Limitations	20	Trial limitations, addressing sources of potential bias,			

		imprecision, and, if relevant, multiplicity of analyses			
Generalizability	21	Generalizability (external validity, applicability) of the trial findings			
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence			
Other information					
Registration	23	Registration number and name of trial registry			
Protocol	24	Where the full trial protocol can be accessed, if available			
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	CONSORT-AI Extension.	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	

*Indicates page numbers to be completed by authors during protocol development

** 22. Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. The Lancet 2019;394(10205):1225.

Supplementary table 2: Included studies

No.	Authors	Year	Title	Country	Specialty	Disease studied	Sample size	Intervention	Control	Blinding	Primary outcome	Trial registration
1.	Sadasivam, et al.	2016	Impact of a Collective Intelligence Tailored Messaging System on Smoking Cessation: The Perspect Randomized Experiment	United States	Medicine	Smoking addiction	120	AI recommended motivational messages	Standard tailored messages	Single (study staff)	Influence of messages	NR
2.	Morrison, et al.	2017	The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone- Based Stress Management	United States	Psychiatry	Stress	77	Smartphone-based stress management system	Daily/occasional notifications within pre-defined	Not stated	Notification response	ISRCTN67177737

			Intervention: An Exploratory Trial						time frames			
3.	Labovitz, et al.	2017	Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy	United States	Cardiology	Ischemic stroke	28	AI system monitoring	No daily monitoring	Not stated	Adherence to therapy	NCT02599259
4.	Rostill, et al.	2018	Technology integrated health management for dementia	United Kingdom	Psychiatry	Dementia	408	Technology-integrated health management for dementia	No TIHM	Not stated	Alerts	NR
5.	Wang, et al	2018	Real-time automatic detection system increases colonoscopic polyp and adenoma	China	Gastroenterology	Adenoma	1058	AI-aided colonoscopy	Standard colonoscopy	None	Adenoma detection rate	ChiCTR-DDD-17012221

			detection rates: a prospective randomised controlled study									
6.	Lin, et al.	2019	Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial	China	Ophthalmology	Cataract	350	AI-assisted cataract detection	Normal clinic	Double	Diagnostic accuracy for congenital cataracts	NCT03240848

7.	Voss, et al.	2019	Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder A Randomized Clinical Trial	United States	Psychiatry	Autism	474	AI-driven behavioral intervention	Applied behavioral analysis therapy	Single	SRS-II, EGG, VABS-II, NEPSY-II socialization scores	NCT03569176
8.	Wu, et al.	2019	Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots	China	Gastroenterology	Upper GI lesions	324	AI-aided esophagogastroduodenoscopy	Standard esophagogastroduodenoscopy	Single	Blind spot rate	ChiCTR1800014809

			during esophagogastroduodenoscopy									
9.	Wang, et al.	2020	Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study	China	Gastroenterology	Adenoma	1046	AI-aided colonoscopy	Sham	Double	Adenoma detection rate	ChiCTR1800017675
10.	Persell, et al.	2020	Effect of Home Blood Pressure Monitoring via a Smartphone Hypertension Coaching	United States	Medicine	Hypertension	333	AI-driven coaching app	Blood pressure tracking app	None	Systolic blood pressure at 6 months	NCT03288142

			Application or Tracking Application on Adults With Uncontrolled Hypertension A Randomized Clinical Trial									
11.	Wijnberge, et al.	2020	Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension	Netherlands	Cardiology	Intra-operative hypotension	68	AI-driven early warning system for intraoperative hypotension	Standard care	None	Time-weighted average of intraoperative hypotension	NCT03376347

			During Elective Noncardiac Surgery The HYPE Randomized Clinical Trial									
12.	Pavel, et al.	2020	A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial	United Kingdom	Neurology	Neonatal seizures	264	Automated seizure detection algorithm	Conventional EEG	Single	Diagnostic accuracy of healthcare professionals with aid of algorithm	NCT02431780
13.	Nimri, et al.	2020	Insulin dose optimization using an automated artificial intelligence-based decision support	Israel	Endocrinology	Diabetes	108	AI-based decision support system	Physician guided care	Single	Time of glucose level within target range	NCT03003806

			system in youths with type 1 diabetes									
14.	Liu, et al.	2020	The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study	China	Gastroenterology	Adenoma	790	AI-aided colonoscopy	Routine colonoscopy	None	Adenoma detection rate	ChiCTR1800018058
15.	Gong, et al.	2020	Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised	China	Gastroenterology	Adenoma	704	AI-aided colonoscopy	Unassisted colonoscopy	Single	Adenoma detection rate	ChiCTR1900021984

			controlled study									

1. Sadasivam RS, Borglund EM, Adams R, Marlin BM, Houston TK. Impact of a collective intelligence tailored messaging system on smoking cessation: the Perspect randomized experiment. *JMIR* 2016;18(11):e285.
2. Morrison LG, Hargood C, Pejovic V, Geraghty AW, Lloyd S, Goodman N, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. *PloS one* 2017;12(1):e0169162.
3. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke* 2017;48(5):1416-9.
4. Rostill H, Nilforooshan R, Morgan A, Barnaghi P, Ream E, Chrysanthaki T. Technology integrated health management for dementia. *Br J Community Nurs* 2018;23(10):502-8.
5. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68(10):1813-9.
6. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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.

7. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr* 2019;173(5):446-54.

8. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* 2019;68(12):2161-9.

9. Wang P, Liu X, Berzin TM, Brown JRG, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020;5(4):343-51.

10. Persell SD, Peprah YA, Lipiszko D, Lee JY, Li JJ, Ciolino JD, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw Open* 2020;3(3):e200255-e.

11. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323(11):1052-60.

12. Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *The Lancet Child & Adolescent Health* 2020;4(10):740-9.

13. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nature medicine* 2020;26(9):1380-4.

14. Liu P, Wang P, Glissen Brown JR, Berzin TM, Zhou G, Liu W, et al. The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study. *Therap Adv Gastroenterol* 2020;13:1756284820979165.

15. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol* 2020;5(4):352-61.

APPENDIX 1

EMBASE:

1) *deep learning/
2) *artificial intelligence/
3) *machine learning/
4) 1 or 2 or 3

PubMed:

1. Artificial intelligence
2. Machine learning
3. Deep learning
4. 1 OR 2 OR 3

Restricted to:

- Article type: Randomized Control Trial
- Publication Date: 1/01/2015 to 31/12/2020
- Species: Human
- Language: English

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Page 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page 3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 4
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 5
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 5-6
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 5
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 5 + appendix 1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 5-6
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 5-6
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 6
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	NA
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Page 5
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	14
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Figure 1
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	NA
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Table 2
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Page 6
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	NA
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Not reported

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	NA
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 14, figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Not reported
Study characteristics	17	Cite each included study and present its characteristics.	Table 2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Not reported
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Not reported
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Not reported
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Table 3
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Not reported
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Not reported
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Not reported
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Not reported
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Page 27
	23b	Discuss any limitations of the evidence included in the review.	Pahe 27-31
	23c	Discuss any limitations of the review processes used.	Page 31
	23d	Discuss implications of the results for practice, policy, and future research.	Page 31
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Page 14
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Page 14
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Page 1
Competing	26	Declare any competing interests of review authors.	Not

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
interests			reported
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Not reported

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71
For more information, visit: http://www.prisma-statement.org/

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-061519.R1
Article Type:	Original research
Date Submitted by the Author:	14-Jul-2022
Complete List of Authors:	Shahzad, Rida; Shahzad Eye Hospital Ayub, Bushra; Aga Khan University Hospital Siddiqui, M. A. Rehman; Aga Khan University Hospital, Department of Ophthalmology and Visual Sciences
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Qualitative research
Keywords:	STATISTICS & RESEARCH METHODS, Clinical trials < THERAPEUTICS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title: Quality of reporting of randomised controlled trials in artificial intelligence in health care: a systematic review

Authors:

Rida Shahzad[1]

Bushra Ayub[2]

M. A. Rehman Siddiqui[1,3,4]

Authors Affiliations:

1. Shahzad Eye Hospital, Karachi, Pakistan
2. Aga Khan University Hospital, Karachi, Pakistan
3. Department of Ophthalmology and Visual Sciences, Aga Khan University Hospital, Karachi, Pakistan
4. The Eye Centre, South City Hospital, Karachi, Pakistan

Corresponding author:

M. A. Rehman Siddiqui

Department of Ophthalmology and Visual Sciences,

Aga Khan University Hospital,

National Stadium Road, P.O. Box 3500,

Karachi, Pakistan

Email: rehman.siddiqui@gmail.com

Funding: None

Keywords:

CONSORT, CONSORT-AI, artificial intelligence, deep learning, machine learning, randomised controlled trials.

Word count: 2430 words

ABSTRACT

Objectives: The aim of this study was to evaluate the quality of reporting of randomised controlled trials (RCTs) of artificial intelligence (AI) in health care against Consolidated Standards of Reporting Trials – Artificial Intelligence (CONSORT-AI) guidelines.

Design: Systematic review

Data sources: We used PubMed and EMBASE databases to include studies from 2015 to 2021.

Eligibility criteria: We included RCTs which used AI as the intervention. Studies were in the English language. 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. Reporting is variable in existing RCTs, therefore caution must be taken when generalising the findings of these RCTs in real-world settings.

ARTICLE SUMMARY

Strengths and Limitations of the Study:

- This systematic review covers RCTs of AI pertaining to all medical fields.
- The CONSORT-AI guideline was used to assess reporting quality of some RCTs predating its publication to set a baseline for future studies.
- A limitation of this study is the utilization of two databases.

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 CONSORT (Consolidated Standards of Reporting Trials) 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 rigor 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 recently published 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 2020, 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

We performed a systematic review of RCTs of AI for medical conditions published from January 2015 to December 2021. 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 term “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 (Appendix 1). The records were screened by two independent investigators (RS and BA) for potential inclusion. The abstracts of RCTs using artificial intelligence, deep learning, and machine learning were further evaluated for possible inclusion. Protocols, conference abstracts, studies on robotics, and post-hoc analyses of randomised controlled trials 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 randomization. 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.

The CONSORT-AI checklist of 43 items (Supplementary 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 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 protocol of this study was submitted to PROSPERO in February 2021. The only deviation from the submitted protocol was extension of the search until December 2021 to keep this review up-to-date.

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 (Supplementary 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 (n = 1) and dentistry (n = 1). Studies were from different parts of the world, including China (n = 16), United States (n = 14), Japan (n = 3), United Kingdom (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.

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 analyzed	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
Generalizability	21	42	0	0	0
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

Patient and Public Involvement

No patient or public was involved in this study.

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.

Our results reinforce previously published findings. In a systematic review conducted by Liu et al, it was seen that sufficient reporting and external validation was 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 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 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 emphasize 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 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 was 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 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 generalizability of their findings in the clinical setting. The generalizability 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 emphasize 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 to 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. The majority of the included studies were published before the CONSORT-AI checklist was widely available. Hence, some authors may not be aware of ideal information to report about an RCT of AI.

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 must be employed by healthcare service providers and policymakers when using these studies to inform decision-making.

Contributorship

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 a combined discussion with MARS. The manuscript was prepared by RS and BA and edited by MARS. All authors reviewed and approved the final manuscript.

Funding Statement

No funding

Competing of interests

None declared

Ethics approval statement

This study is a systematic review which did not involve human subjects. Therefore, no ethics approval was required.

Data sharing statement

No additional data available.

REFERENCES:

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89-94.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-8.
3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 2018;15(11):e1002686.
4. Tyler NS, Mosquera-Lopez CM, Wilson LM, Dodier RH, Branigan DL, Gabo VB, et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metab* 2020;2(7):612-9.
5. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. *Radiol* 2020;296(1):216-24.
6. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA* 2016 Dec 13;316(22):2366-7.
7. von der Emde L, Pfau M, Dysli C, Thiele S, Möller PT, Lindner M, Schmid M, Fleckenstein M, Holz FG, Schmitz-Valckenberg S. Artificial intelligence for morphology-based function prediction in neovascular age-related macular degeneration. *Sci Rep* 2019 Jul 31;9(1):1-2.

8. Devalla SK, Liang Z, Pham TH, Boote C, Strouthidis NG, Thiery AH, Girard MJ. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol* 2020 Mar 1;104(3):301-11.

9. Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *Br Med J* 2021 Oct 20;375.

10. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *British Medical Journal Publishing Group*; 1996.

11. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *Br Med J* 2001;323(7303):42-6.

12. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.

13. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276(8):637-9.

14. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials* 2010;11(1):1-8.

15. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Ann Int Med* 2017;167(1):40-7.

16. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *The Lancet* 2008;371(9609):281-3.

17. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. *Ann Int Med* 2006;144(5):364-7.

18. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013;309(8):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(1):248-54.

20. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br Med J* 2020;368.

21. Liu X, Rivera SC, Faes L, Di Ruffano LF, Yau C, Keane10 PA, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019;25:1467-8.

22. Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. *The Lancet* 2019;394(10205):1225.

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(6):e271-e97

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(12):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(3):e034568.

26. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. *Radiol Artif Intell* 2020 Jan 22;2(1):e190015.

27. Ibrahim H, Liu X, Rivera SC, Moher D, Chan AW, Sydes MR, Calvert MJ, Denniston AK. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 2021 Dec;22(1):1-5.

28. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN. Do no harm: a roadmap for responsible machine learning for health care. *Nat Me* 2019 Sep;25(9):1337-40.

29. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ* 2020 Apr 1;98(4):251.

30. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018 Dec 4;320(21):2199-200.

31. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Dec;17(1):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 2020 Oct 8 (pp. 74-83). Springer, Cham.

33. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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.

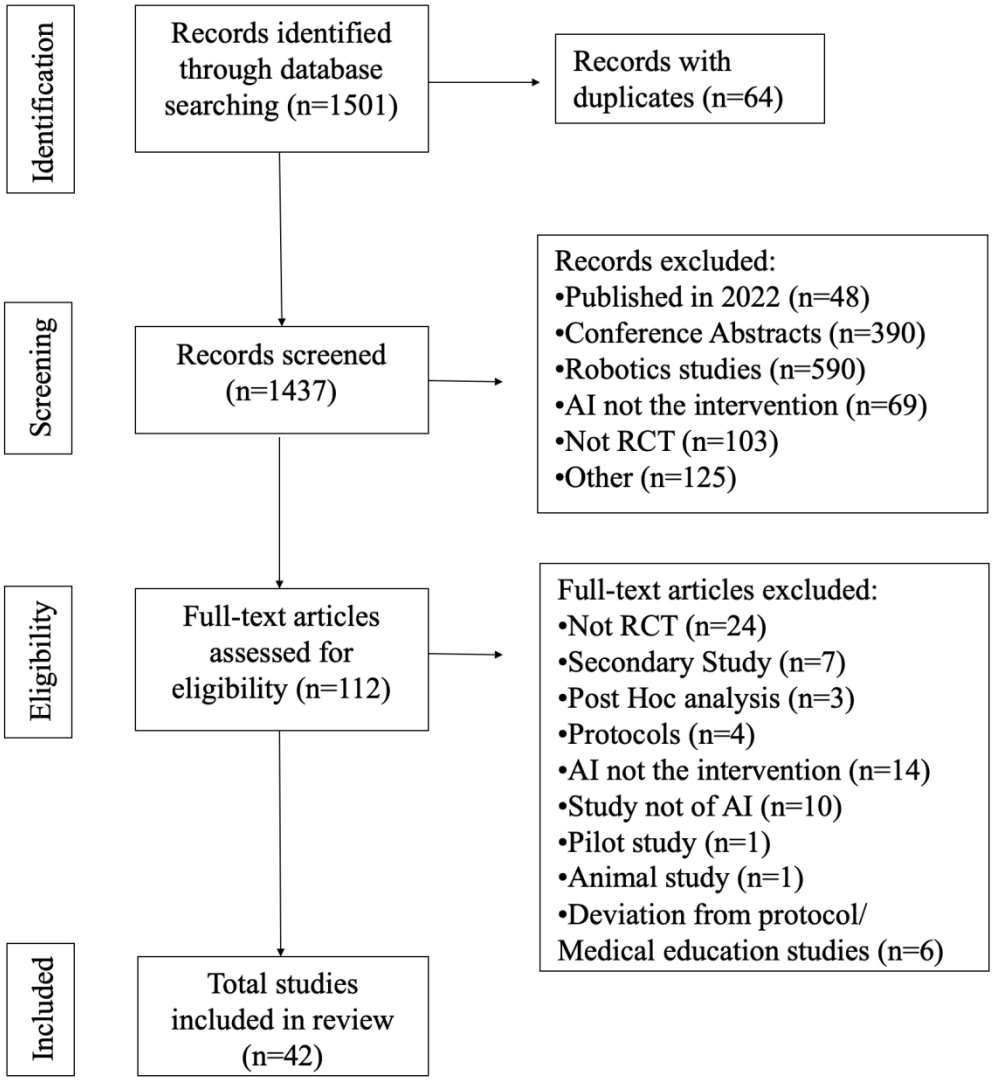
34. Zhou Q, Cao YH, Chen ZH. Optimizing the study design of clinical trials to identify the efficacy of artificial intelligence tools in clinical practices. *EClinicalMedicine* 2019 Nov 1;16:10-1.

FIGURE LEGENDS:

Figure 1: PRISMA flowchart

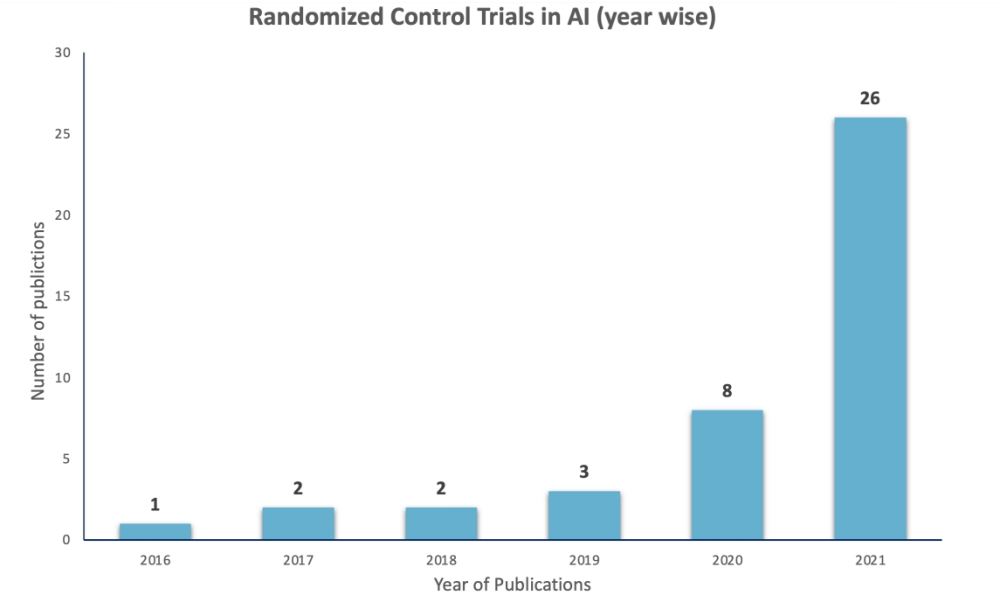
Figure 2: Year-wise distribution of RCTs in AI

Figure 3: Percentage of AI RCTs in different countries and specialties



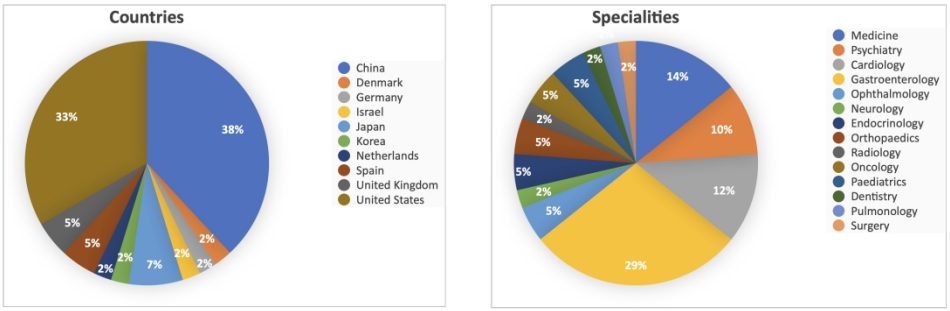
PRISMA flow chart

293x349mm (144 x 144 DPI)



Year-wise distribution of RCTs in AI

403x244mm (144 x 144 DPI)



Percentage of AI RCTs in different countries	Percentage of AI RCTs in different specialities

Percentage of AI RCTs in different countries and specialties

608x295mm (144 x 144 DPI)

Supplementary table 1: CONSORT-AI checklist**

Section	Item	CONSORT 2010 item	CONSORT-AI item		Addressed on page no.*
Title and abstract					
Title and abstract	1a	Identification as a randomized trial in the title	CONSORT-AI a,b Elaboration	(i) 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 (for specific guidance see CONSORT for abstracts)		(ii) 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	CONSORT-AI a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare	

				professionals, patients, public).	
	2b	Specific objectives or hypotheses			
Methods					
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio			
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons			
Participants	4a	Eligibility criteria for participants	CONSORT-AI a (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.	
			CONSORT-AI a (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.	
	4b	Settings and locations where the data were collected	CONSORT-AI b Extension	Describe how the AI intervention was integrated into the trial setting, including any onsite or on site requirements.	

Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	CONSORT-AI (i) Extension	State which version of the AI algorithm was used.	
			CONSORT-AI (ii) Extension	Describe how the input data were acquired and selected for the AI intervention.	
			CONSORT-AI (iii) Extension	Describe how poor quality or unavailable input data were assessed and handled.	
			CONSORT-AI (iv) Extension.	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.	
			CONSORT-AI (v) Extension	Specify the output of the AI intervention	
			CONSORT-AI (vi) Extension	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome			

		measures, including how and when they were assessed			
	6b	Any changes to trial outcomes after the trial commenced, with reasons			
Sample size	7a	How sample size was determined			
	7b	When applicable, explanation of any interim analyses and stopping guidelines			
Sequence generation	8a	Method used to generate the random allocation sequence			
	8b	Type of randomization; details of any restriction (such as blocking and block size)			
Randomization					
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the			

		sequence until interventions were assigned			
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions			
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how			
	11b	If relevant, description of the similarity of interventions			
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes			
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses			
Results					

Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome			
	13b	For each group, losses and exclusions after randomization, together with reasons			
Recruitment	14a	Dates defining the periods of recruitment and follow-up			
	14b	Why the trial ended or was stopped			
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group			
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups			
Outcomes and estimation	17a	For each primary and secondary outcome,			

		results for each group, and the estimated effect size and its precision (such as % confidence interval)			
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended			
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory			
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	CONSORT-AI 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					
Limitations	20	Trial limitations, addressing sources of potential bias,			

		imprecision, and, if relevant, multiplicity of analyses			
Generalizability	21	Generalizability (external validity, applicability) of the trial findings			
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence			
Other information					
Registration	23	Registration number and name of trial registry			
Protocol	24	Where the full trial protocol can be accessed, if available			
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	CONSORT-AI Extension.	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	

*Indicates page numbers to be completed by authors during protocol development

** 22. Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. The Lancet 2019;394(10205):1225.

Supplementary table 2: Included studies

No.	Authors	Year	Title	Country	Specialty	Disease studied	Sample size	Intervention	Control	Blinding	Primary outcome	Trial registration
1.	Sadasivam, et al.	2016	Impact of a Collective Intelligence Tailored Messaging System on Smoking Cessation: The Perspect Randomized Experiment	United States	Medicine	Smoking addiction	120	AI recommended motivational messages	Standard tailored messages	Single (study staff)	Influence of messages	NR
2.	Morrison, et al.	2017	The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone- Based Stress Management	United States	Psychiatry	Stress	77	Smartphone-based stress management system	Daily/occasional notifications within pre-defined	Not stated	Notification response	ISRCTN67177737

			Intervention: An Exploratory Trial						time frames			
3.	Labovitz, et al.	2017	Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy	United States	Cardiology	Ischemic stroke	28	AI system monitoring	No daily monitoring	Not stated	Adherence to therapy	NCT02599259
4.	Rostill, et al.	2018	Technology integrated health management for dementia	United Kingdom	Psychiatry	Dementia	408	Technology-integrated health management for dementia	No TIHM	Not stated	Alerts	NR
5.	Wang, et al	2018	Real-time automatic detection system increases colonoscopic polyp and adenoma	China	Gastroenterology	Adenoma	1058	AI-aided colonoscopy	Standard colonoscopy	None	Adenoma detection rate	ChiCTR-DDD-17012221

			detection rates: a prospective randomised controlled study									
6.	Lin, et al.	2019	Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial	China	Ophthalmology	Cataract	350	AI-assisted cataract detection	Normal clinic	Double	Diagnostic accuracy for congenital cataracts	NCT03240848

7.	Voss, et al.	2019	Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder A Randomized Clinical Trial	United States	Psychiatry	Autism	474	AI-driven behavioral intervention	Applied behavioral analysis therapy	Single	SRS-II, EGG, VABS-II, NEPSY-II socialization scores	NCT03569176
8.	Wu, et al.	2019	Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots	China	Gastroenterology	Upper GI lesions	324	AI-aided esophagogastroduodenoscopy	Standard esophagogastroduodenoscopy	Single	Blind spot rate	ChiCTR1800014809

			during esophagogastroduodenoscopy									
9.	Wang, et al.	2020	Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study	China	Gastroenterology	Adenoma	1046	AI-aided colonoscopy	Sham	Double	Adenoma detection rate	ChiCTR1800017675
10.	Persell, et al.	2020	Effect of Home Blood Pressure Monitoring via a Smartphone Hypertension Coaching	United States	Medicine	Hypertension	333	AI-driven coaching app	Blood pressure tracking app	None	Systolic blood pressure at 6 months	NCT03288142

			Application or Tracking Application on Adults With Uncontrolled Hypertension A Randomized Clinical Trial									
11.	Wijnberge, et al.	2020	Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension	Netherlands	Cardiology	Intra-operative hypotension	68	AI-driven early warning system for intraoperative hypotension	Standard care	None	Time-weighted average of intraoperative hypotension	NCT03376347

			During Elective Noncardiac Surgery The HYPE Randomized Clinical Trial									
12.	Pavel, et al.	2020	A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial	United Kingdom	Neurology	Neonatal seizures	264	Automated seizure detection algorithm	Conventional EEG	Single	Diagnostic accuracy of healthcare professionals with aid of algorithm	NCT02431780
13.	Nimri, et al.	2020	Insulin dose optimization using an automated artificial intelligence-based decision support	Israel	Endocrinology	Diabetes	108	AI-based decision support system	Physician guided care	Single	Time of glucose level within target range	NCT03003806

			system in youths with type 1 diabetes									
14.	Liu, et al.	2020	The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study	China	Gastroenterology	Adenoma	790	AI-aided colonoscopy	Routine colonoscopy	None	Adenoma detection rate	ChiCTR1800018058
15.	Gong, et al.	2020	Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised	China	Gastroenterology	Adenoma	704	AI-aided colonoscopy	Unassisted colonoscopy	Single	Adenoma detection rate	ChiCTR1900021984

			controlled study									
16.	Luo, et al.	2020	Artificial Intelligence-Assisted Colonoscopy for Detection of Colon Polyps: a Prospective, Randomized Cohort Study	China	Gastroenterology	Polyps	150	AI-aided colonoscopy	Unaided colonoscopy	None	Polyp detection rate	NCT04712626 5
17.	Anan, at al.	2021	Effects of an Artificial Intelligence-Assisted Health Program on Workers With Neck/Shoulder Pain/Stiffness and Low Back Pain:	Japan	Orthopaedics	Neck/shoulder and back pain	94	AI-assisted health program	Usual care routine	None	Pain level	(UMIN-CTR) 000033894

			Randomized Controlled Trial									
18.	Blomberg, et al.	2021	Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial	Denmark	Cardiology	Cardiac arrest	654	AI-led alerts	Normal protocol	Double	Recognition of cardiac arrest	NCT04219306
19.	Chen, et al.	2021	The Role of Deep Learning-Based Echocardiography in the Diagnosis and Evaluation of the Effects of Routine Anti-Heart-	China	Cardiology	Heart failure	80	AI-based echocardiography	Routine echocardiography	None	Mortality and rehospitalization rate	NR

			Failure Western Medicines in Elderly Patients with Acute Left Heart Failure									
20.	Eng, et al.	2021	Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment	United States	Radiology	Skeletal age	1903	AI diagnostic aid	Without aid	None	Mean absolute difference between the skeletal age	NCT03530098
21.	Harada, et al.	2021	Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy	Japan	Medicine	Various medical conditions	22	AI-assisted differential diagnosis	Without AI assistance	Single	Diagnostic accuracy	UMIN000042881

			of physicians: An open-label randomized controlled study									
22.	Hassoon, et al.	2021	Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors	United States	Oncology	Different cancer types	42	AI coaching	Written information	Single	Change in steps per day	NCT03212079
23.	Jayakumar, et al.	2021	Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on	United States	Orthopaedics	Osteoarthritis	129	AI-enabled patient decision aid	Educational material	None	Knee OA Decision Quality	NCT03956004

			Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults with Knee Osteoarthritis: A Randomized Clinical Trial									
24.	Kamba, et al.	2021	Reducing adenoma miss rate of colonoscopy assisted by artificial intelligence: a multicenter randomized controlled trial	Japan	Gastroenterology	Adenoma	358	AI-aided colonoscopy	Unaided colonoscopy	None	Adenoma miss rate	jRCTs032190061

25.	Luna, et al.	2021	Artificial intelligence application versus physical therapist for squat evaluation: a randomized controlled trial	United States	Medicine	Physical therapy	30	AI-assisted exercise application	Unassisted exercise	Single	Successful squats	NCT04624594
26.	Medina, et al.	2021	Electrophysiological brain changes associated with cognitive improvement in a pediatric attention deficit hyperactivity disorder digital artificial intelligence-driven intervention:	Spain	Psychiatry	Attention deficit hyperactivity disorder	29	AI-driven cognitive stimulation program	Commerical video games	Single	Conners CPT (CPT-III) score	ISRCTN71041318

			Randomized controlled trial									
27.	Mertens, et al.	2021	Artificial intelligence for caries detection: Randomized trial	Germany	Dentistry	Caries	22	AI-based diagnostic support	Unaided diagnosis	None	Accuracy metrics	DRKS00022357
28.	Prochaska, et al.	2021	A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic	United States	Medicine	Substance-related disorders	180	AI relational conversational agent	No intervention during the study	None	Past-month substance use occasions	NCT04096001
29.	Rafferty, et al.	2021	A novel mobile app (heali) for disease treatment in participants with irritable bowel	United States	Gastroenterology	Irritable bowel syndrome	58	AI dietry mobile app	Educational material	None	Quality of life score	NCT04256551

			syndrome: Randomized controlled pilot trial									
30.	Seok, et al.	2021	A personalized 3d-printed model for obtaining informed consent process for thyroid surgery: A randomized clinical study using a deep learning approach with mesh-type 3d modeling	Korea	Endocrinology	Thyroid lesions	53	AI 3D-printed thyroid model	Without model	None	Patient general knowledge and satisfaction	KCT0005069
31.	Seol, et al.	2021	Artificial intelligence-assisted clinical decision support for childhood asthma management: A	United States	Paediatrics	Childhood asthma	184	AI-assisted decision support	Usual asthma care	Single	Asthma exacerbation occurrence within one year	NCT02865967

			randomized clinical trial									
32.	Strombla d, et al.	2021	Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources: A Randomized Clinical Trial	United States	Surgery	Gynaecol ogical and colorecta l surgery	683	AI-assisted surgical predictions	Standard estimatio n process	None	Accurate surgery duration prediction	NCT0347137 7
33.	Turino, et al.	2021	Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system	Spain	Pulmonolo gy	Obstructi ve sleep apnea	60	Intelligent monitoring system	Standard manage ment	None	Complianc e to CPAP	NCT0311695 8

			based on machine learning aiming to improve continuous positive airway pressure treatment compliance: Randomized controlled trial									
34.	Wang, et al.	2021	Utilization of Ultrasonic Image Characteristics Combined with Endoscopic Detection on the Basis of Artificial Intelligence Algorithm in Diagnosis of Early Upper	China	Gastroenterology	Early gastric cancer	80	Endoscopy with AI-based ultrasound imaging	Endoscopy alone	None	Detection rate of upper gastric cancer	NR

			Gastrointestinal Cancer									
35.	Wu, et al.	2021	Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial	China	Gastroenterology	Early gastric cancer	1050	AI-aided endoscopy	Endoscopy alone	Single	Number of blind spots during endoscopy	ChiCTR1800018403
36.	Wu, et al.	2021	Effect of a deep learning-based system on the miss rate of gastric neoplasms during	China	Gastroenterology	Gastric neoplasm	1886	AI-assisted endoscopy	Unaided endoscopy	Single	Gastric neoplasm miss rate	ChiCTR2000034453

			upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial									
37.	Xu, et al.	2021	The Clinical Value of Explainable Deep Learning for Diagnosing Fungal Keratitis Using in vivo Confocal Microscopy Images	China	Ophthalmology	Fungal keratitis	1089	AI-assisted image reading	Unassisted image reading	None	Accuracy	NR
38.	Xu, et al.	2021	Artificial intelligence-assisted colonoscopy: A prospective, multicenter, randomized	China	Gastroenterology	Polyp	2352	AI-assisted colonoscopy	Conventional colonoscopy	None	Polyp detection rate	ChiCTR1800015607

			controlled trial of polyp detection									
39.	Yao, et al.	2021	Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial	United States	Cardiology	Low ejection fraction	358	AI-enabled electrocardiograms	Usual care	None	New diagnosis of low ejection fraction	NCT04000087
40.	Zeng, et al.	2021	Long-Term Assessment of Rehabilitation Treatment of Sports through Artificial Intelligence Research	China	Medicine	Sports health management	150	AI-based personalized sports management service system	General management	None	Blood glucose, blood pressure, lipids	NR

41.	Zhang, et al.	2021	Artificial Intelligence Algorithm-Based Ultrasound Image Segmentation Technology in the Diagnosis of Breast Cancer Axillary Lymph Node Metastasis	China	Oncology	Breast cancer	90	AI-based ultra sound image segmentation	Routine ultrasound	Double	Accuracy metrics	NR
42.	Zhang, et al.	2021	Value of Rehabilitation Training for Children with Cerebral Palsy Diagnosed and Analyzed by Computed Tomography	China	Paediatrics	Cerebral palsy	73	AI-assisted analysis of brain images	Original images	None	Cerebral artery blood flow velocity (VP) and Vasscular pulse index (PI)	NR

			Imaging Information Features under Deep Learning									

1. Sadasivam RS, Borglund EM, Adams R, Marlin BM, Houston TK. Impact of a collective intelligence tailored messaging system on smoking cessation: the Perspect randomized experiment. *JMIR* 2016;18(11):e285.
2. Morrison LG, Hargood C, Pejovic V, Geraghty AW, Lloyd S, Goodman N, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. *PloS one* 2017;12(1):e0169162.
3. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke* 2017;48(5):1416-9.
4. Rostill H, Nilforooshan R, Morgan A, Barnaghi P, Ream E, Chrysanthaki T. Technology integrated health management for dementia. *Br J Community Nurs* 2018;23(10):502-8.
5. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68(10):1813-9.

6. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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.

7. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr* 2019;173(5):446-54.

8. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* 2019;68(12):2161-9.

9. Wang P, Liu X, Berzin TM, Brown JRG, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020;5(4):343-51.

10. Persell SD, Peprah YA, Lipiszko D, Lee JY, Li JJ, Ciolino JD, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw Open* 2020;3(3):e200255-e.

11. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323(11):1052-60.

12. Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *The Lancet Child & Adolescent Health* 2020;4(10):740-9.

13. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nature medicine* 2020;26(9):1380-4.

14. Liu P, Wang P, Glissen Brown JR, Berzin TM, Zhou G, Liu W, et al. The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study. *Therap Adv Gastroenterol* 2020;13:1756284820979165.

15. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol* 2020;5(4):352-61.

16. Luo Y, Zhang Y, Liu M, Lai Y, Liu P, Wang Z, Xing T, Huang Y, Li Y, Li A, Wang Y. Artificial intelligence-assisted colonoscopy for detection of colon polyps: a prospective, randomized cohort study. *Journal of Gastrointestinal Surgery* 2021;25(8):2011-8.

17. Anan T, Kajiki S, Oka H, et al. Effects of an Artificial Intelligence-Assisted Health Program on Workers With Neck/Shoulder Pain/Stiffness and Low Back Pain: Randomized Controlled Trial. *JMIR Mhealth Uhealth* 2021;9(9):e27535.

18. 1. Blomberg SN, Christensen HC, Lippert F, et al. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw Open* 2021;4(1):e2032320.

19. Chen J, Gao Y. The Role of Deep Learning-Based Echocardiography in the Diagnosis and Evaluation of the Effects of Routine Anti-Heart-Failure Western Medicines in Elderly Patients with Acute Left Heart Failure. *Journal of Healthcare Engineering* 2021;2021

20. Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: A prospective multicenter randomized controlled trial. *Radiology* 2021;301(3):692-99.

21. Harada Y, Katsukura S, Kawamura R, et al. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: An open-label randomized controlled study. *International Journal of Environmental Research and Public Health* 2021;18(4):1-10.

22. Hassoon A, Baig Y, Naiman DQ, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. *npj Digital Medicine* 2021;4(1)

23. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults with Knee Osteoarthritis: A Randomized Clinical Trial. *JAMA Network Open* 2021;4(2)

24. Kamba S, Tamai N, Saitoh I, et al. Reducing adenoma miss rate of colonoscopy assisted by artificial intelligence: a multicenter randomized controlled trial. *Journal of Gastroenterology* 2021;56(8):746-57.

25. Luna A, Casertano L, Timmerberg J, et al. Artificial intelligence application versus physical therapist for squat evaluation: a randomized controlled trial. *Scientific reports* 2021;11(1):18109.

26. Medina R, Bouhaben J, de Ramon I, et al. Electrophysiological brain changes associated with cognitive improvement in a pediatric attention deficit hyperactivity disorder digital artificial intelligence-driven intervention: Randomized controlled trial. *Journal of Medical Internet Research* 2021;23(11)

27. Mertens S, Krois J, Cantu AG, et al. Artificial intelligence for caries detection: Randomized trial. *J Dent* 2021;115:103849.

28. Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug and Alcohol Dependence* 2021;227

29. Rafferty AJ, Hall R, Johnston CS. A novel mobile app (heali) for disease treatment in participants with irritable bowel syndrome: Randomized controlled pilot trial. *Journal of Medical Internet Research* 2021;23(3)

30. Seok J, Yoon S, Ryu CH, et al. A personalized 3d-printed model for obtaining informed consent process for thyroid surgery: A randomized clinical study using a deep learning approach with mesh-type 3d modeling. *Journal of Personalized Medicine* 2021;11(6)

31. Seol HY, Shrestha P, Muth JF, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: A randomized clinical trial. *PLoS ONE* 2021;16(8 August)

32. Stromblad CT, Baxter-King RG, Meisami A, et al. Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources: A Randomized Clinical Trial. *JAMA Surgery* 2021;156(4):315-21.

33. Turino C, Benitez ID, Rafael-Palou X, et al. Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: Randomized controlled trial. *Journal of Medical Internet Research* 2021;23(10)

34. Wang L, Song H, Wang M, et al. Utilization of Ultrasonic Image Characteristics Combined with Endoscopic Detection on the Basis of Artificial Intelligence Algorithm in Diagnosis of Early Upper Gastrointestinal Cancer. *J Healthc Eng* 2021;2021:2773022.

35. Wu L, He X, Liu M, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. *Endoscopy* 2021;53(12):1199-207

36. Wu L, Shang R, Sharma P, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. *The Lancet Gastroenterology and Hepatology* 2021;6(9):700-08.

37. Xu F, Jiang L, He W, et al. The Clinical Value of Explainable Deep Learning for Diagnosing Fungal Keratitis Using in vivo Confocal Microscopy Images. *Frontiers in Medicine* 2021;8

38. Xu L, He X, Zhou J, et al. Artificial intelligence-assisted colonoscopy: A prospective, multicenter, randomized controlled trial of polyp detection. *Cancer Medicine* 2021;10(20):7184-93.

39. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nature Medicine* 2021;27(5):815-19.

40. Zeng C, Huang Y, Yu L, et al. Long-Term Assessment of Rehabilitation Treatment of Sports through Artificial Intelligence Research. *Comput Math Methods Med* 2021;2021:4980718.

41. Zhang L, Jia Z, Leng X, et al. Artificial Intelligence Algorithm-Based Ultrasound Image Segmentation Technology in the Diagnosis of Breast Cancer Axillary Lymph Node Metastasis. *J Healthc Eng* 2021;2021:8830260.

42. Zhang X, Wang Z, Liu J, et al. Value of Rehabilitation Training for Children with Cerebral Palsy Diagnosed and Analyzed by Computed Tomography Imaging Information Features under Deep Learning. *J Healthc Eng* 2021;2021:6472440.

Search Strategy:

FOR EMBASE:

1) *deep learning/
2) *artificial intelligence/
3) *machine learning/
4) 1 or 2 or 3

For PubMed:

1. Artificial intelligence
2. Machine learning
3. Deep learning
4. 1 OR 2 OR 3

Restricted to:

- Article type: Randomized Control Trial
- Publication Date: 1/01/2015 to 31/12/2021
- Species: Human
- Language: English

Reporting checklist for systematic review (with or without a meta-analysis).

Based on the PRISMA guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the PRISMAreporting guidelines, and cite them as:

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews

		Reporting Item	Page Number
Title			
Title	#1	Identify the report as a systematic review	1
Abstract			
Abstract	#2	Report an abstract addressing each item in the PRISMA 2020 for Abstracts checklist	3
Introduction			
Background/rationale	#3	Describe the rationale for the review in the context of existing knowledge	4
Objectives	#4	Provide an explicit statement of the objective(s) or question(s) the review addresses	5
Methods			

Eligibility criteria	#5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses	6
Information sources	#6	Specify all databases, registers, websites, organisations, reference lists, and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted	6
Search strategy	#7	Present the full search strategies for all databases, registers, and websites, including any filters and limits used	6
Selection process	#8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and, if applicable, details of automation tools used in the process	6
Data collection process	#9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and, if applicable, details of automation tools used in the process	7
Data items	#10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (for example, for all measures, time points, analyses), and, if not, the methods used to decide which results to collect	7
Study risk of bias assessment	#11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and, if applicable, details of automation tools used in the process	n/a
Effect measures	#12	Specify for each outcome the effect measure(s) (such as risk ratio, mean difference) used in the synthesis or presentation of results	7
Synthesis methods	#13a	Describe the processes used to decide which studies were eligible for each synthesis (such as tabulating the study	n/a

		intervention characteristics and comparing against the planned groups for each synthesis (item #5))	
Synthesis methods	#13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics or data conversions	n/a
Synthesis methods	#13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses	n/a
Synthesis methods	#13d	Describe any methods used to synthesise results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used	n/a
Synthesis methods	#13e	Describe any methods used to explore possible causes of heterogeneity among study results (such as subgroup analysis, meta-regression)	n/a
Synthesis methods	#13f	Describe any sensitivity analyses conducted to assess robustness of the synthesised results	n/a
Reporting bias assessment	#14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases)	n/a
Certainty assessment	#15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome	n/a
Data items	#10b	List and define all other variables for which data were sought (such as participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information	n/a
Results			
Study selection	#16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram (http://www.prisma-statement.org/PRISMAStatement/FlowDiagram)	7
Study selection	#16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded	7

Study characteristics	#17	Cite each included study and present its characteristics	supplementary table 2
Risk of bias in studies	#18	Present assessments of risk of bias for each included study	n/a
Results of individual studies	#19	For all outcomes, present for each study (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (such as confidence/credible interval), ideally using structured tables or plots	8
Results of syntheses	#20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies	n/a
Results of syntheses	#20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (such as confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect	n/a
Results of syntheses	#20c	Present results of all investigations of possible causes of heterogeneity among study results	n/a
Results of syntheses	#20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesised results	n/a
Risk of reporting biases in syntheses	#21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed	n/a
Certainty of evidence	#22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed	n/a
Discussion			
Results in context	#23a	Provide a general interpretation of the results in the context of other evidence	11
Limitations of included studies	#23b	Discuss any limitations of the evidence included in the review	14
Limitations of the review methods	#23c	Discuss any limitations of the review processes used	14
Implications	#23d	Discuss implications of the results for practice, policy, and future research	14

Other information

Registration and protocol	#24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered	7
Registration and protocol	#24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared	7
Registration and protocol	#24c	Describe and explain any amendments to information provided at registration or in the protocol	7
Support	#25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review	15
Competing interests	#26	Declare any competing interests of review authors	15
Availability of data, code, and other materials	#27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review	supplementary table 2

Notes:

- 17: supplementary table 2
- 27: supplementary table 2 The PRISMA checklist is distributed under the terms of the Creative Commons Attribution License CC-BY. This checklist was completed on 12. July 2022 using https://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-061519.R2
Article Type:	Original research
Date Submitted by the Author:	16-Aug-2022
Complete List of Authors:	Shahzad, Rida; Shahzad Eye Hospital Ayub, Bushra; Aga Khan University Hospital Siddiqui, M. A. Rehman; Aga Khan University Hospital, Department of Ophthalmology and Visual Sciences
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Qualitative research
Keywords:	STATISTICS & RESEARCH METHODS, Clinical trials < THERAPEUTICS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Authors:

Rida Shahzad[1]

Bushra Ayub[2]

M. A. Rehman Siddiqui[1,3,4]

Affiliations:

1. Shahzad Eye Hospital, Karachi, Pakistan
2. Aga Khan University Hospital, Karachi, Pakistan
3. Department of Ophthalmology and Visual Sciences, Aga Khan University Hospital, Karachi, Pakistan
4. The Eye Centre, South City Hospital, Karachi, Pakistan

Correspondence to:

M. A. Rehman Siddiqui

Department of Ophthalmology and Visual Sciences,

Aga Khan University Hospital,

National Stadium Road, P.O. Box 3500,

Karachi, Pakistan

Email: rehman.siddiqui@gmail.com

Keywords: CONSORT, CONSORT-AI, artificial intelligence, deep learning, machine learning, randomised controlled trials

Word count: 2518 words

ABSTRACT

Objectives: The aim of this study was to evaluate the quality of reporting of randomised controlled trials (RCTs) of artificial intelligence (AI) in health care against Consolidated Standards of Reporting Trials – Artificial Intelligence (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.

ARTICLE SUMMARY

Strengths and limitations of this study

- This systematic review assesses the reporting of randomised trials of artificial intelligence interventions across medical fields from 2015 to 2021 against 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 CONSORT (Consolidated Standards of Reporting Trials) 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 rigor 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 September 9, 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 2015-2020; the literature search was later updated to include publications until December 2021. RCTs of AI in healthcare is 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 term "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 (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 artificial intelligence, deep learning, and machine learning were further evaluated for possible inclusion. Protocols, conference abstracts, studies on robotics, and post-hoc analyses of randomised controlled trials 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 (Supplementary 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 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 (Supplementary 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 (n = 1) and dentistry (n = 1). Studies were from different parts of the world, including China (n = 16), United States (n = 14), Japan (n = 3), United Kingdom (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.

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	2a, 2a(i)	41	1	0	0

objectives	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
Generalizability	21	42	0	0	0
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

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.

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 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 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 emphasize 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 was 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 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 generalizability of their findings in the clinical setting. The generalizability 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 emphasize 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 to 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.

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 a combined discussion with MARS. The manuscript was prepared by RS and BA and edited by MARS. All authors reviewed and approved the final manuscript.

Funding

None.

Competing interests

None declared.

Ethics approval statement

This study is a systematic review which did not involve human subjects. Therefore, no ethics approval was required.

Data availability statement

No additional data available.

REFERENCES

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89-94.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-8.
3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 2018;15(11):e1002686.

4. Tyler NS, Mosquera-Lopez CM, Wilson LM, Dodier RH, Branigan DL, Gabo VB, et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat Metab* 2020;2(7):612-9.

5. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. *Radiol* 2020;296(1):216-24.

6. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA* 2016 Dec 13;316(22):2366-7.

7. von der Emde L, Pfau M, Dysli C, Thiele S, Möller PT, Lindner M, Schmid M, Fleckenstein M, Holz FG, Schmitz-Valckenberg S. Artificial intelligence for morphology-based function prediction in neovascular age-related macular degeneration. *Sci Rep* 2019 Jul 31;9(1):1-2.

8. Devalla SK, Liang Z, Pham TH, Boote C, Strouthidis NG, Thiery AH, Girard MJ. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol* 2020 Mar 1;104(3):301-11.

9. Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *Br Med J* 2021 Oct 20;375.

10. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *British Medical Journal Publishing Group*; 1996.

11. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *Br Med J* 2001;323(7303):42-6.

12. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.

13. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276(8):637-9.

14. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials* 2010;11(1):1-8.

15. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Ann Int Med* 2017;167(1):40-7.

16. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *The Lancet* 2008;371(9609):281-3.

17. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. *Ann Int Med* 2006;144(5):364-7.

18. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013;309(8):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(1):248-54.

20. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br Med J* 2020;368.

21. Liu X, Rivera SC, Faes L, Di Ruffano LF, Yau C, Keane10 PA, et al. 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(6):e271-e97

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(12):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(3):e034568.

26. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. *Radiol Artif Intell* 2020 Jan 22;2(1):e190015.

27. Ibrahim H, Liu X, Rivera SC, Moher D, Chan AW, Sydes MR, Calvert MJ, Denniston AK. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 2021 Dec;22(1):1-5.

28. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN. Do no harm: a roadmap for responsible machine learning for health care. *Nat Me* 2019 Sep;25(9):1337-40.

29. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ* 2020 Apr 1;98(4):251.

30. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018 Dec 4;320(21):2199-200.

31. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Dec;17(1):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 2020 Oct 8 (pp. 74-83). Springer, Cham.

33. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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 YH, Chen ZH. Optimizing the study design of clinical trials to identify the efficacy of artificial intelligence tools in clinical practices. *EClinicalMedicine* 2019 Nov 1;16:10-1.

FIGURE TITLES:

Figure 1: PRISMA flowchart

Figure 2: Year-wise distribution of RCTs in AI

Figure 3: Percentage of AI RCTs in different countries and specialties

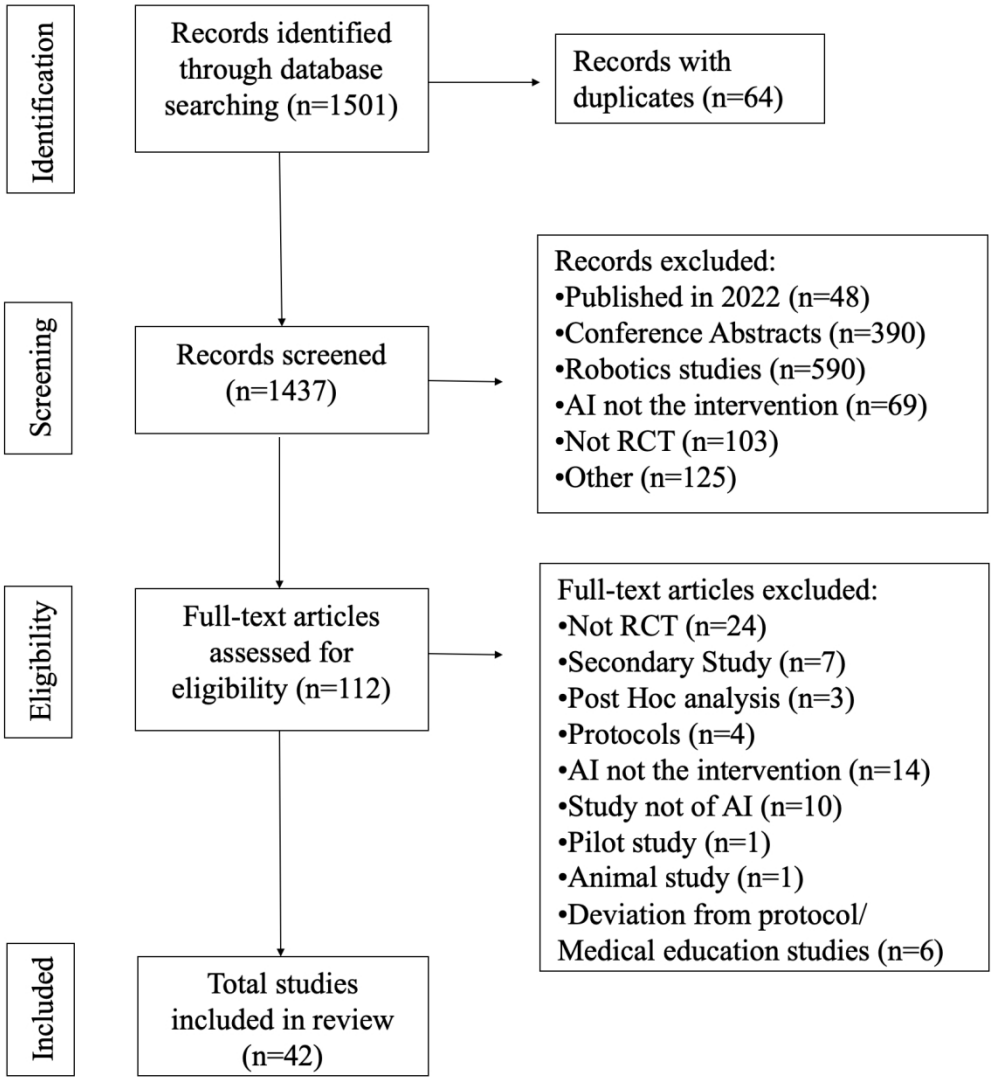


Figure 1: PRISMA flowchart

293x349mm (144 x 144 DPI)

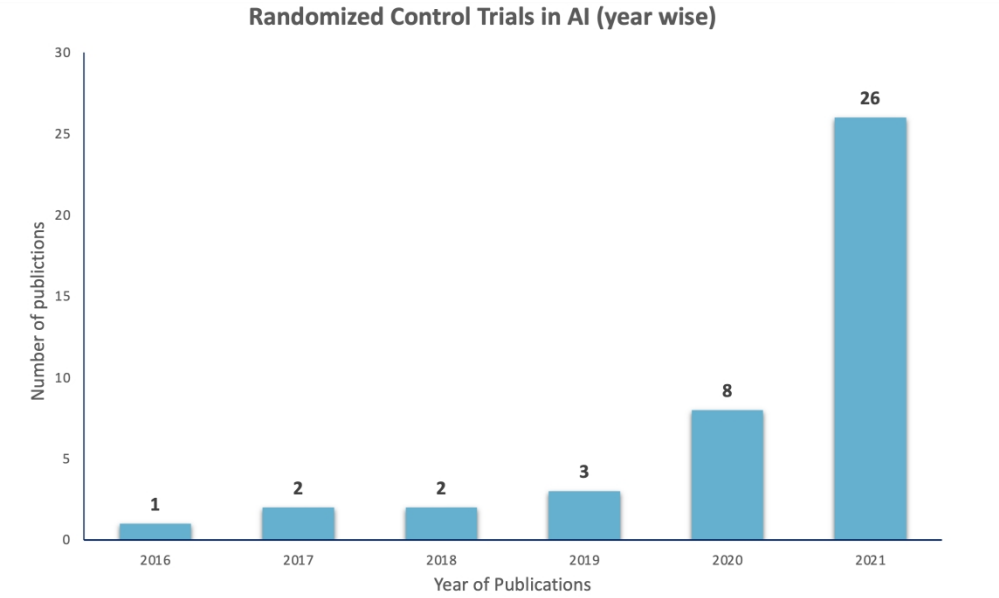


Figure 2: Year-wise distribution of RCTs in AI

403x244mm (144 x 144 DPI)

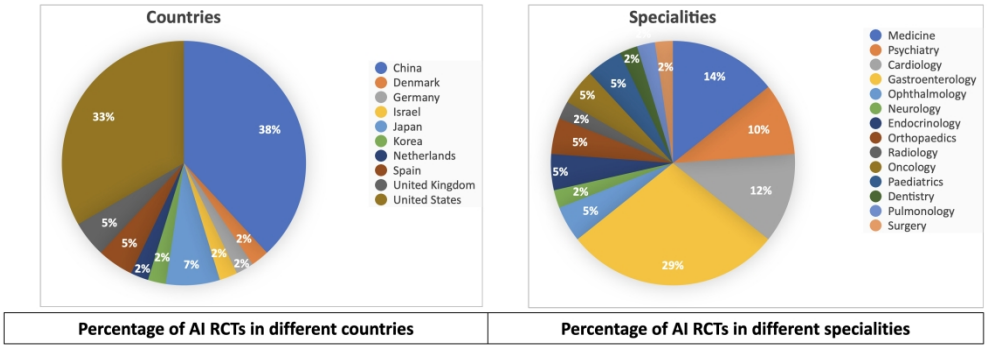


Figure 3: Percentage of AI RCTs in different countries and specialties

608x295mm (144 x 144 DPI)

Supplementary table 1: CONSORT-AI checklist**

Section	Item	CONSORT 2010 item	CONSORT-AI item		Addressed on page no.*
Title and abstract					
Title and abstract	1a	Identification as a randomized trial in the title	CONSORT-AI a,b Elaboration	(i) 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 (for specific guidance see CONSORT for abstracts)		(ii) 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	CONSORT-AI a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare	

				professionals, patients, public).	
	2b	Specific objectives or hypotheses			
Methods					
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio			
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons			
Participants	4a	Eligibility criteria for participants	CONSORT-AI a (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.	
			CONSORT-AI a (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.	
	4b	Settings and locations where the data were collected	CONSORT-AI b Extension	Describe how the AI intervention was integrated into the trial setting, including any onsite or on site requirements.	

Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	CONSORT-AI (i) Extension	State which version of the AI algorithm was used.	
			CONSORT-AI (ii) Extension	Describe how the input data were acquired and selected for the AI intervention.	
			CONSORT-AI (iii) Extension	Describe how poor quality or unavailable input data were assessed and handled.	
			CONSORT-AI (iv) Extension.	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.	
			CONSORT-AI (v) Extension	Specify the output of the AI intervention	
			CONSORT-AI (vi) Extension	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome			

		measures, including how and when they were assessed			
	6b	Any changes to trial outcomes after the trial commenced, with reasons			
Sample size	7a	How sample size was determined			
	7b	When applicable, explanation of any interim analyses and stopping guidelines			
Sequence generation	8a	Method used to generate the random allocation sequence			
	8b	Type of randomization; details of any restriction (such as blocking and block size)			
Randomization					
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the			

		sequence until interventions were assigned			
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions			
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how			
	11b	If relevant, description of the similarity of interventions			
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes			
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses			
Results					

Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome			
	13b	For each group, losses and exclusions after randomization, together with reasons			
Recruitment	14a	Dates defining the periods of recruitment and follow-up			
	14b	Why the trial ended or was stopped			
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group			
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups			
Outcomes and estimation	17a	For each primary and secondary outcome,			

		results for each group, and the estimated effect size and its precision (such as % confidence interval)			
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended			
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory			
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	CONSORT-AI 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					
Limitations	20	Trial limitations, addressing sources of potential bias,			

		imprecision, and, if relevant, multiplicity of analyses			
Generalizability	21	Generalizability (external validity, applicability) of the trial findings			
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence			
Other information					
Registration	23	Registration number and name of trial registry			
Protocol	24	Where the full trial protocol can be accessed, if available			
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	CONSORT-AI Extension.	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	

*Indicates page numbers to be completed by authors during protocol development

** 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.

Supplementary table 2: Included studies

No.	Authors	Year	Title	Country	Specialty	Disease studied	Sample size	Intervention	Control	Blinding	Primary outcome	Trial registration
1.	Sadasivam, et al.	2016	Impact of a Collective Intelligence Tailored Messaging System on Smoking Cessation: The Perspect Randomized Experiment	United States	Medicine	Smoking addiction	120	AI recommended motivational messages	Standard tailored messages	Single (study staff)	Influence of messages	NR
2.	Morrison, et al.	2017	The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone- Based Stress Management	United States	Psychiatry	Stress	77	Smartphone-based stress management system	Daily/occasional notifications within pre-defined	Not stated	Notification response	ISRCTN67177737

			Intervention: An Exploratory Trial						time frames			
3.	Labovitz, et al.	2017	Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy	United States	Cardiology	Ischemic stroke	28	AI system monitoring	No daily monitoring	Not stated	Adherence to therapy	NCT02599259
4.	Rostill, et al.	2018	Technology integrated health management for dementia	United Kingdom	Psychiatry	Dementia	408	Technology-integrated health management for dementia	No TIHM	Not stated	Alerts	NR
5.	Wang, et al	2018	Real-time automatic detection system increases colonoscopic polyp and adenoma	China	Gastroenterology	Adenoma	1058	AI-aided colonoscopy	Standard colonoscopy	None	Adenoma detection rate	ChiCTR-DDD-17012221

			detection rates: a prospective randomised controlled study									
6.	Lin, et al.	2019	Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial	China	Ophthalmology	Cataract	350	AI-assisted cataract detection	Normal clinic	Double	Diagnostic accuracy for congenital cataracts	NCT03240848

7.	Voss, et al.	2019	Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder A Randomized Clinical Trial	United States	Psychiatry	Autism	474	AI-driven behavioral intervention	Applied behavioral analysis therapy	Single	SRS-II, EGG, VABS-II, NEPSY-II socialization scores	NCT03569176
8.	Wu, et al.	2019	Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots	China	Gastroenterology	Upper GI lesions	324	AI-aided esophagogastroduodenoscopy	Standard esophagogastroduodenoscopy	Single	Blind spot rate	ChiCTR1800014809

			during esophagogastroduodenoscopy									
9.	Wang, et al.	2020	Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study	China	Gastroenterology	Adenoma	1046	AI-aided colonoscopy	Sham	Double	Adenoma detection rate	ChiCTR1800017675
10.	Persell, et al.	2020	Effect of Home Blood Pressure Monitoring via a Smartphone Hypertension Coaching	United States	Medicine	Hypertension	333	AI-driven coaching app	Blood pressure tracking app	None	Systolic blood pressure at 6 months	NCT03288142

			Application or Tracking Application on Adults With Uncontrolled Hypertension A Randomized Clinical Trial									
11.	Wijnberge, et al.	2020	Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension	Netherlands	Cardiology	Intra-operative hypotension	68	AI-driven early warning system for intraoperative hypotension	Standard care	None	Time-weighted average of intraoperative hypotension	NCT03376347

			During Elective Noncardiac Surgery The HYPE Randomized Clinical Trial									
12.	Pavel, et al.	2020	A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial	United Kingdom	Neurology	Neonatal seizures	264	Automated seizure detection algorithm	Conventional EEG	Single	Diagnostic accuracy of healthcare professionals with aid of algorithm	NCT02431780
13.	Nimri, et al.	2020	Insulin dose optimization using an automated artificial intelligence-based decision support	Israel	Endocrinology	Diabetes	108	AI-based decision support system	Physician guided care	Single	Time of glucose level within target range	NCT03003806

			system in youths with type 1 diabetes									
14.	Liu, et al.	2020	The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study	China	Gastroenterology	Adenoma	790	AI-aided colonoscopy	Routine colonoscopy	None	Adenoma detection rate	ChiCTR1800018058
15.	Gong, et al.	2020	Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised	China	Gastroenterology	Adenoma	704	AI-aided colonoscopy	Unassisted colonoscopy	Single	Adenoma detection rate	ChiCTR1900021984

			controlled study									
16.	Luo, et al.	2020	Artificial Intelligence-Assisted Colonoscopy for Detection of Colon Polyps: a Prospective, Randomized Cohort Study	China	Gastroenterology	Polyps	150	AI-aided colonoscopy	Unaided colonoscopy	None	Polyp detection rate	NCT04712626 5
17.	Anan, at al.	2021	Effects of an Artificial Intelligence-Assisted Health Program on Workers With Neck/Shoulder Pain/Stiffness and Low Back Pain:	Japan	Orthopaedics	Neck/shoulder and back pain	94	AI-assisted health program	Usual care routine	None	Pain level	(UMIN-CTR) 000033894

			Randomized Controlled Trial									
18.	Blomberg, et al.	2021	Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial	Denmark	Cardiology	Cardiac arrest	654	AI-led alerts	Normal protocol	Double	Recognition of cardiac arrest	NCT04219306
19.	Chen, et al.	2021	The Role of Deep Learning-Based Echocardiography in the Diagnosis and Evaluation of the Effects of Routine Anti-Heart-	China	Cardiology	Heart failure	80	AI-based echocardiography	Routine echocardiography	None	Mortality and rehospitalization rate	NR

			Failure Western Medicines in Elderly Patients with Acute Left Heart Failure									
20.	Eng, et al.	2021	Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment	United States	Radiology	Skeletal age	1903	AI diagnostic aid	Without aid	None	Mean absolute difference between the skeletal age	NCT03530098
21.	Harada, et al.	2021	Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy	Japan	Medicine	Various medical conditions	22	AI-assisted differential diagnosis	Without AI assistance	Single	Diagnostic accuracy	UMIN000042881

			of physicians: An open-label randomized controlled study									
22.	Hassoon, et al.	2021	Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors	United States	Oncology	Different cancer types	42	AI coaching	Written information	Single	Change in steps per day	NCT03212079
23.	Jayakumar, et al.	2021	Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on	United States	Orthopaedics	Osteoarthritis	129	AI-enabled patient decision aid	Educational material	None	Knee OA Decision Quality	NCT03956004

			Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults with Knee Osteoarthritis: A Randomized Clinical Trial									
24.	Kamba, et al.	2021	Reducing adenoma miss rate of colonoscopy assisted by artificial intelligence: a multicenter randomized controlled trial	Japan	Gastroenterology	Adenoma	358	AI-aided colonoscopy	Unaided colonoscopy	None	Adenoma miss rate	jRCTs032190061

25.	Luna, et al.	2021	Artificial intelligence application versus physical therapist for squat evaluation: a randomized controlled trial	United States	Medicine	Physical therapy	30	AI-assisted exercise application	Unassisted exercise	Single	Successful squats	NCT04624594
26.	Medina, et al.	2021	Electrophysiological brain changes associated with cognitive improvement in a pediatric attention deficit hyperactivity disorder digital artificial intelligence-driven intervention:	Spain	Psychiatry	Attention deficit hyperactivity disorder	29	AI-driven cognitive stimulation program	Commerical video games	Single	Conners CPT (CPT-III) score	ISRCTN71041318

			Randomized controlled trial									
27.	Mertens, et al.	2021	Artificial intelligence for caries detection: Randomized trial	Germany	Dentistry	Caries	22	AI-based diagnostic support	Unaided diagnosis	None	Accuracy metrics	DRKS00022357
28.	Prochaska, et al.	2021	A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic	United States	Medicine	Substance-related disorders	180	AI relational conversational agent	No intervention during the study	None	Past-month substance use occasions	NCT04096001
29.	Rafferty, et al.	2021	A novel mobile app (heali) for disease treatment in participants with irritable bowel	United States	Gastroenterology	Irritable bowel syndrome	58	AI dietry mobile app	Educational material	None	Quality of life score	NCT04256551

			syndrome: Randomized controlled pilot trial									
30.	Seok, et al.	2021	A personalized 3d-printed model for obtaining informed consent process for thyroid surgery: A randomized clinical study using a deep learning approach with mesh-type 3d modeling	Korea	Endocrinology	Thyroid lesions	53	AI 3D-printed thyroid model	Without model	None	Patient general knowledge and satisfaction	KCT0005069
31.	Seol, et al.	2021	Artificial intelligence-assisted clinical decision support for childhood asthma management: A	United States	Paediatrics	Childhood asthma	184	AI-assisted decision support	Usual asthma care	Single	Asthma exacerbation occurrence within one year	NCT02865967

			randomized clinical trial									
32.	Strombla d, et al.	2021	Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources: A Randomized Clinical Trial	United States	Surgery	Gynaecol ogical and colorecta l surgery	683	AI-assisted surgical predictions	Standard estimatio n process	None	Accurate surgery duration prediction	NCT0347137 7
33.	Turino, et al.	2021	Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system	Spain	Pulmonolo gy	Obstructi ve sleep apnea	60	Intelligent monitoring system	Standard manage ment	None	Complianc e to CPAP	NCT0311695 8

			based on machine learning aiming to improve continuous positive airway pressure treatment compliance: Randomized controlled trial									
34.	Wang, et al.	2021	Utilization of Ultrasonic Image Characteristics Combined with Endoscopic Detection on the Basis of Artificial Intelligence Algorithm in Diagnosis of Early Upper	China	Gastroenterology	Early gastric cancer	80	Endoscopy with AI-based ultrasound imaging	Endoscopy alone	None	Detection rate of upper gastric cancer	NR

			Gastrointestinal Cancer									
35.	Wu, et al.	2021	Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial	China	Gastroenterology	Early gastric cancer	1050	AI-aided endoscopy	Endoscopy alone	Single	Number of blind spots during endoscopy	ChiCTR1800018403
36.	Wu, et al.	2021	Effect of a deep learning-based system on the miss rate of gastric neoplasms during	China	Gastroenterology	Gastric neoplasm	1886	AI-assisted endoscopy	Unaided endoscopy	Single	Gastric neoplasm miss rate	ChiCTR2000034453

			upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial									
37.	Xu, et al.	2021	The Clinical Value of Explainable Deep Learning for Diagnosing Fungal Keratitis Using in vivo Confocal Microscopy Images	China	Ophthalmology	Fungal keratitis	1089	AI-assisted image reading	Unassisted image reading	None	Accuracy	NR
38.	Xu, et al.	2021	Artificial intelligence-assisted colonoscopy: A prospective, multicenter, randomized	China	Gastroenterology	Polyp	2352	AI-assisted colonoscopy	Conventional colonoscopy	None	Polyp detection rate	ChiCTR1800015607

			controlled trial of polyp detection									
39.	Yao, et al.	2021	Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial	United States	Cardiology	Low ejection fraction	358	AI-enabled electrocardiograms	Usual care	None	New diagnosis of low ejection fraction	NCT04000087
40.	Zeng, et al.	2021	Long-Term Assessment of Rehabilitation Treatment of Sports through Artificial Intelligence Research	China	Medicine	Sports health management	150	AI-based personalized sports management service system	General management	None	Blood glucose, blood pressure, lipids	NR

41.	Zhang, et al.	2021	Artificial Intelligence Algorithm-Based Ultrasound Image Segmentation Technology in the Diagnosis of Breast Cancer Axillary Lymph Node Metastasis	China	Oncology	Breast cancer	90	AI-based ultra sound image segmentation	Routine ultrasound	Double	Accuracy metrics	NR
42.	Zhang, et al.	2021	Value of Rehabilitation Training for Children with Cerebral Palsy Diagnosed and Analyzed by Computed Tomography	China	Paediatrics	Cerebral palsy	73	AI-assisted analysis of brain images	Original images	None	Cerebral artery blood flow velocity (VP) and Vasscular pulse index (PI)	NR

			Imaging Information Features under Deep Learning									

1. Sadasivam RS, Borglund EM, Adams R, Marlin BM, Houston TK. Impact of a collective intelligence tailored messaging system on smoking cessation: the Perspect randomized experiment. *JMIR* 2016;18(11):e285.
2. Morrison LG, Hargood C, Pejovic V, Geraghty AW, Lloyd S, Goodman N, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. *PloS one* 2017;12(1):e0169162.
3. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke* 2017;48(5):1416-9.
4. Rostill H, Nilforooshan R, Morgan A, Barnaghi P, Ream E, Chrysanthaki T. Technology integrated health management for dementia. *Br J Community Nurs* 2018;23(10):502-8.
5. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68(10):1813-9.

6. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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.

7. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr* 2019;173(5):446-54.

8. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* 2019;68(12):2161-9.

9. Wang P, Liu X, Berzin TM, Brown JRG, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020;5(4):343-51.

10. Persell SD, Peprah YA, Lipiszko D, Lee JY, Li JJ, Ciolino JD, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw Open* 2020;3(3):e200255-e.

11. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323(11):1052-60.

12. Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *The Lancet Child & Adolescent Health* 2020;4(10):740-9.

13. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nature medicine* 2020;26(9):1380-4.

14. Liu P, Wang P, Glissen Brown JR, Berzin TM, Zhou G, Liu W, et al. The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study. *Therap Adv Gastroenterol* 2020;13:1756284820979165.

15. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol* 2020;5(4):352-61.

16. Luo Y, Zhang Y, Liu M, Lai Y, Liu P, Wang Z, Xing T, Huang Y, Li Y, Li A, Wang Y. Artificial intelligence-assisted colonoscopy for detection of colon polyps: a prospective, randomized cohort study. *Journal of Gastrointestinal Surgery* 2021;25(8):2011-8.

17. Anan T, Kajiki S, Oka H, et al. Effects of an Artificial Intelligence-Assisted Health Program on Workers With Neck/Shoulder Pain/Stiffness and Low Back Pain: Randomized Controlled Trial. *JMIR Mhealth Uhealth* 2021;9(9):e27535.

18. 1. Blomberg SN, Christensen HC, Lippert F, et al. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw Open* 2021;4(1):e2032320.

19. Chen J, Gao Y. The Role of Deep Learning-Based Echocardiography in the Diagnosis and Evaluation of the Effects of Routine Anti-Heart-Failure Western Medicines in Elderly Patients with Acute Left Heart Failure. *Journal of Healthcare Engineering* 2021;2021

20. Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: A prospective multicenter randomized controlled trial. *Radiology* 2021;301(3):692-99.

21. Harada Y, Katsukura S, Kawamura R, et al. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: An open-label randomized controlled study. *International Journal of Environmental Research and Public Health* 2021;18(4):1-10.

22. Hassoon A, Baig Y, Naiman DQ, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. *npj Digital Medicine* 2021;4(1)

23. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults with Knee Osteoarthritis: A Randomized Clinical Trial. *JAMA Network Open* 2021;4(2)

24. Kamba S, Tamai N, Saitoh I, et al. Reducing adenoma miss rate of colonoscopy assisted by artificial intelligence: a multicenter randomized controlled trial. *Journal of Gastroenterology* 2021;56(8):746-57.

25. Luna A, Casertano L, Timmerberg J, et al. Artificial intelligence application versus physical therapist for squat evaluation: a randomized controlled trial. *Scientific reports* 2021;11(1):18109.

26. Medina R, Bouhaben J, de Ramon I, et al. Electrophysiological brain changes associated with cognitive improvement in a pediatric attention deficit hyperactivity disorder digital artificial intelligence-driven intervention: Randomized controlled trial. *Journal of Medical Internet Research* 2021;23(11)

27. Mertens S, Krois J, Cantu AG, et al. Artificial intelligence for caries detection: Randomized trial. *J Dent* 2021;115:103849.

28. Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug and Alcohol Dependence* 2021;227

29. Rafferty AJ, Hall R, Johnston CS. A novel mobile app (heali) for disease treatment in participants with irritable bowel syndrome: Randomized controlled pilot trial. *Journal of Medical Internet Research* 2021;23(3)

30. Seok J, Yoon S, Ryu CH, et al. A personalized 3d-printed model for obtaining informed consent process for thyroid surgery: A randomized clinical study using a deep learning approach with mesh-type 3d modeling. *Journal of Personalized Medicine* 2021;11(6)

31. Seol HY, Shrestha P, Muth JF, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: A randomized clinical trial. *PLoS ONE* 2021;16(8 August)

32. Stromblad CT, Baxter-King RG, Meisami A, et al. Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources: A Randomized Clinical Trial. *JAMA Surgery* 2021;156(4):315-21.

33. Turino C, Benitez ID, Rafael-Palou X, et al. Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: Randomized controlled trial. *Journal of Medical Internet Research* 2021;23(10)

34. Wang L, Song H, Wang M, et al. Utilization of Ultrasonic Image Characteristics Combined with Endoscopic Detection on the Basis of Artificial Intelligence Algorithm in Diagnosis of Early Upper Gastrointestinal Cancer. *J Healthc Eng* 2021;2021:2773022.

35. Wu L, He X, Liu M, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. *Endoscopy* 2021;53(12):1199-207

36. Wu L, Shang R, Sharma P, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. *The Lancet Gastroenterology and Hepatology* 2021;6(9):700-08.

37. Xu F, Jiang L, He W, et al. The Clinical Value of Explainable Deep Learning for Diagnosing Fungal Keratitis Using in vivo Confocal Microscopy Images. *Frontiers in Medicine* 2021;8

38. Xu L, He X, Zhou J, et al. Artificial intelligence-assisted colonoscopy: A prospective, multicenter, randomized controlled trial of polyp detection. *Cancer Medicine* 2021;10(20):7184-93.

39. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nature Medicine* 2021;27(5):815-19.

40. Zeng C, Huang Y, Yu L, et al. Long-Term Assessment of Rehabilitation Treatment of Sports through Artificial Intelligence Research. *Comput Math Methods Med* 2021;2021:4980718.

41. Zhang L, Jia Z, Leng X, et al. Artificial Intelligence Algorithm-Based Ultrasound Image Segmentation Technology in the Diagnosis of Breast Cancer Axillary Lymph Node Metastasis. *J Healthc Eng* 2021;2021:8830260.

42. Zhang X, Wang Z, Liu J, et al. Value of Rehabilitation Training for Children with Cerebral Palsy Diagnosed and Analyzed by Computed Tomography Imaging Information Features under Deep Learning. *J Healthc Eng* 2021;2021:6472440.

Search Strategy:

FOR EMBASE:

1) *deep learning/
2) *artificial intelligence/
3) *machine learning/
4) 1 or 2 or 3

For PubMed:

1. Artificial intelligence
2. Machine learning
3. Deep learning
4. 1 OR 2 OR 3

Restricted to:

- Article type: Randomized Control Trial
- Publication Date: 1/01/2015 to 31/12/2021
- Species: Human
- Language: English

Reporting checklist for systematic review (with or without a meta-analysis).

Based on the PRISMA guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the PRISMAreporting guidelines, and cite them as:

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews

		Reporting Item	Page Number
Title			
Title	#1	Identify the report as a systematic review	1
Abstract			
Abstract	#2	Report an abstract addressing each item in the PRISMA 2020 for Abstracts checklist	3
Introduction			
Background/rationale	#3	Describe the rationale for the review in the context of existing knowledge	4
Objectives	#4	Provide an explicit statement of the objective(s) or question(s) the review addresses	5
Methods			

Eligibility criteria	#5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses	6
Information sources	#6	Specify all databases, registers, websites, organisations, reference lists, and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted	6
Search strategy	#7	Present the full search strategies for all databases, registers, and websites, including any filters and limits used	6
Selection process	#8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and, if applicable, details of automation tools used in the process	6
Data collection process	#9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and, if applicable, details of automation tools used in the process	7
Data items	#10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (for example, for all measures, time points, analyses), and, if not, the methods used to decide which results to collect	7
Study risk of bias assessment	#11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and, if applicable, details of automation tools used in the process	n/a
Effect measures	#12	Specify for each outcome the effect measure(s) (such as risk ratio, mean difference) used in the synthesis or presentation of results	7
Synthesis methods	#13a	Describe the processes used to decide which studies were eligible for each synthesis (such as tabulating the study	n/a

		intervention characteristics and comparing against the planned groups for each synthesis (item #5))	
Synthesis methods	#13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics or data conversions	n/a
Synthesis methods	#13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses	n/a
Synthesis methods	#13d	Describe any methods used to synthesise results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used	n/a
Synthesis methods	#13e	Describe any methods used to explore possible causes of heterogeneity among study results (such as subgroup analysis, meta-regression)	n/a
Synthesis methods	#13f	Describe any sensitivity analyses conducted to assess robustness of the synthesised results	n/a
Reporting bias assessment	#14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases)	n/a
Certainty assessment	#15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome	n/a
Data items	#10b	List and define all other variables for which data were sought (such as participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information	n/a
Results			
Study selection	#16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram (http://www.prisma-statement.org/PRISMAStatement/FlowDiagram)	7
Study selection	#16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded	7

Study characteristics	#17	Cite each included study and present its characteristics	supplementary table 2
Risk of bias in studies	#18	Present assessments of risk of bias for each included study	n/a
Results of individual studies	#19	For all outcomes, present for each study (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (such as confidence/credible interval), ideally using structured tables or plots	8
Results of syntheses	#20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies	n/a
Results of syntheses	#20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (such as confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect	n/a
Results of syntheses	#20c	Present results of all investigations of possible causes of heterogeneity among study results	n/a
Results of syntheses	#20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesised results	n/a
Risk of reporting biases in syntheses	#21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed	n/a
Certainty of evidence	#22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed	n/a
Discussion			
Results in context	#23a	Provide a general interpretation of the results in the context of other evidence	11
Limitations of included studies	#23b	Discuss any limitations of the evidence included in the review	14
Limitations of the review methods	#23c	Discuss any limitations of the review processes used	14
Implications	#23d	Discuss implications of the results for practice, policy, and future research	14

Other information

Registration and protocol	#24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered	7
Registration and protocol	#24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared	7
Registration and protocol	#24c	Describe and explain any amendments to information provided at registration or in the protocol	7
Support	#25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review	15
Competing interests	#26	Declare any competing interests of review authors	15
Availability of data, code, and other materials	#27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review	supplementary table 2

Notes:

- 17: supplementary table 2
- 27: supplementary table 2 The PRISMA checklist is distributed under the terms of the Creative Commons Attribution License CC-BY. This checklist was completed on 12. July 2022 using https://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai