Graefe's Archive for Clinical and Experimental Ophthalmology
https://doi.org/10.1007/s00417-024-06553-3

REVIEW ARTICLE

Adherence of studies involving artificial intelligence in the analysis of ophthalmology electronic medical records to AI-specific items from the CONSORT-AI guideline: a systematic review

Niveditha Pattathil[1] · Tin-Suet Joan Lee[2] · Ryan S. Huang[2] · Eleanor R. Lena[2] · Tina Felfeli[3,4]

Received: 18 August 2023 / Revised: 9 June 2024 / Accepted: 12 June 2024

Abstract

Purpose In the context of ophthalmologic practice, there has been a rapid increase in the amount of data collected using electronic health records (EHR). Artificial intelligence (AI) offers a promising means of centralizing data collection and analysis, but to date, most AI algorithms have only been applied to analyzing image data in ophthalmologic practice. In this review we aimed to characterize the use of AI in the analysis of EHR, and to critically appraise the adherence of each included study to the CONSORT-AI reporting guideline.

Methods A comprehensive search of three relevant databases (MEDLINE, EMBASE, and Cochrane Library) from January 2010 to February 2023 was conducted. The included studies were evaluated for reporting quality based on the AI-specific items from the CONSORT-AI reporting guideline.

Results Of the 4,968 articles identified by our search, 89 studies met all inclusion criteria and were included in this review. Most of the studies utilized AI for ocular disease prediction ($n=41$, 46.1%), and diabetic retinopathy was the most studied ocular pathology ($n=19$, 21.3%). The overall mean CONSORT-AI score across the 14 measured items was 12.1 (range 8–14, median 12). Categories with the lowest adherence rates were: describing handling of poor quality data (48.3%), specifying participant inclusion and exclusion criteria (56.2%), and detailing access to the AI intervention or its code, including any restrictions (62.9%).

Conclusions In conclusion, we have identified that AI is prominently being used for disease prediction in ophthalmology clinics, however these algorithms are limited by their lack of generalizability and cross-center reproducibility. A standardized framework for AI reporting should be developed, to improve AI applications in the management of ocular disease and ophthalmology decision making.

Keywords CONSORT · Artificial intelligence · Disease prediction · Reporting guidelines · Electronic medical records

✉ Tina Felfeli
tina.felfeli@mail.utoronto.ca

Niveditha Pattathil
npattathil@qmed.ca

Tin-Suet Joan Lee
tsjoan.lee@mail.utoronto.ca

Ryan S. Huang
ry.huang@mail.utoronto.ca

Eleanor R. Lena
rebecca.lena@mail.utoronto.ca

[1] Queen's University School of Medicine, Kingston, Canada

[2] Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[3] Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, ON, Canada

[4] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

Published online: 02 July 2024

Key messages

What is known:

- Within the field of ophthalmology, there is increasing interest in exploring the potential of artificial intelligence (AI) in the analysis of electronic health data, however the reporting quality of the studies currently published is not known.

What this study adds:

- This systematic review aimed to evaluate the adherence of studies applying AI to the analysis of electronic health data within the field of ophthalmology and vision science to the AI-specific items from the CONSORT-AI reporting guideline.
- There is a large number of studies that have been published to date on the applications of AI to electronic health data within ophthalmology and vision science, however adherence to guideline items is suboptimal. CONSORT-AI items with the lowest levels of adherence were describing the handling of data of poor quality (48.3%), outlining inclusion and exclusion criteria for participants (56.2%), and describing how to access the AI intervention, including any restrictions (62.9%).
- High variability in study methodology and reporting frameworks limit the ability to conduct comprehensive cross-study comparison of AI applications to electronic health data.

Introduction

Electronic health records (EHRs) and electronic medical records (EMRs) have largely become the norm for storing patient data in medicine, enabling the storage of vast amounts of data that were previously unfeasible using paper records. These large volumes of patient data have the potential for secondary utility in bolstering clinical decision making and generating predictive disease stratification models [1]. Yet, due to the heterogeneity in the collected variables, information recording, and data input, it is challenging to use this data to derive meaningful clinical research conclusions [1]. Artificial intelligence (AI) provides a promising solution to analyze these vast amounts of data and has been successful in several fields. For example, within cardiology, AI has been used with EHR/EMR data to assist in the early detection of heart failure and predict the onset of congestive heart failure [2, 3]. In the realm of ophthalmology, AI and machine learning approaches have been utilized to predict the risk of complications post-cataract surgery, conduct risk assessment of diabetic retinopathy, and improve the diagnosis of conditions such as glaucoma and age-related macular degeneration [4–7].

Despite the promising potential of AI in ophthalmology, reporting standards across studies are not consistent, leading to a lack of clarity and transparency in the literature. As a result, there have been efforts to develop standardized guidelines for AI-specific study reporting. For example, the Consolidated Standards of Reporting Trials (CONSORT) statement provides the basic guidelines for reporting in randomized trials. The Consolidated Standards of Reporting Trials—Artificial Intelligence (CONSORT-AI) extension guideline was developed to provide guidance for reporting in randomized controlled trials (RCTs) specifically evaluating interventions with an AI component, ensuring that the results are transparent, reproducible, and comparable across the literature [8]. Herein we completed a critical analysis of all studies applying AI to data from electronic health and medical records within the field of ophthalmology and vision science. Furthermore, as there are no AI-specific generalized guidelines for non-RCT studies, we used the relevant AI elements from the CONSORT-AI checklist in order to critically appraise the adherence of each included study to the reporting guideline.

Methods

This is a systematic review of all studies applying AI to the analysis of patient data from EHRs/EMRs within the field of ophthalmology and vision science from January 1, 2010 to April 17, 2022 with a search update run on February 23, 2023. This review was conducted in accordance with the Preferred Reporting Items for a Systematic Review and Meta-analysis (PRISMA) guidelines. The protocol was prospectively registered in PROSPERO (registration number: CRD42022303128). A comprehensive search of the relevant databases MEDLINE, EMBASE, and Cochrane Library was done in consultation with an experienced librarian. A combination of keywords and Medical Subject Headings related to concepts of EHRs/EMRs, ophthalmology and AI were used to build the search strategy (Appendix 1).

Primary English studies that focused on human subjects published after January 2010 were eligible, including observational studies, case reports, and population studies. Articles were included if they provided outcomes regarding the value of AI in the analysis of patient EHR/EMR data, with or without imaging data, in any of the following ocular conditions: corneal disease, lens disease, glaucoma, retinal disease, scleral diseases, uveal diseases, choroid diseases, ocular neoplasms, strabismus, eyelid diseases, and ophthalmic emergencies. Studies were excluded if they solely focused on AI in the evaluation of ophthalmic imaging data, were in a language other than English, or were in the form of a review article, meta-analysis, conference abstract, editorial, short communication, guideline, or research letter. The authors of articles whose full-text was not available were contacted directly to request full-text versions.

Screening and data extraction

Two authors (T.J.L, R.S.H) independently conducted an initial title-abstract screening followed by full-text screening of all articles. All conflicts were resolved by consensus in consultation with a third author (E.R.L). Data from the final set of articles included in the review were extracted and recorded in a predetermined datasheet by two authors (T.J.L, R.S.H). Findings extracted from published reports included basic study characteristics, aspects of AI model construction, AI performance, and AI reporting domains. We collected information on baseline variables including: country, study design, purpose of study, disease outcome, sample size, reporting of socioeconomic status. Studies were evaluated for AI reporting based on 14 items from the AI-specific elements from the CONSORT-AI reporting guideline.

CONSORT-AI checklist

All included studies were scored independently by two authors (T.J.L, R.S.H) using 14 AI-specific items from the CONSORT-AI checklist. Each item was given equal weight, scoring 1 point each. The resulting mark was termed the 'CONSORT-AI score.' After initial scoring, any conflicts were resolved by consensus. The AI-specific items were from across the domains of: Title and Abstract, Background and Objectives, Methods, Results, and Other Information (Code Availability). The specific reporting requirements were: 1) indicating that the intervention involves AI and specifying the type of model; 2) stating the intended use of the AI intervention; 3) explaining the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users; 4) stating the inclusion and exclusion criteria at the level of participants; 5) stating the inclusion and exclusion criteria at the level of the input data; 6) describing how the AI intervention was integrated into the trial setting; 7) stating which version of the AI algorithm was used; 8) describing how the input data were acquired and selected for the AI intervention; 9) describing how poor quality or unavailable input data were assessed and handled; 10) specifying whether there human-AI interaction in the handling of the input data, and what level of expertise was required of users; 11) specifying the output of the AI intervention; 12) explaining how the AI intervention's outputs contributed to decision-making or other elements of clinical practice; 13) describing results of any analysis of performance errors and how errors were identified, where applicable; and 14) stating whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use. Inter-rater reliability was assessed using Cohen's kappa.

Risk of *Bias* assessment

Two independent reviewers (T.J.L, R.S.H) evaluated the potential for bias in the included studies using the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) tool. The assessment covered various bias domains for each study, including confounding factors, processes for selecting participants, classification of interventions, deviations from planned interventions, missing data, measurement of outcomes, and the selection of reported results.

Results

The search strategy yielded a total of 4,968 citations (Fig. 1). Following deduplication and screening, 89 studies met the inclusion and exclusion criteria. The characteristics of the included studies are summarized in Supplemental Table 1.

Studies were predominantly generated by the US ($n=26$, 32.5%) and China ($n=14$, 17.5%). The majority of studies retrieved patient data from either clinical records (33.75%) or health records databases (41.25%). Clinical records were defined as records collected from individual clinical practice sites, while health records databases were defined as large-scale repositories storing aggregated health information across multiple health systems or regions. The number of participants included in the AI algorithm ranged from 20 patients to 407,573 patients [9, 10]. The most commonly used AI modality was machine learning ($n=72$, 80.9%). Most studies used AI for ocular disease prediction ($n=41$, 46.1%), and diabetic retinopathy was the most studied ocular pathology ($n=19$, 21.3%).

The overall mean CONSORT-AI score across the 14 measured items was 12.1 (range 8–14, median 12). Following the initial round of scoring, there was conflict on 68 items (5.5%). The inter-rater concordance for CONSORT-AI scoring had a kappa score of 0.89. The compliance rates of the included studies to each of the individual AI-specific items from the CONSORT-AI reporting guideline are shown in Table 1 and organized as a heatmap in Supplemental Fig. 1. The categories

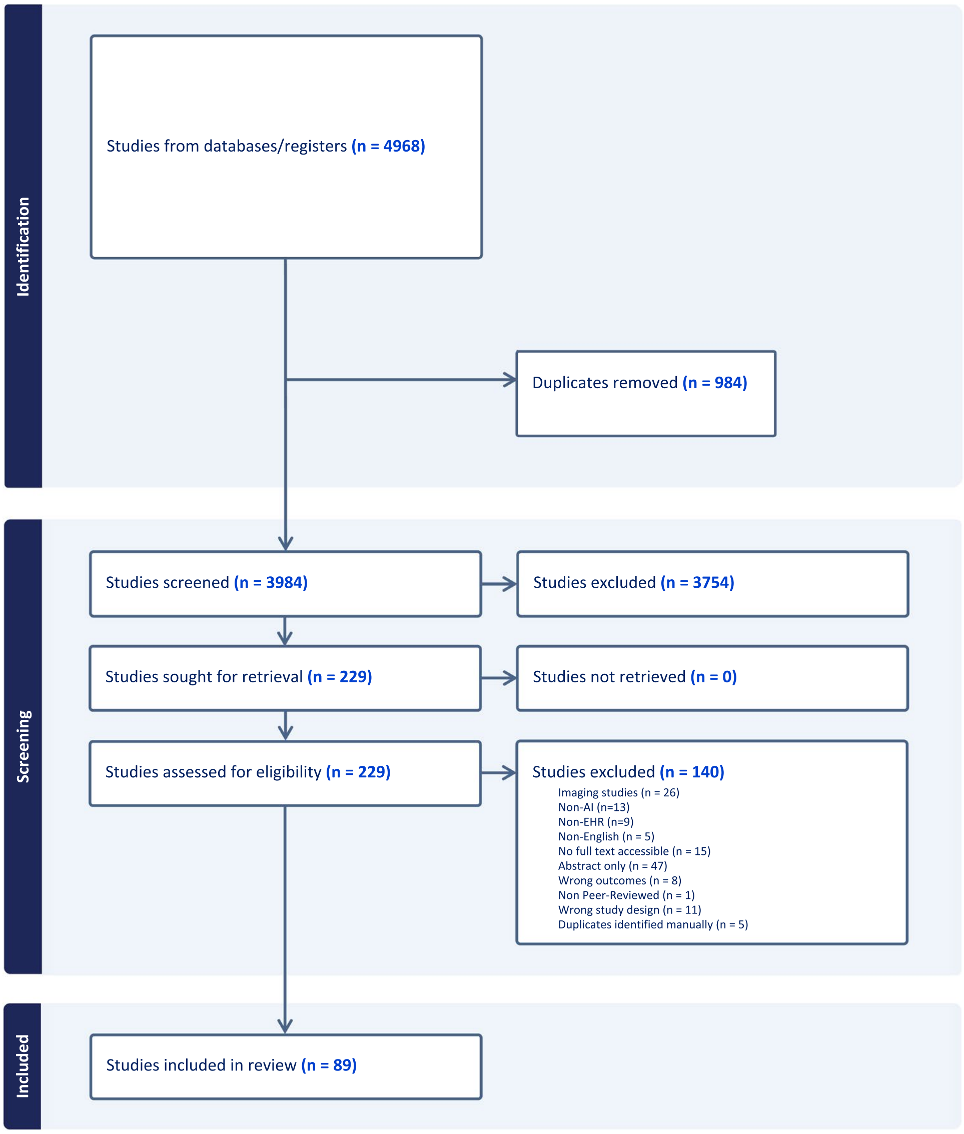

Identification
Studies from databases/registers (n = 4968)
Duplicates removed (n = 984)
Screening
Studies screened (n = 3984)
Studies excluded (n = 3754)
Studies sought for retrieval (n = 229)
Studies not retrieved (n = 0)
Studies assessed for eligibility (n = 229)
Studies excluded (n = 140)
Imaging studies (n = 26)
Non-AI (n=13)
Non-EHR (n=9)
Non-English (n = 5)
No full text accessible (n = 15)
Abstract only (n = 47)
Wrong outcomes (n = 8)
Non Peer-Reviewed (n = 1)
Wrong study design (n = 11)
Duplicates identified manually (n = 5)
Included
Studies included in review (n = 89)

Fig. 1 PRISMA flowchart diagram for study identification and selection

with the lowest adherence rates were: describing how poor quality or unavailable input data were assessed and handled (48.3%), reporting the inclusion and exclusion criteria of participants (56.2%), and providing further information as to whether and how the AI intervention and/or its code could be accessed, as well as any restrictions to access or re-use the modality (62.9%). The best performed categories were specifying the output of the AI intervention (100%), explaining how the AI intervention's outputs contributed to decision-making or other elements of clinical practice (100%), stating the intended use of the AI intervention within the trial in the title and/or abstract (98.9%), explaining the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (97.8%), describing how the input data were acquired and selected for the AI intervention (97.8%), stating which version of the AI algorithm was used (96.6%), describing the results of any analysis of performance errors and how errors were identified (96.6%), indicating that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specifying the type of model (93.3%), and stating the inclusion and exclusion criteria at the level of the input data (92.1%). Almost all studies reported their sources of funding, if applicable ($n = 80$, 89.9%). The majority of studies did not include socioeconomic status (SES) characteristics of patients within their study reporting ($n = 67$, 75.2%).

Based on ROBINS-I risk of bias assessment tool, risk of bias was "low" for 49 (55%) and associated with "some concerns" for 40 (45%) studies (Fig. 2). The majority of concerns with regards to risk of bias domains were identified to be "confounding," "selection of participants in the study," and "missing data."

Table 1 Compliance of included studies to CONSORT-AI

Section	Criteria	Number of Articles	Percent of Articles
Title and Abstract	(i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model	83	93.3%
	(ii) State the intended use of the AI intervention within the trial in the title and/or abstract	88	98.9%
Background and objectives	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)	87	97.8%
Methods (Participants)	State the inclusion and exclusion criteria at the level of participants	50	56.2%
	State the inclusion and exclusion criteria at the level of the input data	82	92.1%
	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements	76	85.4%
Methods (Interventions)	State which version of the AI algorithm was used	86	96.6%
	Describe how the input data were acquired and selected for the AI intervention	87	97.8%
	Describe how poor quality or unavailable input data were assessed and handled	43	48.3%
	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users	75	84.3%
	Specify the output of the AI intervention	89	100%
	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice	89	100%
Results	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, explain why not	86	96.6%
Funding	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use	56	62.9%
Total	Y/Total Domains	1077	86.4%

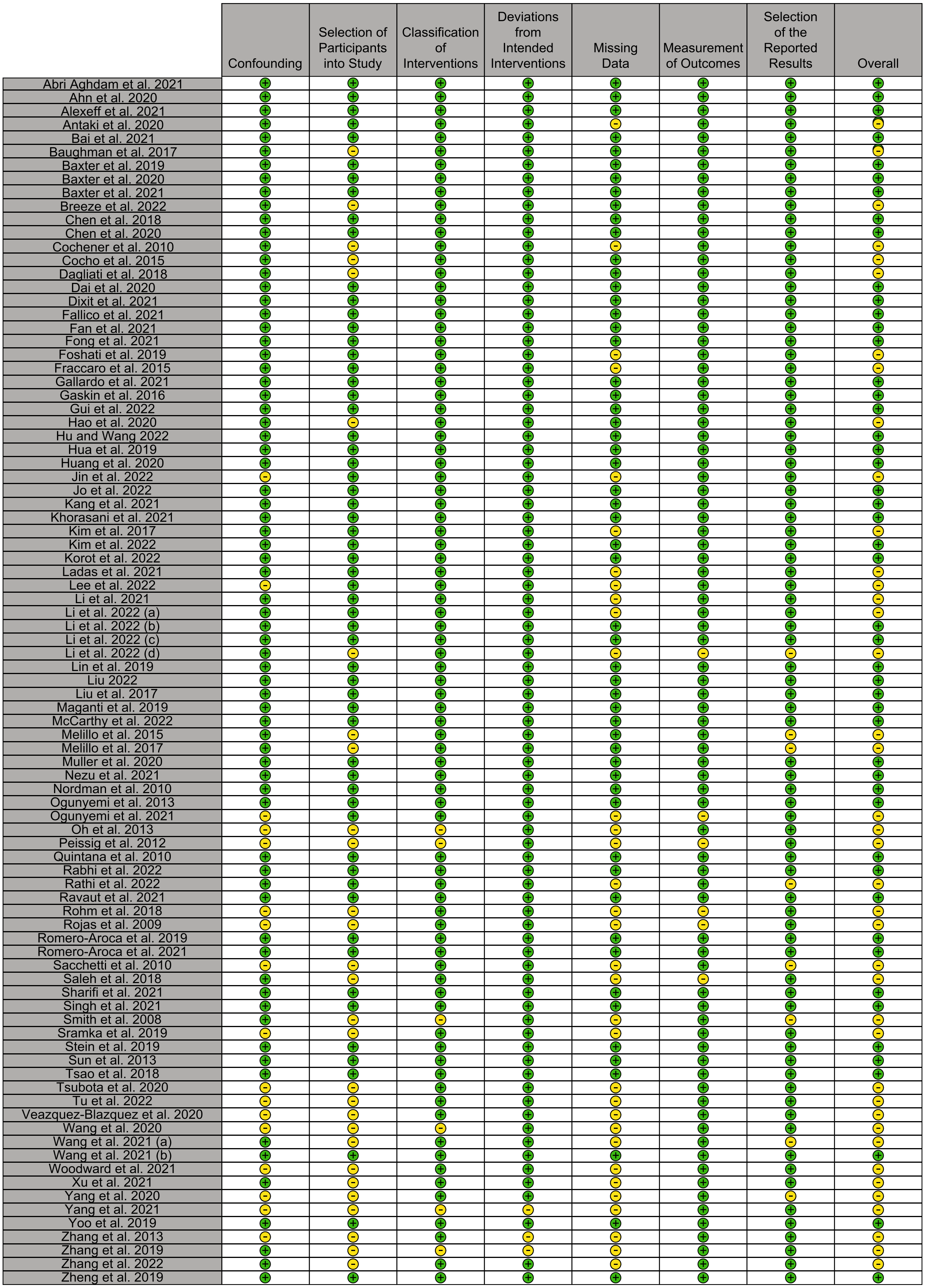
Y/Total Domains represents the number of affirmative responses (yes) across all criteria domains. There are a total of 1246 domains evaluated across the 14 criteria in 89 studies

Discussion

With this review, we aimed to assess the reporting quality of studies utilizing AI and EMRs within ophthalmology by examining their adherence to 14 AI-specific items from the CONSORT-AI reporting standards for studies involving AI. Our study found a total of 89 studies that utilized AI with EMRs in ophthalmology. The mean CONSORT-AI score of the articles was 12.1/14 (range 8–14, median 12). Out of the 89 articles total in our review, 14 (15.7%) of the articles received a score of 100%.

To the best of our knowledge, our review is the first comprehensive study to evaluate the adherence of articles utilizing AI with EMRs in ophthalmology using AI-specific items from the CONSORT-AI reporting guideline. The adherence of the studies we examined was generally high for the 14 AI-specific items assessed from the CONSORT-AI reporting guideline, with an average adherence score of 86.4% (range 48.3–100%). However, the criteria with the lowest adherence were describing how poor quality or unavailable input data were assessed and handled (48.3%), reporting the inclusion and exclusion criteria of participants (56.2%), and providing information as to whether and how the AI intervention and/or its code could be accessed, as well as any restrictions to access or re-use the modality (62.9%). A similar recent study on the adherence of randomized controlled trials using AI in ophthalmology to the CONSORT-AI checklist found suboptimal reporting across certain domains, with an average adherence of 53% (range 37%–78%) for the included articles [11].

As research utilizing artificial intelligence continues to expand rapidly, tools for the evaluation of research output are necessary in order to maintain a high quality of reporting standards amongst publications. Reporting guidelines help to ensure scientific validity, clarity in the arrangement of results presented, greater reproducibility, and adherence to a consistent and ethical set of standards amongst researchers utilizing AI. This push towards standardization in reporting is already reflected in the current literature with the recent production of several guidelines for the reporting and quality assessment of AI studies. Currently, there are several guidelines that have been published for authors to reference when publishing research within the field of AI. For randomized trials, CONSORT-AI and SPIRIT-AI are the corresponding AI extensions for CONSORT (Consolidated Standards of Reporting Trials) and SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) [8, 12]. There are some options that have been developed for different types



	Confounding	Selection of Participants into Study	Classification of Interventions	Deviations from Intended Interventions	Missing Data	Measurement of Outcomes	Selection of the Reported Results	Overall
Abri Aghdam et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Ahn et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Alexeff et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Antaki et al. 2020	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Bai et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Baughman et al. 2017	⊕	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Baxter et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Baxter et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Baxter et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Breeze et al. 2022	⊕	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Chen et al. 2018	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Chen et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Cochener et al. 2010	⊕	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Cocho et al. 2015	⊕	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Dagliati et al. 2018	⊕	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Dai et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Dixit et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Fallico et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Fan et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Fong et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Foshati et al. 2019	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Fraccaro et al. 2015	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Gallardo et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Gaskin et al. 2016	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Gui et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Hao et al. 2020	⊕	⊖	⊕	⊕	⊕	⊕	⊕	⊖
Hu and Wang 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Hua et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Huang et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Jin et al. 2022	⊖	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Jo et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Kang et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Khorasani et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Kim et al. 2017	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Kim et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Korot et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Ladas et al. 2021	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Lee et al. 2022	⊖	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Li et al. 2021	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Li et al. 2022 (a)	⊕	⊕	⊕	⊕	⊖	⊕	⊕	⊖
Li et al. 2022 (b)	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Li et al. 2022 (c)	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Li et al. 2022 (d)	⊕	⊖	⊕	⊕	⊖	⊖	⊖	⊖
Lin et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Liu 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Liu et al. 2017	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Maganti et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
McCarthy et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Melillo et al. 2015	⊕	⊖	⊕	⊕	⊕	⊕	⊖	⊖
Melillo et al. 2017	⊕	⊖	⊕	⊕	⊕	⊕	⊖	⊖
Muller et al. 2020	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Nezu et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Nordman et al. 2010	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Ogunyemi et al. 2013	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Ogunyemi et al. 2021	⊖	⊕	⊕	⊕	⊖	⊖	⊕	⊖
Oh et al. 2013	⊖	⊖	⊖	⊕	⊖	⊕	⊕	⊖
Peissig et al. 2012	⊖	⊖	⊖	⊕	⊖	⊖	⊕	⊖
Quintana et al. 2010	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Rabhi et al. 2022	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Rathi et al. 2022	⊕	⊕	⊕	⊕	⊖	⊕	⊖	⊖
Ravaut et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Rohm et al. 2018	⊖	⊖	⊕	⊕	⊖	⊖	⊕	⊖
Rojas et al. 2009	⊖	⊖	⊕	⊕	⊖	⊖	⊕	⊖
Romero-Aroca et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Romero-Aroca et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Sacchetti et al. 2010	⊖	⊖	⊕	⊕	⊖	⊕	⊖	⊖
Saleh et al. 2018	⊕	⊖	⊕	⊕	⊖	⊖	⊕	⊖
Sharifi et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Singh et al. 2021	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Smith et al. 2008	⊕	⊖	⊖	⊕	⊖	⊕	⊖	⊖
Sramka et al. 2019	⊖	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Stein et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Sun et al. 2013	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Tsao et al. 2018	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Tsubota et al. 2020	⊖	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Tu et al. 2022	⊖	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Veazquez-Blazquez et al. 2020	⊖	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Wang et al. 2020	⊖	⊖	⊖	⊕	⊖	⊕	⊕	⊖
Wang et al. 2021 (a)	⊕	⊖	⊕	⊕	⊖	⊕	⊖	⊖
Wang et al. 2021 (b)	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Woodward et al. 2021	⊖	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Xu et al. 2021	⊕	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Yang et al. 2020	⊖	⊖	⊕	⊕	⊖	⊕	⊖	⊖
Yang et al. 2021	⊖	⊖	⊖	⊖	⊖	⊕	⊕	⊖
Yoo et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Zhang et al. 2013	⊖	⊖	⊕	⊖	⊖	⊕	⊕	⊖
Zhang et al. 2019	⊕	⊖	⊖	⊖	⊖	⊕	⊕	⊖
Zhang et al. 2022	⊕	⊖	⊕	⊕	⊖	⊕	⊕	⊖
Zheng et al. 2019	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕

◂**Fig. 2** The ROBINS-I traffic light plots of the domain-level judgements for each individual result of publication quality and are formatted according to the risk-of-bias assessment tool used to perform the assessments. Green indicates "low risk" of bias and yellow indicates "some concerns."

of non-randomized studies. For example, for diagnostic accuracy studies, STARD-AI is the AI-specific version of the Standards for Reporting of Diagnostic Accuracy Study (STARD) [13]. For prediction model studies on diagnosis and prognosis, there are three upcoming guidelines in development, called QUADAS-AI, TRIPOD-AI and PROBAST-AI [14, 15]. They are the AI versions of the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies), TRIPOD (Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis) statement and the PROBAST (Prediction model Risk Of Bias ASsessment Tool) [14, 15]. Once complete, these checklists will provide guidelines for reporting standards as well as risk of bias assessment, which will be very useful for meta-analyses comparing between various AI studies.

With regards to performance assessment of AI-specific models, MI-CLAIM (Minimum Information about Clinical Artificial Intelligence Modelling) focuses on technical reproducibility of clinical AI modeling research, and MINIMAR (MINimum Information for Medical AI Reporting) provides guidance on proper data source usage and model evaluation [16, 17]. For the evaluation of early-stage AI decision support systems, DECIDE-AI is a reporting guideline for the evaluation of these clinical evaluations, helping to facilitate critical appraisal of these studies and replicability of their findings [19]. There are also guidelines specific to certain topics of research within AI, such as CLAIM (Checklist for Artificial Intelligence in Medical Imaging) which is for studies applying machine learning to medical imaging, and the Radiomics Quality Score (RQS) that is specific to publications on radiomics [20, 21]. There are also initiatives targeted towards AI model development, such as FUTURE-AI, which is a checklist for use within the conceptualization and development stage. FUTURE-AI is based on six central principles (Fairness, Universality, Traceability, Usability, Robustness and Explainability (FUTURE)) that focuses on assessing AI model optimization for real-word practice [18].

Therefore, depending on the type of non-randomized study, there are several potential options for reporting guidelines that could serve as a valuable reference for authors when developing their manuscripts. However, a consolidated generalized checklist applicable for all non-randomized studies would be ideal to provide a standardized framework for AI reporting, and help facilitate easier comparison between different types of non-randomized studies. In the interim, non-RCT studies can utilize the previously mentioned guidelines or even the CONSORT-AI framework as a reference by which to ensure that their reporting includes all relevant details, allowing for greater translation into clinical settings and standardization in the way results are reported between different AI models.

Our study completed a comprehensive search of the literature in order to identify all eligible articles within the field of ophthalmology that have applied AI to the analysis of EHR/EMRs. We were also able to assess the presence of certain characteristics, specifically the purpose and type of the AI model, the ophthalmological disease focused on, the data source used, study design type, country of origin, and whether baseline SES and funding source were reported. These discrepancies highlight a need for standardization of AI reporting guidelines which will enable better reproducibility of AI methodologies and allow for generalizability of results across various ophthalmologic centers. Lastly, certain restrictions in our inclusion criteria including English-language publications and other forms of secondary literature may have limited the identification of additional studies and perspectives on the topic.

Conclusion

Artificial intelligence (AI) offers considerable promise in leveraging large, heterogeneous patient health data sets to inform clinical practice in the management of medical conditions and disease. The digitization and electronic storage of medical information has provided a favourable setting for this application of AI and machine learning. The application of AI techniques in ophthalmology continues to rapidly progress, with new initiatives being developed in a wide variety of areas within the field. However, there is still a lack of standardization in reporting the results of these studies, which can make it difficult to compare and evaluate different AI models. The CONSORT-AI framework holds promise as an effective guideline for the transparent and comprehensive reporting of AI studies, by helping to standardize reporting across key aspects such as the study design, participant characteristics, interventions, outcomes, and statistical analysis. By adhering to the AI-specific reporting guidelines, researchers can improve the clarity and completeness of their reporting, allowing readers to better assess the quality and validity of their study. Standardized and transparent reporting of AI studies in ophthalmology will ultimately aid in application of AI for enhanced diagnosis and management of ocular conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00417-024-06553-3.

Funding No funding was received for this research.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Conflict of interest No conflicts of interest.

References

1. Lin W-C, Chen JS, Chiang MF, Hribar MR (2020) Applications of artificial intelligence to electronic health record data in ophthalmology. Transl Vision Sci Technol 9:13. https://doi.org/10.1167/tvst.9.2.13
2. Cheng Y, Wang F, Zhang P, Hu J (2016) Risk prediction with electronic health records: a deep learning approach. In: Venkatasubramanian SC, Meira W (eds) Proceedings of the 2016 SIAM international conference on data mining 2016 Jun 30. Society for Industrial and Applied Mathematics Publications, p 432–440. https://doi.org/10.1137/1.9781611974348.49
3. Choi E, Schuetz A, Stewart WF, Sun J (2017) Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 24:361–370. https://doi.org/10.1093/jamia/ocw112
4. Gaskin GL, Pershing S, Cole TS, Shah NH (2016) Predictive modeling of risk factors and complications of cataract surgery. Eur J Ophthalmol 26:328–337. https://doi.org/10.5301/ejo.5000706
5. Saleh E, Błaszczyński J, Moreno A et al (2018) Learning ensemble classifiers for diabetic retinopathy assessment. Artif Intell Med 85:50–63. https://doi.org/10.1016/j.artmed.2017.09.006
6. Chaganti S, Nabar KP, Nelson KM et al (2017) Phenotype analysis of early risk factors from electronic medical records improves image-derived diagnostic classifiers for optic nerve pathology. Proc SPIE Int Soc Opt Eng 10138:101380F. https://doi.org/10.1117/12.2254618
7. Fraccaro P, Nicolo M, Bonetto M et al (2015) Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. BMC Ophthalmol 15:1–9. https://doi.org/10.1186/1471-2415-15-10
8. Schulz KF, Altman DG, Moher D (2011) CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomized trials. Ann Intern Med 154:291–292. https://doi.org/10.7326/0003-4819-154-4-201102150-00017
9. Cocho L, Fernández I, Calonge M et al (2015) Gene expression-based predictive models of graft versus host disease-associated dry eye. Invest Ophthalmol Vis Sci 56:4570–4581. https://doi.org/10.1167/iovs.15-16736
10. Breeze F, Hossain RR, Mayo M, McKelvie J (2023) Predicting ophthalmic clinic non-attendance using machine learning: Development and validation of models using nationwide data. Clin Exp Ophthalmol 51(8):764–774. https://doi.org/10.1111/ceo.14310
11. Pattathil N, Zhao JZL, Sam-Oyerinde O, Felfeli T (2023) Adherence of randomised controlled trials using artificial intelligence in ophthalmology to CONSORT-AI guidelines: a systematic review and critical appraisal. BMJ Health Care Inform 30:e100757. https://doi.org/10.1136/bmjhci-2023-100757
12. Lekadir K, Osuala R, Gallin C et al (2021) FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv preprint arXiv:2109.09658
13. Sounderajah V, Ashrafian H, Golub RM et al (2021) Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 11:e047709
14. Sounderajah V, Ashrafian H, Rose S et al (2021) A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 27:1663–1665
15. Collins GS, Dhiman P, Navarro CLA et al (2021) Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 11:e048008
16. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH (2020) MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 27:2011–2015. https://doi.org/10.1093/jamia/ocaa088
17. Norgeot B, Quer G, Beaulieu-Jones BK et al (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 26:1320–1324. https://doi.org/10.1038/s41591-020-1041-y
18. Lekadir K, Osuala R, Gallin C et al (2021) FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging
19. Vasey B, Nagendran M, Campbell B et al (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 28:924–933. https://doi.org/10.1038/s41591-022-01772-9

20 Mongan J, Moy L, Kahn CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell 2:200029. https://doi.org/10.1148/ryai.2020200029

21. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.