Artificial Intelligence in Medicine: A Systematic Review of Guidelines for the Reporting and Interpretation of Studies

Zsombor Zrubka[1], Levente Kovács[2], Hossein Motahari Nezhad[3], János Czere[4], László Gulácsi[1], Márta Péntek[1]

[1] Health Economics Research Center, University Research and Innovation Center; Doctoral School of Innovation Management, Óbuda University, Bécsi út 96/b 1034 Budapest, Hungary; zrubka.zsombor@uni-obuda.hu; gulacsi@uni-obuda.hu; pentek.marta@uni-obuda.hu

[2] Physiological Controls Research Center, University Research and Innovation Center, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, kovacs@uni-obuda.hu

[3] Health Economics Research Center, University Research and Innovation Center, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, motahari.hossein@uni-obuda.hu

[4] Doctoral School of Innovation Management, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, czere.janos@phd.uni-obuda.hu

Abstract: Reporting guidelines, developed for medical artificial intelligence (AI) studies, are structured tools that address general and/or AI-specific methodological and reporting issues. We aimed to systematically review published medical AI reporting guidelines and checklists, and evaluate aspects that can support the choice of the tool, in a particular research context. We searched PubMed, Scopus, and Web of Science thru February 2023, as well as, Citations and Google. From 821 records, and additional sources, 24 guidelines were identified (4 narrative guidelines, 7 general reporting checklists, 4 study design specific checklists, and 9 clinical area specific checklists). 13 studies reported the guideline development methods, 10 guidelines were registered in the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network. In 224 sections, the guidelines contained 704 items in total. The number of items per checklist varied between 10 and 66. The guidelines' structure and level of detail varied significantly which makes difficult for researchers to follow how detailed and standardized a medical AI study design and report should be. The robustness of development process and support from the literature suggests that the AI extension of checklist for randomized controlled trials (CONSORT-AI guideline) as the most established tool. Such AI extensions of clinical study guidelines may not cover all the application fields of AI in medicine. In certain research contexts, an established checklist for main clinical study types, and a general AI-based checklist may be used in parallel to provide most useful guidance in designing, writing and interpreting medical AI studies.

Keywords: artificial intelligence; machine learning; medical studies; guideline; checklist; usability; explainability

1 Introduction

Artificial intelligence (AI) and machine learning (ML), unless specified otherwise, hereunder collectively referred to as AI, have been researched in academia since the mid-20th Century. Recent advances in computing power, data storage, research methodology, and skilled human resources have accelerated their application of AI in medicine, with the potential to transform healthcare and the life sciences industries [1] [2].

The use of AI is rapidly developing in many clinical areas including endocrinology [3], cardiology [4], neurology, oncology, haematology, nephrology, gastroenterology, orthopaedics, and rheumatology, clinical approaches such as medical imaging [5] [6], precision medicine, genomics, and telemedicine, and care components such as triage, diagnosis, prognosis, monitoring, and treatment [7, 8]. The growing importance of AI in medicine is well illustrated by the fact that the number of approved AI-based devices by the US Food and Drug Administration (FDA) increased from 9 in 2015 to 77 in 2019 with 24 new devices approved in the first quarter of 2020. Furthermore, a total of 240 AI-based medical devices were approved in Europe between 2015 and 2020 [9].

With the rapid advancement of AI-based medical technologies, there has been an increasing need for evidence on their risks and benefits, based on high-quality clinical trials. Methods for observational and interventional studies (clinical trials) in clinical epidemiology have been established and used in evidence-based medicine in the past decades [10-12]. In parallel, significant improvements have been made to standardize how clinical studies are designed and reported [13]. The impact of reporting quality should not be underestimated. First, good quality data can only be reported from well-designed, conducted, and well-documented studies. Hence, reporting requirements may positively influence trial design [13]. Second, poor reporting of a high-quality clinical trial may hamper its clinical usability, and vice versa, inadequate, or incomplete reporting of a poor-quality clinical trial may obscure its weaknesses and can lead to incorrect medical decisions. Third, systematic literature reviews and meta-analyses, which represent the highest level of medical evidence, require well-structured, and transparently reported data [14].

The need for standardization has driven the development of reporting guidelines and checklists. While we acknowledge the difference, we will refer to reporting guidelines and checklists interchangeably in this paper. (Guidelines without checklists will be denoted as narrative guidelines.) The compliance with relevant reporting guidelines is a criterion for publication in many medical journals, which

facilitates their uptake among researchers [13]. For instance, the Consolidated Standards of Reporting Trials (CONSORT) checklist for randomized controlled trials has been simultaneously published in 9 medical journals and translated to 12 languages [15]. It includes 25 items, indicating the information that should be provided about each section of the study, with 27 extension versions for specific clinical or methodological applications. Reporting guidelines in the CONSORT family follow a standardized structure and undergo rigorous development process [16], recommended by the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network, 'An international initiative that seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines' [17].

The development and therapeutic applications of AI algorithms presented new complexities and potential sources of bias that had not been addressed in former reporting guidelines [18]. Although some improvement over time has been observed in some areas [19], several studies reported alarming results about the methodological quality and reporting of medical AI studies. Despite the comparable diagnostic accuracy of deep learning algorithms and healthcare professionals in the interpretation of medical imaging, the absence of external validation and inadequate reporting standards undermine the credibility of these results [20]. Most medical imaging studies reported between 2010-2019 that compared the diagnostic accuracy of deep-learning algorithms with human experts had high risk of bias, deviated from reporting standards, and lacked data and code sharing [21]. External validation was performed in only 6% of 516 radiological-AI studies reported in 2018, and none of the studies featured all important design elements of clinical AI validation [22]. Likewise, none of the papers from 2015-2018 on the use of diagnostic ML models used a reporting guideline, and most of them lacked adequate details on participants [23].

As a response, research teams, international organizations, and publishers have established reporting guidelines for medical AI studies to ensure that results are replicable, transparent, and provide sufficient information for inclusion in future evidence syntheses [18, 24, 25]. AI reporting guidelines are diverse in terms of their target audiences, aims, scope, structure, and the rigorousness of their development process. Therefore, with all the good intentions, AI reporting guidelines present a new challenge to authors and physicians alike. Which one to follow? What represents best practice in reporting? Have I missed a new reporting guideline that fits better my research context?

Given the rapid development in the field [26-30], the present research aims to provide a comprehensive systematic literature review on medical AI reporting guidelines. Specifically, we aim to analyze the goals, target audiences, development process, focus area, structure, and usage of guidelines. With our review we aim to aid researchers in the choice of the reporting guideline that represents best practice in their research context in a particular clinical area or research setting.

2 Methods

2.1 Search Strategy

We searched PubMed, Scopus, and the Web of Science databases for potentially relevant publications from the inception to September 28, 2022. The search was updated on February 12, 2023. The detailed search syntaxes are provided in the Supplementary File (https://osf.io/bz9f7/).

The reference list of the included studies was also reviewed to find any other eligible documents. Furthermore, in February 2023, a literature search was conducted using the Google search engine to identify additional pertinent studies using the following keywords: checklist, guideline, reporting, standard, recommendations, artificial intelligence, machine learning, deep learning, medical, medicine, health, clinician, doctor, and healthcare.

2.2 Selection Criteria

Articles were included if they presented an original reporting guideline applicable for medical AI research. Both narrative guidelines, and detailed checklists were considered. Peer reviewed publications in any language were considered as eligible without restriction on publication date or publication type. Studies were excluded if they published a guideline development research protocol or the use of reporting guidelines for the assessment of medical AI research.

2.3 Study Selection

Records were deduplicated and imported into an Excel spreadsheet. Two reviewers independently screened the titles and abstracts of the publications against the inclusion criteria. Potentially eligible records were subjected to full-text screening by two independent reviewers. In both stages, disagreements between reviewers were resolved by consensus.

2.4 Data Extraction and Synthesis

To characterize guidelines, we extracted their name, the title and year of publication, the journal, the first author's name and country of affiliation. We also recorded the clinical area or study design in focus. The target audience of the guidelines was divided into three distinct groups: A) application developers, B) clinicians and model users, and C) authors, reviewers, and editors. Furthermore, we recorded if details of the development process were reported and whether the

guideline was registered on the EQUATOR Network website. If reported, we extracted seven key components of guideline development from EQUATOR framework. These involved a literature review, a Delphi consensus survey, expert consensus meeting, pilot testing, obtaining funding, a policy for periodic updates, and endorsement by a journal or professional society [16, 17, 31]. Furthermore, as a proxy indicator of the use of the guidelines in academic research, the number of citations for each was retrieved from the Google Scholar database on July 18, 2023.

When analyzing the structure of guidelines, we extracted all elements of their structure that is one level above the reporting items. We compared the overall structure and level of detail, while a detailed comparative analysis of their content at the reporting item level was beyond the scope of this research. We counted the total number of reporting items presented in the guidelines. We considered the smallest units of text in the guideline structure (e.g., phrase, sentence, or paragraph) as reporting items. Although some guidelines described multiple reporting elements within an item, we considered them as part of the respective item without further breakdown.

We synthesized results using simple descriptive statistics. The associations between variables were explored via appropriate bivariate tests after checking the distributional assumptions. For reporting, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [32].

3 Results

3.1 Search Results

The search in PubMed, Scopus, and Web of Science yielded a total of 1397 hits. After the elimination of 567 duplicates, 821 records were screened by title and abstract. We checked 193 full-text publications against the predefined eligibility criteria, resulting in the inclusion of 20 studies. In addition, four studies were identified through the search of the reference lists of included articles and the complementary Google search. In total, 24 studies were included for further investigation. Details of the article screening and selection are shown in Fig. 1. The list of excluded studies is provided in the Supplementary File (https://osf.io/bz9f7/).

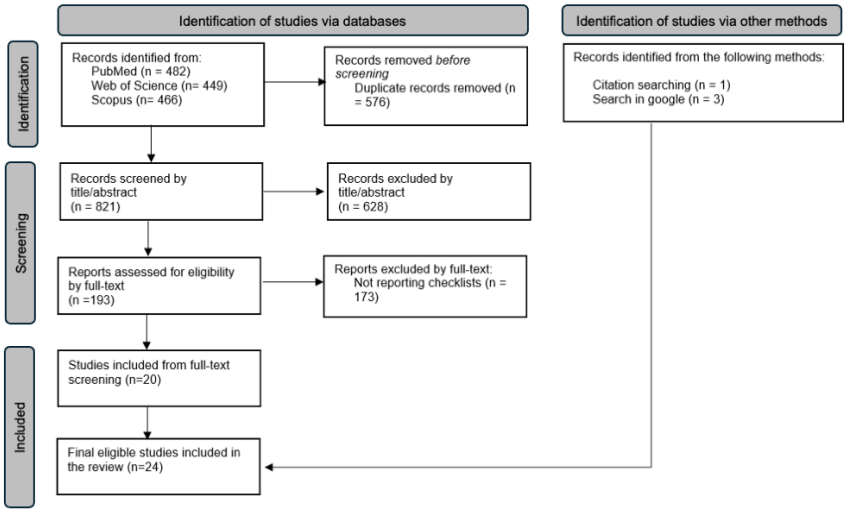

Figure 1
PRISMA flowchart of article selection and screening process

3.2 Main Characteristics of the Publications

Table 1 summarizes the 24 articles involved for further analysis. The first guideline, CHARMS [33], focusing on the appraisal of systematic reviews of predictive modelling studies was published in 2014, followed by a guideline on the reporting of ML predictive models by Luo et al [34] from 2016. Most guidelines were subsequently published in 2020 (n=9, 38%) and 2021 (n=9, 38%), followed by 2022 (n=3, 13%) and one guideline in 2023.

The first authors' affiliation was mainly from the United States (n=9, 38%), followed by the United Kingdom (n=4, 17%), Australia (n=2, 8%) and Canada (n=2, 8%). Other guidelines were published by authors with affiliation from France, Italy, the Netherlands, Sweden, Spain, Germany, and Switzerland. The 24 guidelines were published in 22 journals, with Nature Medicine (n=3, 13%) being the most common.

The CHARMS guideline [33] received the highest number of citations on Google Scholar (N=1080), followed by Luo et al [34] which was cited 571 times. Four guidelines, CONSORT-AI [35], SPIRIT-AI [36], CLAIM [37], and MI-CLAIM [38], attracted more citations than the average (N=178). They were cited 523, 444, 442, and 212 times, respectively. All highly cited guidelines included a point-by-point checklist.

Table 1

Summary of the 24 included reporting guidelines for medical AI studies

First author (Year) / Country of affiliation	Name of the guideline / N of reporting items	Focus of the guideline	Journal	Purpose of the guideline	Google Scholar citations[a]
Narrative guidelines					
Buvat (2021) / France [39]	T.R.U.E. / 4	Nuclear medicine	The Journal of Nuclear Medicine	To aid the identification of studies reporting ground-breaking developments in AI-based research in nuclear medicine.	15
Stevens (2020) / US [40]	NS[b] / 19	ML in clinical research	Circulation: Cardiovascular Quality and Outcomes	A guideline for transparent and systematic presentation of outcomes from ML analyses, addressed primarily for clinical researchers. Designed to supplement current clinical reporting requirements.	90
Faes (2020) / UK [41]	NS / 8	ML clinical studies	Translational Vision Science & Technology	Improve the quality of research on the therapeutic use of ML by equipping clinicians and researchers with the tools they need to conduct their own rigorous assessments.	111
Bates (2020) / US [42]	NS / 8	Clinical research of AI-based interventions	Annals of Internal Medicine	Suggestions for reporting standards to enable the assessment of the incremental benefits of ML and AI, and remove barriers from their clinical adoption.	53
General reporting checklists					
Al-Zaiti (2022) / US [26]	ROBUST-ML / 30	ML in clinical studies	European Heart Journal - Digital Health	Increase physicians' understanding of ML by equipping them with the information and tools required to comprehend and evaluate clinical research focusing on ML.	10
Cabitza (2021) / Italy [24]	NS / 55	ML in clinical studies	International Journal of Medical Informatics	To analyse the scientific rigor of a medical ML contribution and the reliability of its findings qualitatively.	93
Olczak (2021) / Sweden [43]	CAIR / 40	Clinical AI research	Acta Orthopaedica	Clinical reporting guidelines for artificial intelligence and ML; guidance on selecting appropriate outcome indicators.	33
Scott (2021) / Australia [44]	NS / 12	ML algorithm in healthcare	BMJ Health Care Informatics	To evaluate the clinical usefulness of ML technologies in healthcare.	58
Hernandez-Boussard (2020) / US [45]	MINIMAR / 21	AI in healthcare	Journal of the American Medical Informatics Association	To facilitate the diffusion of algorithms across healthcare systems, enable transparency to address any biases and unintended effects, and encourage the use of secondary resources through promoting external validation and encouraging the use of secondary resources.	145
Norgeot (2020) / US [38]	MI-CLAIM / 19	Clinical AI modelling	Nature Medicine	To propose a baseline for reporting to guarantee transparency and practicality in the use of AI in healthcare.	212
Luo (2016) / Australia [34]	NS / 52	ML predictive models in biomedical research	Journal of Medical Internet Research	To develop guidelines for the application of prediction models based on ML in healthcare settings.	571
Checklists for specific study designs					
Liu (2020) / UK [35]	CONSORT-AI / 49	Randomized clinical trials involving interventions with AI component	BMJ Nature Medicine Lancet Digital Health	To establish a standard for the reporting of clinical trials employing artificial intelligence-based therapies.	523
Rivera (2020) / UK [36]	SPIRIT-AI / 66	Clinical study protocol involving AI method	BMJ Nature Medicine Lancet Digital Health	To enhance the comprehensiveness of clinical trial protocol documentation.	444
Vasey (2022) / UK [29]	DECIDE-AI / 38	Early-stage clinical evaluation of AI-driven decision support systems	BMJ Nature Medicine	To facilitate the evaluation of research and the reproducibility of their results in healthcare studies using AI-based decision support systems.	96

Moons (2014) / Netherlands [33]	CHARMS / 35	Systematic reviews of prediction modelling studies	PLOS Medicine	To assist with the formulation of a review question and evaluation of all forms of primary prediction modeling studies for systematic reviews.	1080
Checklists for specific clinical areas					
Daneshjou (2021) / US [46]	CLEAR Derm / 25	Image-based AI in dermatology	JAMA Dermatology	To synthesize the minimal current material to serve as a guide for dermatological AI developers and reviewers.	42
Haller (2022) / Switzerland [28]	R-AI-DIOLOGY / 21	AI in clinical neuroradiology	Neuroradiology	To assist neuroradiologists in evaluating an AI tool for clinical neuroradiology applications.	5
Kwong (2021) / Canada [47]	STREAM-URO / 29	ML in urology	European Urology Focus	To improve ML literacy in the field of urology by establishing a standard for reporting ML applications.	18
Mongan (2020) / US [37]	CLAIM / 42	AI in medical imaging	Radiology: Artificial Intelligence	A recommended method for reporting medical imaging studies.	442
Mörch (2020) / Canada [48]	Canada protocol / 36	AI in mental health	Artificial Intelligence In Medicine	Focusing on mental health and suicide prevention, this study explores methods to more effectively identify and respond to ethical challenges in AI.	13
Schwendicke (2021) / Germany [49]	NS / 31	AI in dental research	Journal of Dentistry	Instructions for the design, execution, and reporting of research with dental AI.	102
Sengupta (2020) / US [50]	PRIME / 28	Cardiovascular Imaging-Related ML	Cardiovascular Imaging	To ensure that the ML models used in cardiovascular imaging studies are reported consistently, this comprehensive guide and associated checklist have been developed.	97
Naqa (2021) / US [51]	CLAMP / 26	AI in medical physics	Medical Physics	To guarantee rigorous and repeatable research of AI / ML in the area of medical physics, introducing a new, necessary checklist for AI / ML applications in Medical Physics (CLAMP).	20
Cerdá-Alberich (2023) / Spain [27]	MAIC–10 / 10	AI in medical images	Insights into Imaging	A guide for examining publications related to AI in medical imaging, with a focus on study design and evaluation.	4

a On July 18, 2023; b NS: Not specified

3.3 Guideline Development Process

The development details, main characteristics, user groups, focus areas and structure of the guidelines are reported in Table 2. Thirteen articles (54%) provided methodological details about the guideline development process. Ten guidelines (42%) were registered in the EQUATOR website. Two (8%) guidelines were extensions to existing EQUATOR guidelines (CONSORT-AI, SPIRIT-AI), the rest were standalone guidelines (92%). From the seven main components of guideline development defined by the EQUATOR Network, a literature review, a Delphi survey, an expert consensus meeting, and pilot testing were reported by 13 (54%), 6 (25%), 7 (29%), and 6 (25%) guidelines, respectively. Eleven guidelines (46%) reported a funding source, 4 guidelines, albeit vaguely, referred to a future update policy, and 6 guidelines (25%) were adapted or endorsed by a journal or professional organization. The development of guidelines involved on average 2.3 (SD 2.0, range 0-7) out of the 7 investigated components.

The development process was most comprehensive for CONSORT-AI [35] with all seven development steps completed, followed by SPIRIT-AI [36], which involved all key steps except the reporting of endorsement by a journal or professional

organization. The development of DECIDE-AI [29] was also comprehensive, but no reference was made about intentions to update it in the future.

The mean (SD) development steps of narrative guidelines, general clinical checklists, study design specific checklists and clinical area specific checklists were 0.5 (0.6), 1 (1.2), 5.3 (1.7), and 2.6 (1.6), respectively, with significant difference between the groups (ANOVA, $F_{3,20}=10.53$, $p<0.001$). Registration on the EQUATOR Network and obtaining funding were associated with more comprehensive development processes, while the endorsement by journals or professional societies was not an indicator of methodological rigor. The mean (SD) development components (except the grouping variable) in registered studies were 3.4 (2.3) versus 1.4 (1.4) in studies not registered in EQUATOR (Welch's t test, $p=0.031$). Funded studies featured 2.6 (1.9), whereas non-funded studies featured 1.1 (1.3) development steps (Welch's t test, $p=.034$). The difference between endorsed and not endorsed guidelines was not significant (Welch's t test, $p = 0.95$). Registration on EQUATOR or endorsement was not associated with higher citation counts. However, the Google Scholar citation count was higher for funded guidelines (mean: 303.8, SD: 102.9) than for those without funding (mean: 71.9, SD: 17.1) (Welch's t-test, $p=0.049$).

3.4 Main Characteristics of the Guidelines

Four papers (17%) were narrative guidelines (34-37) and twenty (83%) comprised a checklist. Seven checklists (30%) were formulated as general AI reporting standards without specific focus on any particular research domain [24, 26, 34, 35, 38, 40-45].

Four guidelines (17%) explicitly stated their focus on distinct study designs, encompassing randomized controlled trials [35] (30) clinical trial protocols [36], early-stage clinical evaluation of AI-driven decision support systems [29], and systematic reviews of prediction modelling studies [33]. Two of these are AI-related extensions of well-established checklists, namely the CONSORT for randomized trials [15] and the SPIRIT for clinical trial protocols [52].

Table 2

Characterization of the included reporting guidelines

Characteristic	Category	T.R.U.E.	Stevens (2020)	Faes (2020)	Bates (2020)	ROBUST-ML	Cabitza (2021)	CAIR	Scott (2021)	MINIMAR	MI-CLAIM	Luo (2016)	CONSORT-AI	SPIRIT-AI	DECIDE-AI	CHARMS	CLEAR Derm	R-AI-DIOLOGY	STREAM-URO	CLAIM	Canada protocol	Schwendicke (2021)	PRIME	CLAMP	MAIC-10
Guideline develop-	Development methods reported						✓		✓			✓	✓	✓	✓	✓	✓		✓		✓	✓	✓		✓
	Registered in EQUATOR website		✓								✓	✓	✓	✓	✓				✓	✓		✓	✓		
	Literature review						✓		✓			✓	✓	✓	✓	✓	✓		✓		✓	✓	✓		✓

ment process	Delphi survey											✓	✓	✓	✓					✓	✓			
	Expert consensus meeting												✓	✓	✓	✓	✓		✓					✓
	Pilot testing												✓	✓	✓	✓						✓			✓
	Funded					✓		✓				✓	✓	✓	✓	✓	✓		✓	✓			✓		✓
	Update policy was stated												✓	✓									✓		
	Journal / Society endorsement	✓					✓						✓							✓		✓	✓	✓	
Target audience	Authors, reviewers, editors	✓	✓	✓			✓			✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓
	Clinicians and model users			✓		✓		✓	✓	✓	✓							✓							
	Application developers																✓		✓		✓				
Type	Narrative	✓	✓	✓	✓																				
	Checklist					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Focus	General	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	Study design	Randomized clinical trial											✓												
		Clinical trial protocol												✓											
		Early stage clinical evaluation													✓										
		Systematic review														✓									
	Clinical area	Clinical imaging															✓		✓			✓		✓
		Dentistry																			✓				
		Cardiovascular medicine		✓		✓																✓			
		Cardiovascular imaging																				✓			
		Medical physics																					✓		
		Mental health																		✓					
		Dermatology															✓								
		Urology																	✓						
		Neuroradiology															✓								
		Nuclear medicine	✓																						
		Ophthalmology			✓																				
		Orthopaedics						✓																	
Structure	IMRAD							✓				✓	✓	✓	✓				✓	✓			✓	✓
	Machine Learning Pipeline		✓	✓	✓	✓	✓		✓	✓	✓					✓	✓	✓					✓		
	Other	✓																			✓	✓			

Nine guidelines (38%), were designed to address various clinical areas including urology [47], neuroradiology [28], mental health [48], medical physics [51], medical imaging [27] [37], dermatology [46], dentistry [49], and cardiovascular imaging [50]. While discussing general AI-related standards, through the journal or the elaboration examples, five (21%) more guidelines could be indirectly associated with nuclear medicine [39], cardiovascular medicine [26] [40], ophthalmology [41], and orthopaedics [43]. Some areas overlapped, such as cardiovascular imaging, medical imaging, and cardiovascular medicine.

Of the total publications analyzed, 20 (74%) were designed for authors, reviewers, and editors, while eight (30%) were tailored to clinicians and model users, and three (11%) were intended for application developers. Two guidelines (7%) did not specify their intended audience [34] [42].

3.5 Analysis of the Structure of Reporting Guidelines

The structure, and level of detail of both the narrative guidelines and point-by-point checklists showed significant heterogeneity.

Nine guidelines followed the usual IMRAD (Introduction, Methods, Results, and Discussion) structure of research articles with one or more additional sections (e.g., Title/Abstract, Statements/Other information) or omissions [27, 34-37, 43, 47, 51, 53]. The number of reporting items within the IMRAD group ranged between 10 [27] and 66 [36] with a median of 40.

Albeit with greater variation, the structure of other 12 guidelines followed the ML pipeline method of clinical AI studies as summarized by MI-CLAIM (i.e., Study design, Data and optimization, Model performance, Model examination, Reproducibility) with frequent additions of partial clinical information domains (e.g., Participants, Outcomes, or Clinical Deployment) or omissions [24, 26, 28, 33, 38, 40-42, 44-46, 50]. The number of items within the ML group ranged between 8 [41] [42] and 55 [24] (median: 21).

The third group comprised three guidelines with structures not fitting into either the IMRAD or ML pipeline frameworks [39, 48, 49] with items ranging between 4 [39] and 36 [48].

The structure of checklist differed considerably within subgroups. The most similar structure was observed across the checklists for specific study designs, as three (CONSORT-AI, SPIRIT-AI, DECIDE-AI) out of the four followed IMRAD with additional sections such at Title/Abstract and Statements/Other information. Although the fourth checklist (CHARMS) in this subgroup was also designed to be used by authors, namely researchers performing systematic review studies, its structure corresponded more the ML pipeline focusing on details of data and the model (26 items), with less emphasis on participants and results (9 items).

From the general checklists two articles followed the IMRAD structure [34] [43] and five the ML pipeline [24, 26, 38, 44, 45]. Among checklists for specific clinical areas four followed IMRAD [27, 37, 47, 51], and three the ML pipeline [28, 46, 50]. CLAIM integrated the ML workflow elements within the IMRAD format [37].

Altogether, the 24 guidelines contained 704 items in 224 sections. Many items were complex statements covering more than one reporting element. The mean number of items (i.e., depth of detail) differed significantly by the type and focus of guidelines. The mean (SD) item count of narrative guidelines, general checklists, checklists for specific study design and checklists for specific clinical areas were 9.8 (6.4), 32.7 (16.8), 47 (14.0) and 27.6 (9.0) respectively (ANOVA, $F_{3,20}=6.31$, $p=0.003$). However, we found no association between the depth of detail and guideline structure, with mean (SD) item count of 39.1 (16.3) for guidelines with IMRAD, 23.4 (13.0) for those with ML pipeline and 23.7 (17.2) for guidelines with other structure (ANOVA, $F_{2,21}=3.17$, $p=0.063$). Furthermore, we found no association between the type (i.e., narrative, or general) and focus (i.e., study design specific, or clinical area specific checklists) and the structure of guidelines (Fisher's exact test, $p=0.272$). However, the structure of guidelines and the comprehensiveness of their development were associated, with mean (SD) 2.8 (1.9), 0.9 (1.2), and 2.3 (1.5) development steps of guidelines with IMRAD, ML pipeline

or other structures, respectively (ANOVA, $F_{2,21}=3.72$, $p=0.041$). The number of development steps and the number of items showed moderate positive correlation ($r = 0.56$, $p=0.005$).

4 Discussion

According to our knowledge, this is the first systematic review that provides a broad overview of reporting guidelines for medical AI studies. While all 24 included guidelines aim to improve the transparency of reporting, and hence, indirectly improve the quality and clinical utility of medical AI studies, they were heterogenous in terms of their target audiences, focus area, development process and structure. The multiplicity and variety of items in the 24 reporting guidelines also reflect that there is not yet an established methodological framework for designing and reporting AI-based studies.

The proliferation of methodological frameworks and definitions is a general phenomenon in the digital transformation of healthcare [54-57]. Medical AI reporting guidelines are no exception to this trend. With over 700 items, albeit with overlap, the number of concepts that should be covered in an AI study report is overwhelming. A qualitative content analysis found similar heterogeneity in the concepts covered by medical AI reporting guidelines [58].

The guidelines' level of detail varied significantly regardless of whether they followed the IMRAD, ML pipeline or other structure. This variety makes it difficult for authors or clinicians to get a firm grip on what represents best practice in reporting, and how detailed a standardized medical AI study report should be. This may delay the clinical uptake of medical AI technologies, hamper evidence syntheses and consequently impede reliable information retrieval, which is essential for the development of automated systematic literature reviews [59].

Although the study design, clinical area, or a target journal may guide the choice of a reporting guideline, it remains a question, what represents a universal standard for best reporting practices. For a hint, authors may look at, which underwent the most robust development process, namely CONSORT-AI, SPIRIT-AI, and DECIDE-AI with 7, 6 and 5 development steps and 49, 66 and 38 reporting items, respectively. All these guidelines follow the IMRAD structure. Indeed, the more comprehensively developed guidelines were more detailed, and the IMRAD structure was associated with more robust development and greater depth of detail. However, the number of AI-specific items was only 14 in CONSORT-AI, 16 in SPIRIT-AI, and 28 in DECIDE-AI, suggesting that broad consensus supports a lower number of essential items compared to the most comprehensive general AI-related checklists, which featured more than 50 items [24] [34]. CONSORT-AI and SPIRIT-AI focus on randomized clinical trials, while DECIDE-AI is concerned

about clinical implementation pilot studies of AI-based technologies. While the long-awaited TRIPOD-AI checklist [60] for predictive modelling and the STARD-AI checklist [61] for diagnostic accuracy studies are under development, AI extensions of clinical study guidelines may not cover all the application fields of AI in medicine. In certain research contexts, an established checklist for a special study design, and a general AI-based checklist may be combined for optimal results.

The citation count may be another clue for authors to select well-established reporting guidelines. In this regard, CHARMS, the checklist by Luo et al., and CONSORT-AI stand out of the crowd with respectively 1080, 573 and 523 Google Scholar citations. The performance of CONSORT-AI is notable, as this guideline was published in 2020, several years after CHARMS (2014) and the work by Luo et al. (2016). While the checklist by Luo et al. has been tested (and cited) over time, it has not been updated for nine years since its publication [34]. A policy for regular updates was generally lacking from the guidelines, with only four guidelines making remarks about the need for future updates [35, 36, 47, 50], which is an important aspect in an area of rapid methodological development. While the number of citations for these guidelines seems impressive, these numbers also show their low adoption rates in the literature. Since 2016, the publication of the first clinical study reporting guideline, over 110000 AI-related studies were published on PubMed [62], while the total citation count for all identified guidelines was only 4277 over the same period.

The publication outlets of guidelines provided useful guidance in the selection of guidelines, although not without some discrepancies. The ten guidelines registered on the EQUATOR website had more robust development than the non-registered ones, but some registered guidelines did not report details of their development [38] [40]. However, CHARMS, a well-developed and the most cited guideline was not registered on EQUATOR. The citation analysis did not signal a greater uptake (more citations) of registered guidelines. To signal their quality, well developed guidelines should be registered on the EQUATOR website. CONSORT-AI, and SPIRIT-AI were published simultaneously in three prestigious journals (BMJ, Nature Medicine, Lancet Digital Health), and DECIDE-AI was published in two (BMJ, Nature Medicine). However, the author guidelines of BMJ and Nature Medicine recommended the use of CONSORT and reporting guidelines from EQUATOR, but did not mention CONSORT-AI [63] [64]. The author guidelines of Lancet Digital Health recommended CONSORT-AI and SPIRIT-AI, but not MI-CLAIM, which was published in the same journal [65]. Review articles that aimed to cover reporting guidelines for medical AI, usually mentioned CONSORT-AI and SPIRIT AI, along with guidelines under development, such as the STARD-AI, TRIPOD-AI and PROBAST-ML extensions of renowned EQUATOR guidelines. Reviews also mentioned CLAIM (proposed by the Radiological Society of North America), MINIMAR (published in the Journal of the American Medical Informatics Association), DECIDE-AI, MI-CLAIM, CHARMS, CAIR, and the guidelines from Luo, Cabitza, Bates, and Scott [18, 25, 66, 67].

Beyond the eminent examples cited by previous reviews, being the first systematic review in the field, our study introduces in detail a comprehensive list of reporting guidelines and demonstrates the challenges of guideline choice from a user perspective. Our analysis supports also the work of guideline developers. However, the limitation of our study is that some of our analyses required judgement, especially about delineating narrative guidelines or checklists, as well as about the inclusion of some articles, which were somewhat loosely related to study reporting. Potential omissions include a critical checklist to assess patient benefit [68], or an extensive report from the National Academy of Medicine [69]. We also excluded guidelines under development, such as STARD-AI, TRIPOD-AI or PROBAST-AI. Novel guidelines, such as the Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI), published during the publication process of this paper have been omitted [70]. Furthermore, Google Scholar citations do reflect the full utility of the included guidelines within a research community, such as their application in the editorial and review process of medical AI studies.

Conclusions

Currently, there is no professional consensus on the content and structure of the reporting guidelines for medical AI studies. The variety of reporting guidelines poses a challenge for researchers to follow what represents best practice in reporting a medical AI study.

Based on the robustness of development process and support from the literature, the CONSORT-AI extension is the most established tool. However, focusing on randomized controlled trials, CONSORT-AI does not cover the breadth of potential clinical and study design aspects of medical AI studies. While other AI extensions of EQUATOR guidelines are awaited, probably in combination with established EQUATOR guidelines, a modular AI-specific reporting guideline would provide the greatest flexibility for researchers and clinicians to follow the best practice in designing, writing and interpreting medical AI studies. Such guideline should undergo a robust development under the cooperation of powerful organizations and should be widely publicized among the medical AI research community to achieve its desired impact.

Acknowledgement

This project has been supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme.

References

[1] Weissler, E. H. et al.: The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, **22** (1), 2021, p. 537

[2] Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. *Futur. Healthc. J.*, **6** (2), 2019, pp. 94-98

[3] Macsik, P. et al.: Local Binary CNN for Diabetic Retinopathy Classification on Fundus Images. *Acta Polytech. Hungarica*, **19** (7), 2022, pp. 27-45

[4] Piros, P. et al.: Further Evolution of Mortality Prediction with Ensemble-based Models on Hungarian Myocar-dial Infarction Registry. *Acta Polytech. Hungarica*, **20** (4), 2023, pp. 125-140

[5] Chandaran, S. R. et al.: Deep Learning-based Transfer Learning Model in Diagnosis of Diseases with Brain Magnetic Resonance Imaging. *Acta Polytech. Hungarica*, **19** (5), 2022, pp. 127-147

[6] Orosz, G. et al.: Lung Ultrasound Imaging and Image Processing with Artificial Intelligence Methods for Bedside Diagnostic Examinations. *Acta Polytech. Hungarica*, **20** (8), 2023, pp. 69-87

[7] Briganti, G., Le Moine, O.: Artificial Intelligence in Medicine: Today and Tomorrow. *Front. Med.*, **7**, 2020

[8] Busnatu, Ștefan et al.: Clinical Applications of Artificial Intelligence—An Updated Overview. *J. Clin. Med.*, **11** (8), 2022, p. 2265

[9] Muehlematter, U. J. et al.: Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit. Heal.*, **3** (3), 2021, p. e195-e203

[10] Masic, I. et al.: Evidence Based Medicine - New Approaches and Challenges. *Acta Inform. Medica*, **16** (4), 2008, p. 219

[11] Chidambaram, A. G., Josephson, M.: Clinical research study designs: The essentials. *Pediatr. Investig.*, **3** (4), 2019, pp. 245-252

[12] Stephenson, J. M.: Overview of study design in clinical epidemiology. *Sex. Transm. Infect.*, **76** (4), 2000, pp. 244-247

[13] Moher, D.: Reporting guidelines: doing better for readers. *BMC Med.*, **16** (1), 2018, p. 233

[14] Papakostidis, C., Giannoudis, P. V: Meta-analysis. What have we learned? *Injury*, **54**, 2023, pp. S30-S34

[15] Schulz, K. F. et al.: CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.*, **8** (1), 2010, p. 18

[16] Moher, D. et al.: Guidance for Developers of Health Research Reporting Guidelines. *PLoS Med.*, **7** (2), 2010, p. e1000217

[17] EQUATOR: *Enhancing the QUAlity and Transparency Of health Research*. 2023

[18] Shelmerdine, S. C. et al.: Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Heal. Care Informatics*, **28** (1), 2021, p. e100385

[19] Zrubka, Z. et al.: The Reporting Quality of Machine Learning Studies on

Pediatric Diabetes Mellitus: Systematic Review. *J. Med. Internet Res.*, **26**, 2024, p. e47430

[20] Liu, X. 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. Heal.*, **1** (6), 2019, pp. e271-e297

[21] Nagendran, M. et al.: Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 2020, p. m689

[22] Kim, D. W. et al.: Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J. Radiol.*, **20** (3), 2019, p. 405

[23] Yusuf, M. et al.: Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*, **10** (3), 2020, p. e034568

[24] Cabitza, F., Campagner, A.: The need to separate the wheat from the chaff in medical informatics. *Int. J. Med. Inform.*, **153**, 2021, p. 104510

[25] Ibrahim, H. et al.: Reporting guidelines for artificial intelligence in healthcare research. *Clin. Experiment. Ophthalmol.*, **49** (5), 2021, pp. 470-476

[26] Al-Zaiti, S. S. et al.: A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). *Eur. Hear. J. - Digit. Heal.*, **3** (2), 2022, pp. 125-140

[27] Cerdá-Alberich, L. et al.: MAIC–10 brief quality checklist for publications using artificial intelligence and medical images. *Insights Imaging*, **14** (1), 2023, p. 11

[28] Haller, S. et al.: The R-AI-DIOLOGY checklist: a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology. *Neuroradiology*, **64** (5), 2022, pp. 851-864

[29] Vasey, B. et al.: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.*, **28** (5), 2022, pp. 924-933

[30] Zrubka, Z. et al.: *Time to start using checklists for reporting artificial intelligence in health care and biomedical research: a rapid review of available tools*. In: 2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES) IEEE, 2022, pp. 000015-000020

[31] Simera, I. et al.: Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med.*, **8** (1), 2010, p. 24

[32] Page, M. J. et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 2021, p. n71

[33] Moons, K. G. M. et al.: Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med.*, **11** (10), 2014, p. e1001744

[34] Luo, W. et al.: Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J. Med. Internet Res.*, **18** (12), 2016, p. e323

[35] Liu, X. et al.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.*, **26** (9), 2020, pp. 1364-1374

[36] Cruz Rivera, S. et al.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.*, **26** (9), 2020, pp. 1351-1363

[37] Mongan, J. et al.: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.*, **2** (2), 2020, p. e200029

[38] Norgeot, B. et al.: Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.*, **26** (9), 2020, pp. 1320-1324

[39] Buvat, I., Orlhac, F.: The T.R.U.E. Checklist for Identifying Impactful Artificial Intelligence–Based Findings in Nuclear Medicine: Is It True? Is It Reproducible? Is It Useful? Is It Explainable? *J. Nucl. Med.*, **62** (6), 2021, pp. 752-754

[40] Stevens, L. M. et al.: Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circ. Cardiovasc. Qual. Outcomes*, **13** (10), 2020

[41] Faes, L. et al.: A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. *Transl. Vis. Sci. Technol.*, **9** (2), 2020, p. 7

[42] Bates, D. W. et al.: Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence. *Ann. Intern. Med.*, **172** (11_Supplement), 2020, pp. S137-S144

[43] Olczak, J. et al.: Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop.*, **92** (5), 2021, pp. 513-525

[44] Scott, I. et al.: Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Heal. Care Informatics*, **28** (1), 2021, p. e100251

[45] Hernandez-Boussard, T. et al.: MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Informatics Assoc.*, **27** (12), 2020, pp. 2011-2015

[46] Daneshjou, R. et al.: Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology. *JAMA Dermatology*, **158** (1), 2022, p. 90

[47] Kwong, J. C. C. et al.: Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework. *Eur. Urol. Focus*, **7** (4), 2021, pp. 672-682

[48] Mörch, C.-M. et al.: Canada protocol: An ethical checklist for the use of artificial Intelligence in suicide prevention and mental health. *Artif. Intell. Med.*, **108**, 2020, p. 101934

[49] Schwendicke, F. et al.: Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.*, **107**, 2021, p. 103610

[50] Sengupta, P. P. et al.: Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist. *JACC Cardiovasc. Imaging*, **13** (9), 2020, pp. 2017-2035

[51] El Naqa, I. et al.: AI in medical physics: guidelines for publication. *Med. Phys.*, **48** (9), 2021, pp. 4711-4714

[52] Chan, A.-W. et al.: SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*, **346** (jan08 15), 2013, pp. e7586-e7586

[53] Vasey, B. et al.: DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.*, **27** (2), 2021, pp. 186-187

[54] Kolasa, K., Kozinski, G.: How to Value Digital Health Interventions? A Systematic Literature Review. *Int. J. Environ. Res. Public Health*, **17** (6), 2020, p. 2119

[55] Zah, V. et al.: Paying for Digital Health Interventions – What Evidence is Needed? *Acta Polytech. Hungarica*, **19** (9), 2022, pp. 179-199

[56] Burrell, A. et al.: How Useful Are Digital Health Terms for Outcomes Research? An ISPOR Special Interest Group Report. *Value Heal.*, **25** (9), 2022, pp. 1469-1479

[57] Zrubka, Z. et al.: The PICOTS-ComTeC Framework for Defining Digital Health Interventions: An ISPOR Special Interest Group Report. *Value Heal.*, **27** (4), 2024, pp. 383-396

[58] Crossnohere, N. L. et al.: Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks. *J. Med. Internet Res.*, **24** (8), 2022, p. e36823

[59] Tóth, B. et al.: Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. *Syst. Rev.*, **13** (1), 2024, p. 174

[60] Collins, G. S. et al.: 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** (7), 2021, p. e048008

[61] Sounderajah, V. et al.: Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*, **11** (6), 2021, p. e047709

[62] National Library of Medicine: *PubMed*. 2023

[63] Nature medicine: *Clinical Research*. 2023

[64] BMJ: *Article types and preparation*. 2023

[65] The Lancet Digital Health: *Information for Authors*. 2023

[66] Campbell, J. P. et al.: Reporting Guidelines for Artificial Intelligence in Medical Research. *Ophthalmology*, **127** (12), 2020, pp. 1596-1599

[67] Meshaka, R. et al.: Artificial intelligence reporting guidelines: what the pediatric radiologist needs to know. *Pediatr. Radiol.*, **52** (11), 2022, pp. 2101-2110

[68] Vollmer, S. et al.: Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, 2020, p. l6927

[69] Matheny, M. et al. eds.: *Artificial Intelligence in Health Care*. Washington, D.C.: National Academies Press, 2019

[70] Elvidge, J. et al.: *Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI)*. Value in Health, 27 (9), 2024, pp. 1196-1205