Detecting Algorithmic Errors and Patient Harms for Artificial Intelligence (AI) enabled Medical Devices in Randomised Controlled Trials: A systematic review protocol

Aditya U Kale, Henry David Jeffry Hogg, Russell Pearson, Ben Glocker, Su Golder, April Coombe, Justin Waring, Xiaoxuan Liu, David J Moore, Alastair K Denniston

Submitted to: JMIR Research Protocols
on: August 23, 2023

Table of Contents

Detecting Algorithmic Errors and Patient Harms for Artificial Intelligence (AI) enabled Medical Devices in Randomised Controlled Trials: A systematic review protocol

Aditya U Kale[1, 2, 3, 4] MBChB; Henry David Jeffry Hogg[5] PhD; Russell Pearson[6] PhD; Ben Glocker[7, 8] PhD; Su Golder[9] PhD; April Coombe[10] MSc; Justin Waring[11] PhD; Xiaoxuan Liu[1, 2, 3, 4] PhD; David J Moore[10] PhD; Alastair K Denniston[1, 2, 3, 4] PhD

[1]Institute of Inflammation and Ageing University of Birmingham Birmingham GB
[2]University Hospitals Birmingham NHS Foundation Trust Birmingham GB
[3]NIHR Birmingham Biomedical Research Centre Birmingham GB
[4]NIHR Incubator for AI and Digital Health Research Birmingham GB
[5]Population Health Science Institute Faculty of Medical Sciences Newcastle University Newcastle upon Tyne GB
[6]Medicines and Healthcare Products Regulatory Agency London GB
[7]Kheiron Medical Technologies London GB
[8]Department of Computing Imperial College London London GB
[9]Department of Health Sciences University of York York GB
[10]Institute of Applied Health Research University of Birmingham Birmingham GB
[11]Health Services Management Centre University of Birmingham Birmingham GB

Corresponding Author:
Alastair K Denniston PhD
Institute of Inflammation and Ageing
University of Birmingham
Edgbaston
Birmingham
GB

Abstract

Background: AI health technologies have the potential to transform existing clinical workflows and ultimately improve patient outcomes. AI health technologies have shown potential for a range of clinical tasks such as diagnostics, prognostics, and therapeutic decision making such as drug dosing. There is however an urgent need to ensure that AI health technologies remain safe for all populations. Recent literature demonstrates the need for rigorous performance error analysis to identify issues such as algorithmic encoding of spurious correlations (e.g. protected characteristics), or specific failure modes that may lead to patient harm. Guidelines for reporting of studies evaluating AI health technologies (e.g. CONSORT-AI) require mention of performance error analysis, however there is still a lack of understanding around how performance errors should be analysed in clinical studies, and what harms authors should aim to detect and report.

Objective: This systematic review will assess the frequency, severity of AI errors and patient harms in randomised controlled trials (RCTs) investigating AI interventions in clinical settings. The review will also explore how performance errors are analysed including whether analysis includes investigation of subgroup level outcomes.

Methods: This systematic review will identify and select randomised controlled trials assessing AI interventions. Search strategies will be deployed in MEDLINE, EMBASE, Cochrane CENTRAL and clinical trials registries to identify relevant articles. RCTs identified in bibliographic databases will be cross-referenced with clinical trials registries. The primary outcomes of interest are the frequency and severity of AI errors, patient harms and reported adverse events. Quality assessment of RCTs will be based on RoB2. Data analysis will include comparison of error rates and patient harms between study arms and a meta-analysis of the rates of patient harm in control versus intervention arms will be conducted if appropriate.

Results: The project was registered on PROSPERO in February 2023. Preliminary searches have been completed and the search strategy has been designed in consultation with an information specialist (see appendices 1 and 2). Abstract screening will start in August 2024.

Conclusions: Evaluations of AI health technology have shown promising results, however reporting of studies has been variable. Detection, analysis and reporting of performance errors and patient harms is vital to robustly assess the safety of AI interventions in RCTs. Scoping searches have illustrated that reporting of harms is variable, often with no mention of adverse events. The findings of this systematic review will identify the frequency and severity of AI performance errors and patient harms, and generate insights into how errors should be analysed to account for both overall and subgroup performance.

Systematic review registration
PROSPERO CRD42023387747.

(JMIR Preprints 23/08/2023:51614)
DOI: https://doi.org/10.2196/preprints.51614

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Detecting Algorithmic Errors and Patient Harms for Artificial Intelligence (AI) enabled Medical Devices in Randomised Controlled Trials: A systematic review protocol

Authors:

Aditya U Kale[1,2,3,4], Henry David Jeffry Hogg[5], Russell Pearson[6], Ben Glocker[7,8], Su Golder[9], April Coombe[11], Justin Waring[10], Xiaoxuan Liu[1,2,3,4], David J Moore[11*], Alastair K Denniston[1,2,3,4*]

Affiliations:

1. Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK
2. University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
3. NIHR Birmingham Biomedical Research Centre, Birmingham, UK
4. NIHR Incubator for AI and Digital Health Research, Birmingham, UK
5. Population Health Science Institute, Faculty of Medical Science, Newcastle University, Newcastle upon Tyne
6. Medicines and Healthcare Products Regulatory Agency, London, UK
7. Kheiron Medical Technologies, London, UK
8. Imperial College London, Department of Computing, London, UK
9. Department of Health Sciences, University of York, York, UK
10. Health Services Management Centre, University of Birmingham, Birmingham, UK
11. Institute of Applied Health Research, University of Birmingham, Birmingham, UK

*AKD and DJM are joint senior authors

Corresponding author:

Professor Alastair Denniston

e: a.denniston@bham.ac.uk

Phone: +44 (0)121 371 6905

Address:

Institute of Inflammation and Ageing

College of Medical and Dental Sciences

University of Birmingham

Edgbaston
Birmingham
B15 2TT
UK

Abstract

Background

AI medical devices have the potential to transform existing clinical workflows and ultimately improve patient outcomes. AI medical devices have shown potential for a range of clinical tasks such as diagnostics, prognostics, and therapeutic decision making such as drug dosing. There is however an urgent need to ensure that these technologies remain safe for all populations. Recent literature demonstrates the need for rigorous performance error analysis to identify issues such as algorithmic encoding of spurious correlations (e.g. protected characteristics), or specific failure modes that may lead to patient harm. Guidelines for reporting of studies evaluating AI medical devices (e.g. CONSORT-AI) require mention of performance error analysis, however there is still a lack of understanding around how performance errors should be analysed in clinical studies, and what harms authors should aim to detect and report. This systematic review will assess the frequency and severity of AI errors and adverse events in randomised controlled trials (RCTs) investigating AI medical devices as interventions in clinical settings. The review will also explore how performance errors are analysed including whether analysis includes investigation of subgroup level outcomes.

Methods

This systematic review will identify and select randomised controlled trials assessing AI medical devices. Search strategies will be deployed in MEDLINE, Embase, Cochrane CENTRAL and clinical trials registries to identify relevant articles. RCTs identified in bibliographic databases will be cross-referenced with clinical trials registries. The primary outcomes of interest are the frequency and severity of AI errors, patient harms and reported adverse events. Quality assessment of RCTs will be based on version 2 of the Cochrane risk-of-bias tool (RoB2). Data analysis will include comparison of error rates and patient harms between study arms and a meta-analysis of the rates of patient harm in control versus intervention arms will be conducted if appropriate.

Results

The project was registered on PROSPERO in February 2023. Preliminary searches have been completed and the search strategy has been designed in consultation with an information specialist and methodologist. Title and abstract screening started in September 2023. Full text screening is ongoing and data collection and analysis will begin in April 2024.

Discussion

Evaluations of AI medical devices have shown promising results, however reporting of studies has been variable. Detection, analysis and reporting of performance errors and patient harms is vital to robustly assess the safety of AI medical devices in RCTs. Scoping searches have illustrated that reporting of harms is variable, often with no mention of adverse events. The findings of this systematic review will identify the frequency and severity of AI performance errors and patient harms and generate insights into how errors should be analysed to account for both overall and subgroup performance.

Systematic review registration
PROSPERO CRD42023387747.

Word limit: 450

Background

Artificial intelligence (AI), the use of machines to undertake complex processes that would usually require human intelligence, has the potential to transform healthcare.[1,2] The potential benefits of such data-led technologies include a wide range of clinical applications, such as faster diagnosis, prognostics, digital therapeutics, and even detection of novel signals.[3–5] Although there has been a great deal of enthusiasm around AI medical device, performance in virtual test environments is often different to that in the real world.[6–8] There is an urgent need to investigate how such technologies can be evaluated and monitored to ensure clinical benefit and avoid patient harm.[9–12]

AI errors and patient harms

The translation of AI medical device from ‘code to clinic’ is complex and if planned poorly can lead to serious safety concerns.[13,14] Safety assessments involve understanding risks associated with AI medical devices, including what AI errors can arise, how these might lead to patient harms and what failure modes may exist. These concepts are defined below:

Adverse events (AEs)	*“An unfavourable outcome that occurs during or after the use of a drug or other intervention but is not necessarily caused by it”*[15,16]
AI errors	*“Any outputs of the AI system which are inaccurate, including those which are inconsistent with expected performance and those which can result in harm if undetected or detected too late.”*[9]
Failure modes	*“The tendency to malfunction in the presence of certain conditions. Whereas an error can be a single occurrence, failure modes represent errors which will repeatedly occur and often have similar consequences.”*[9]
Patient harms	*“Injury or damage to the health of people” (as defined in ISO 14971-application of risk management for medical devices).)*[17] “*The totality of possible adverse consequences of an intervention or therapy”*[18]

Table 1: Glossary of terms

Performance evaluation and monitoring of AI medical device

AI medical device safety and effectiveness evidence can be generated at various stages in the evaluation process, which can be broadly divided into pre- and post-market. Pre-market evaluation includes a range of study types such as test accuracy studies and randomised controlled trials (RCTs).

Post-market evaluation on the other hand includes local assurance practices and ongoing monitoring. Several study designs exist for generation of effectiveness evidence, with the most robust evidence in terms of minimising bias and objectively measuring the effect of AI interventions on clinical outcomes being derived from prospective, randomised controlled trials.[19] Recent literature demonstrates the importance of in depth performance error analysis including identification of "inhuman errors" (e.g. highly displaced fractures missed by AI), testing for algorithmic encoding of protected characteristics, and conducting exploratory error analyses to identify cases of hidden stratification.[20–22] An AI medical device might be shown to perform well *overall*, however without more rigorous error analysis including exploratory and subgroup analysis, it is not possible to truly understand the clinical impact on patients as individuals. The concept of performance error analysis has been outlined in the recent AI extensions reporting guidelines for clinical trials and trial protocols (CONSORT-AI and SPIRIT-AI).[23,24] Recent systematic reviews demonstrate that that the quality of reporting of RCTs is remains both suboptimal and variable.[25,26] The reviews both demonstrated poor adherence of published RCTs to the CONSORT-AI reporting guidelines. There is still minimal literature specifically describing the reporting and analysis of errors and adverse events, and how performance error analysis is being conducted. There is a need to conduct a literature review in this area to inform future clinical evaluations of AI medical devices and real-world adverse event reporting. This systematic review aims to explore AI errors and adverse event reporting in RCTs of AI interventions.

Purpose

This systematic review will assess the frequency and severity of AI errors and adverse events in randomised controlled trials (RCTs) investigating AI medical devices as interventions in clinical settings. Where reported, data regarding AI system risks, reported errors and how these errors were analysed will be extracted. Our research question is:

What are the characteristics (including frequency and severity) of AI errors and adverse events in RCTs and how are these performance errors analysed?

Aim

The primary aims of this review are to 1) assess the frequency, severity and types of errors and adverse events reported in RCTs of AI medical devices.

Secondary aims of the review include: 1) identifying what analyses are conducted when errors or

harms are reported, 2) reporting the error and AE detection methods used.

Methods

Protocol

This systematic review protocol is written in compliance with the guidelines of Preferred Reporting Items for Systematic Review and Meta-Analysis Protocol (PRISMA-P).[27] The completed systematic review will be reported in line with PRISMA guidance.[28] PRISMA-AI will be used if published prior to the submission of this systematic review.[29]

Systematic review registration

This systematic review protocol is registered on PROSPERO CRD42023387747.

Information sources

The search strategy will be used to search three online bibliographic databases, in addition to clinical trial registries to identify RCTs evaluating AI interventions in clinical settings. Literature searches will not be limited by year to ensure that all AI medical device RCTs are identified.

- Bibliographic databases of published studies
    - MEDLINE (Ovid)
    - Embase (Ovid)
    - Cochrane CENTRAL
- Registers of clinical trials
    - Clinicaltrials.gov
    - WHO International Clinical Trials Registry Platform (ICTRP portal)

Search Strategy

In bibliography databases, free text and index terms will be used to search for RCTs of AI medical devices. Clinical trials registries will be searched using in-built filters to identify RCTs with results. RCTs identified in bibliographic databases will be cross-referenced using clinical trials registries to ensure that all harms data is captured. The search strategy has been developed in consultation with an information specialist (author AC) and further details are included in appendix 1. The searches were executed on 30th June 2023. No date cut-off was applied. Reference lists of included reports will be checked to capture additional RCTs. Additionally, experts in the field will be contacted to identify

reports that were not available from the databases listed above.

Selection criteria

The selection criteria are structured using the Studies, Data, Methods, Outcome measures (SDMO) framework for methodological systematic reviews which was deemed most appropriate and adapted for this study.[30] Studies not published in the English language will be included where translation is available.

Types of studies:

Only RCTs will be included in this systematic review. Other study types including non-randomized clinical trials, observational studies and case studies will be excluded. The review will include trials where randomisation happens at any level (such as cluster randomisation, and cross-over randomised controlled trials).

Types of data:

AI medical device interventions which directly affect patient care will be included, for example diagnostic, prognostic or therapeutic tasks. AI medical devices will be included if their function as described within the trial was consistent with the function of a medical device, i.e. within the range of functions attributed to medical devices as defined by the International Medical Device Regulators Forum (IMDRF).[31] AI medical devices that are deployed for non-clinical tasks will be excluded. RCTs evaluating robotic interventions will also be excluded.

Types of methods:

RCTs with control arms involving a non-AI standard of care will be included. RCTs with only AI enabled control arms will be excluded. Additionally, the review will include trials where error analysis has been conducted.

Outcomes:

RCTs reporting AEs, patient harms (not explicitly reported as AEs) will be included in the final analysis. Studies not involving these outcomes will be examined to extract data relating to the RCT design and characteristics of the AI medical device.

Selection process

Once articles have been identified through the search strategy, the studies will be screened for relevance by title and abstract. The rayyan.ai (Rayyan, USA) systematic review tool will be used to screen results.[32] Irrelevant studies will be removed. This process will be carried out by two reviewers independently and any discrepancies will be resolved by discussion, or referral to an arbitrator.

Articles identified as potentially relevant will then be retrieved and the full text will be assessed for inclusion against the selection criteria described above. During full text screening, the studies will also be assessed for presence of patient harm data or any form of performance error analysis. Those with this data present will be marked for full extraction and risk of bias assessment, and those that do not report this data will be marked for extraction of the RCT design and AI technology characteristics only. This will again be done by two reviewers independently with recourse to arbitration if required.

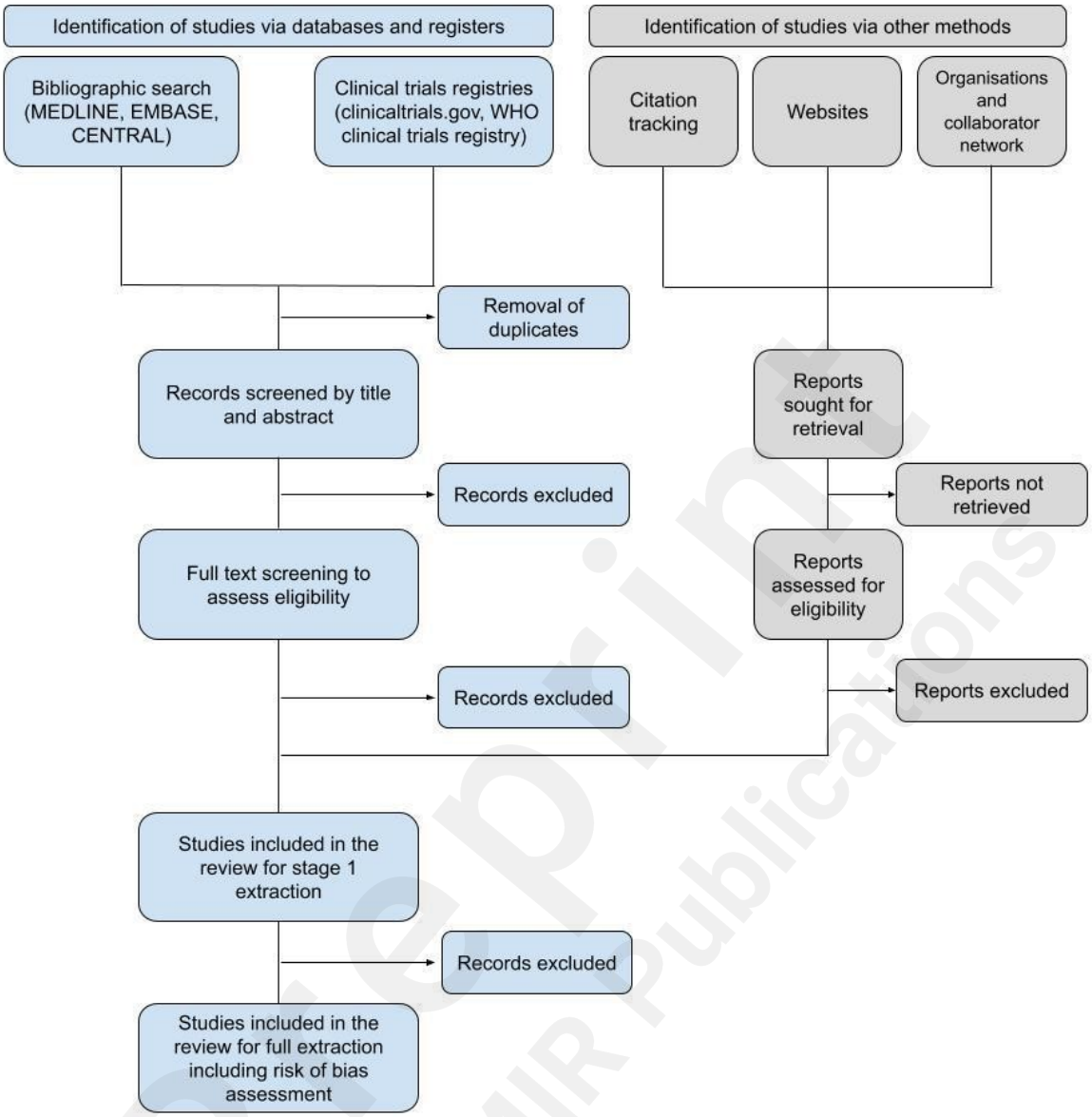


Figure 1: PRISMA flow diagram outline to be populated during the systematic review process. Where RCTs of AI medical devices do not report errors or adverse events, data relating to the type of AI medical device and trial design. This is signposted as stage 1 extraction in the PRISMA diagram. Further details are included in the data extraction section.

Data extraction

The data extraction process will be undertaken using a standardised, piloted data extraction form. Data will be entered into the data extraction form in Microsoft Excel (Microsoft, Washington, UK). This will be done by two reviewers who will complete data extraction independently using the agreed data extraction template. Authors of articles will be contacted for further information and clarification where required. Where available, the following items will be extracted.

Study characteristics:

- Title, authors, publication year, journal, country
- Specialty (medical discipline e.g. radiology, ophthalmology or cardiology)
- Study context (e.g. primary care, hospital care)
- Study design
- Sample size
- Study length (time period)
- Control arm comparator (overview of workflow)
- Baseline characteristic subgroups (e.g. sex, age, ethnicity, socioeconomic details)
- Primary and secondary endpoints

Characteristics of the AI medical device:

- Name of AI medical device
- AI developer (and manufacturer where relevant)
- AI subtype for example ‘recurrent neural network’
- AI intended use and clinical pathway (context)
- AI autonomy level (i.e. the extent to which human oversight is expected). The autonomy level will be graded from one to five based on classification described in the literature.[33]
- Input data
- AI output
- Role in clinical decision-making
- Characteristics of end-user (e.g. clinician or patient)

Outcomes and findings:

- Primary outcomes (to satisfy primary objectives of systematic review):
    - Frequency of AI errors
    - Frequency and severity of adverse events (as classified by relevant regulatory documents including ISO 14971- application of risk management for medical devices) in all study arms[17]
    - Characteristics of error, patient harm and adverse events identified
- Secondary outcomes (to satisfy secondary objectives of systematic review):
    - Types of performance error analysis e.g. subgroup analysis by patient or task

characteristics
- Error and AE detection method described in the study and risk mitigations in place during the RCT

Reporting of adverse events and performance error analysis:

Characteristics of the AI medical device being evaluated will be extracted for all included RCTs. Full data extraction will only be completed for studies reporting some form of adverse events (or possible patient harms not explicitly reported by authors), or details of performance error analysis (item 19 of the CONSORT-AI extension).[23] Performance error analysis is defined as any of: 1) exploratory error analysis, 2) subgroup analysis, 3) adversarial testing.[9]

Quality assessment

Assessment of quality will be carried out for all included studies. Version 2 of the Cochrane risk-of-bias tool (RoB2) for randomised trials will be used to assess studies.[34] Assessment will be undertaken by two reviewers independently with arbitration by a third reviewer where required. The risk is categorised into ‘low’ or ‘high’, or alternatively ‘some concerns’.

Data synthesis

Findings will be synthesised in both narrative and tabular formats. Included studies will be divided into three groups (1, 2a and 2b as shown below) for within-group (and between-group where possible) comparison, based on the AI medical device type and RCT study design.

1. Studies assessing therapeutic AI medical devices (e.g. drug dosing algorithms, AI-enabled psychological therapies)
2. Studies assessing diagnostic or predictive AI medical devices
    a. With ground truth (where ground truth is a reference test e.g. biopsy result, or clinician opinion)
    b. Without ground truth

The synthesis of data will be divided into two sections consistent with the aims outlined in this protocol. The first section is focused on the primary aims of the review: the frequency, severity and types of AI errors and patient harms. The second section is focused on the secondary aims of the review: 1) the reporting of harms data based on the CONSORT harms extension, 2) types of performance error analysis described, and 3) identified subgroups of interest for each health area.

Analysis to achieve primary aims

AI error and patient harm rates will be calculated for each RCT. This data will be compared between and within the identified groups. The following analyses will be considered:

- Reported adverse events with comparison between AI and control arms
    - Frequency and severity of adverse events for each technology, with comparison between AI medical device groups listed above.
    - Whether the adverse event was directly linked to the AI medical device (as assessed by RCT authors).
    - Severity of adverse events will be based on guidance from international standards (ISO 14971- application of risk management for medical devices).[17]
- The frequency of errors e.g. false positive/false negative for diagnostic AI medical devices. If the AI output is reported as likelihood distribution, then the analysis will be directed by the subsequent clinical action taken in response to the AI output. If a ground truth is present in the study, then a comparison can be made.
    - Comparison within and between AI medical device groups listed above. The type of algorithm utilised by the AI medical device will also be included for comparison.
    - If appropriate, a meta-analysis will be conducted investigating harms as a proportion of total outputs for intervention versus control arms. Appropriateness will be defined by assessing the heterogeneity of trial characteristics. Assessment of heterogeneity will include consideration of trial design, primary outcomes and the types of reported adverse events.
- Characterisation of errors and harms for AI medical devices
    - Comparison between AI medical device error rate and erroneous clinical action. For example, if the AI medical device output incorrectly suggests administration of a drug, is this drug actually administered?
    - Harms that are identified but not explicitly reported by authors will also be extracted where possible.

Analysis to achieve secondary aims

- Failure modes- the number of studies describing subgroup and exploratory error analysis will be recorded.
    - Subgroup analysis of AI medical device performance for the clinical task will be

documented. Subgroups of interest described in RCTs will be documented for each medical specialty.

- Exploratory error analysis will be documented with specific focus on the types of scenarios most likely to cause errors for each clinical use case. Described failure modes will be documented for each medical specialty and clinical task.
- The types of performance analysis conducted for each type of AI medical device and clinical discipline will be compared to identify groups with

- Error and AE detection methods will be recorded for each study. The extraction of AI medical device characteristics for all identified RCTs (including those excluded from full extraction) will demonstrate trends in AI medical devices with no adverse events or implicit patient harms. This will allow for identification of areas where adverse event detection methods are particularly underdeveloped, or less frequently utilised. An example of an AE detection method is the use of questionnaires to allow patients to self-report AEs after interaction with an AI enabled mental health chatbot.

Results

The project was registered on PROSPERO in February 2023. Preliminary searches have been completed and the search strategy has been designed in consultation with an information specialist and methodologist (authors AC and DJM). Searches were conducted in June 2023. Title and abstract screening began in September 2023 and finished in February 2024. After deduplication, 11,913 articles were screened resulting in 423 eligible studies for full text screening. Full text screening is currently ongoing and to be completed in April 2024. Data extraction will commence in April 2024. Data Analysis and manuscript drafting will be conducted from May 2024 to June 2024.

Discussion

The potential value of AI medical devices are well recognised, and numerous studies have been published recently relating to model development and evaluation.[35,36] Although AI medical devices show promise, there are still barriers to their deployment at scale. One of the most important related challenges is ensuring that these technologies are effective, safe and inclusive. As an interventional study, RCTs allow measurement of clinically relevant outcomes including patient harms that would not be possible in an in silico study. As a randomised clinical trial, the study design minimises bias, and is therefore considered the gold standard of clinical evidence.

This systematic review aims to assess the frequency and severity of AI errors and adverse events.

Data will be extracted regarding *how* adverse events and AI errors are analysed such as sub-group analysis and identification of failure modes. Investigating the severity and frequency of errors and adverse events in addition to how these are reported in RCTs may provide insights into study design, real-world impacts, and methods for evaluating unintended effects of AI medical devices. The systematic review will not only shed light on which AI medical devices or RCT designs most commonly report AEs, but also on the methods used for AE detection. A summary of these methods will be an important part of the insights generated by this study. The main anticipated limitation of this systematic review is the heterogeneity of outcomes across the different medical disciplines and types of AI medical device. This will be addressed by grouping RCTs based on type of AI medical device and medical specialty where appropriate. The benefits of a broad review in this instance outweigh the limitations given the lack of consensus in the analysis and reporting of AI errors and adverse events. Furthermore, recent literature reviews have demonstrated poor adherence to CONSORT-AI guidelines which indicates a reporting limitation. This means that if no AI errors or adverse events are reported, this will not necessarily stipulate that none had occurred in the study. Finally, AI error may or may not lead to clinical error and there will be other instances where clinical error is introduced by human involvement in the workflow. Mapping of clinical workflows and analysing work system elements will be important, however there might be reporting limitations. Where relevant, authors may be contacted for further information.

There is a growing unmet need for methods enabling detection, analysis and reporting of AI errors and adverse events relating to AI medical device usage. This systematic review aims to be the first of its kind focused on errors and adverse events associated with AI medical devices in healthcare. The impact of this systematic review will be two-fold. Firstly, it will demonstrate current practices in error and adverse event detection, analysis and reporting, forming the basis for further work around best practices for AI harms in RCTs. Secondly, we hope that this work will inform the real-world deployment of AI medical devices, particularly safety monitoring and risk mitigation practices which is an area of significant interest globally. This will be achieved through signposting of best practices for adverse event detection and performance error analysis identified through the review. This is part of a wider programme of work looking at post-market safety monitoring for AI medical devices. A complementary systematic review focusing on AEs reported in regulatory databases is also being conducted.

Acknowledgments

This research is supported by the NIHR Birmingham Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Funding statement

This systematic review is not funded by a research grant.

Conflicts of Interest

The authors declare no conflicts of interest.

Author contributions

All authors have contributed to the design and development of this systematic review. AUK, XL, DJM, and AKD contributed directly to the drafting of the manuscript. AUK, XL, AC, AKD and DJM developed the search strategy. All authors contributed to reviewing and redrafting of this manuscript. AKD and DJM are joint senior authors.

Abbreviations

AI- Artificial Intelligence

AE- Adverse Event

IMDRF- International Medical Device Regulators Forum

ISO- International Organisation for Standardisation

RCT- Randomised Controlled Trial

Multimedia Appendix 1: Development of search strategy for MEDLINE and EMBASE

Multimedia Appendix 2: Search strategies

References

1. Samoili S, Cobo ML, Gomez E, De Prato G, Martinez-Plumed F, Delipetrev B. AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence. 2020 [cited 2022 Aug 24]; Available from: https://eprints.ugd.edu.mk/28047/

2. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017 Dec;2(4):230–43.

3. Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into Systemic Disease through Retinal Imaging-Based Oculomics. Transl Vis Sci Technol. 2020 Feb 12;9(2):6.

4. Zarins CK, Taylor CA, Min JK. Computed fractional flow reserve (FFTCT) derived from coronary CT angiography. J Cardiovasc Transl Res. 2013 Oct;6(5):708–14.

5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019 Oct 29;17(1):195.

6. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022 Dec;6(12):1330–45.

7. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TWV, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep. 2021 Nov 26;11(1):23017.

8. Vaid A, Sawant A, Suarez-Farinas M, Lee J, Kaul S, Kovatch P, et al. Real-world usage diminishes validity of Artificial Intelligence tools [Internet]. bioRxiv. 2022. Available from: https://www.medrxiv.org/content/10.1101/2022.11.17.22282440v1

9. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022 May;4(5):e384–97.

10. Medicines & Healthcare Products Regulatory Agency. Software and AI as a Medical Device Change Programme [Internet]. Gov.uk. 2021 [cited 2022 Jan 12]. Available from: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme

11. Lundström C, Lindvall M. Mapping the Landscape of Care Providers' Quality Assurance Approaches for AI in Diagnostic Imaging. J Digit Imaging [Internet]. 2022 Nov 9; Available from: http://dx.doi.org/10.1007/s10278-022-00731-7

12. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022 May 31;5(1):66.

13. Campbell JP, Mathenge C, Cherwek H, Balaskas K, Pasquale LR, Keane PA, et al. Artificial Intelligence to Reduce Ocular Health Disparities: Moving From Concept to Implementation. Transl Vis Sci Technol. 2021 Mar 1;10(3):19.

14. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019 Mar;28(3):231–7.

15. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons; 2019. 736 p.

16. Chou R, Aronson N, Atkins D, Ismaila AS, Santaguida P, Smith DH, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):502–12.

17. ISO 14971:2019 [Internet]. ISO. 2019 [cited 2022 Aug 25]. Available from: https://www.iso.org/standard/72704.html

18. Ioannidis JPA, Evans SJW, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med. 2004 Nov 16;141(10):781–8.

19. Taylor-Phillips S, Seedat F, Kijauskaite G, Marshall J, Halligan S, Hyde C, et al. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. Lancet Digit Health. 2022 Jul;4(7):e558–65.

20. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. New York, NY, USA: Association for Computing Machinery; 2020. p. 151–9. (CHIL '20).

21. Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health. 2022 May;4(5):e351–8.

22. Glocker B, Jones C, Bernhardt M, Winzeck S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. EBioMedicine. 2023 Mar;89:104467.

23. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.

24. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351–63.

25. Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. BMJ Open. 2022 Sep 5;12(9):e061519.

26. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. JAMA Netw Open. 2022 Sep 1;5(9):e2233946.

27. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015 Jan 1;4(1):1.

28. 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. PLoS Med. 2021 Mar;18(3):e1003583.

29. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med. 2023 Jan;29(1):14–5.

30. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol [Internet]. 2018 Dec;18(1). Available from: http://dx.doi.org/10.1186/s12874-017-0468-4

31. Software as a medical device (SaMD): Key definitions [Internet]. International Medical Device Regulators Forum. [cited 2023 May 14]. Available from: https://www.imdrf.org/documents/software-medical-device-samd-key-definitions

32. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016 Dec 5;5(1):210.

33. Bitterman DS, Aerts HJWL, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health. 2020 Sep;2(9):e447–9.

34. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898.

35. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019 Oct 1;1(6):e271–97.

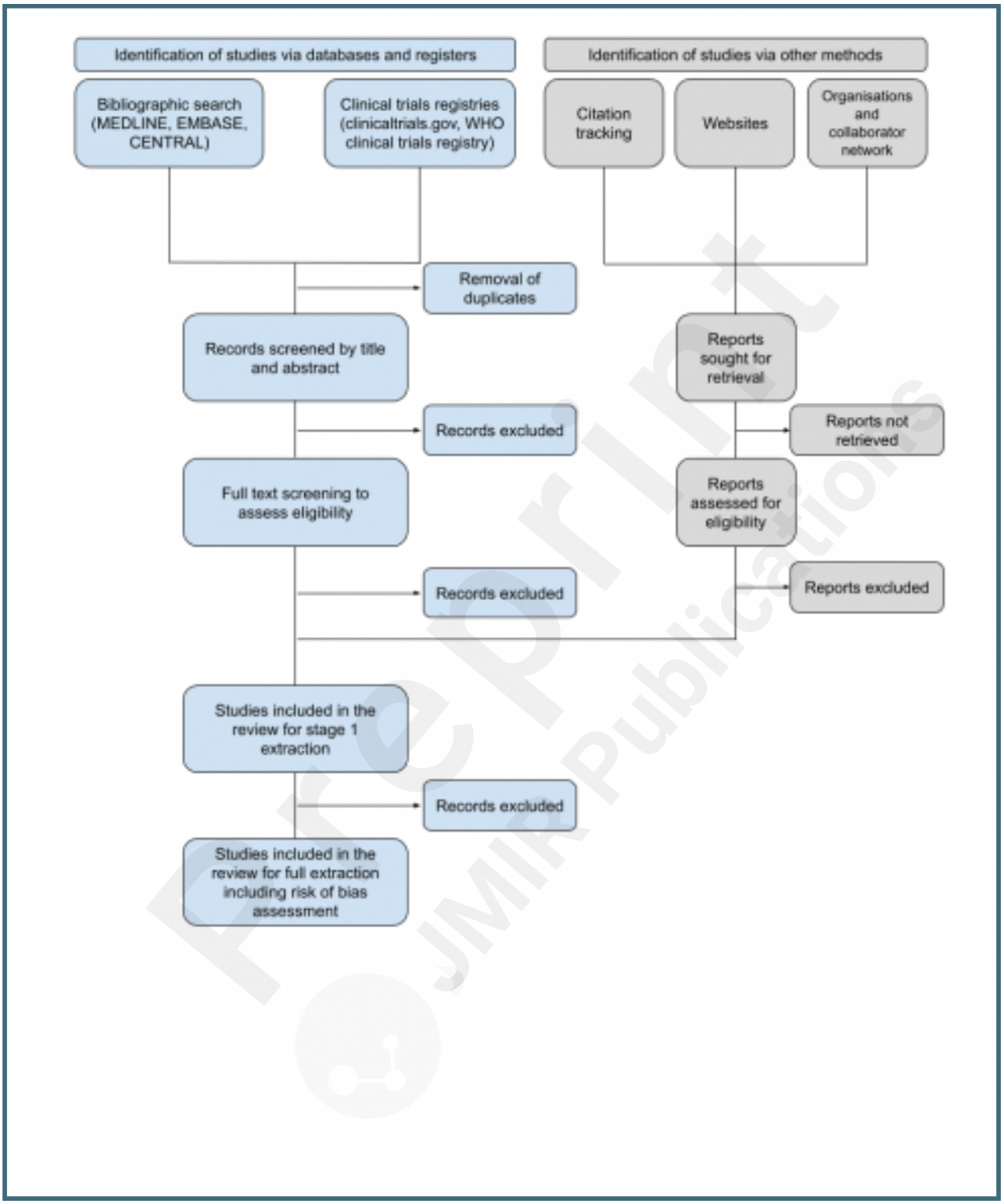
36. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020 Mar 25;368:m689.

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/d04568d73f5a406050ac7453da93c4a3.docx

Figures

PRISMA flow diagram outline to be populated during the systematic review process.



Multimedia Appendixes

Development of search strategy for MEDLINE and EMBASE.
URL: http://asset.jmir.pub/assets/cc6ff5aa7c909888d73fdbdc419e99a8.docx

Search Strategies.
URL: http://asset.jmir.pub/assets/b1baf4d193da92b53b1fe7deb94c8edf.docx