

Review

AI Model Reporting Standards in Health Care: Systematic Review and Consolidation

Ekaterina Akhmad^{1,2}; Daniël Slob^{1,2}, MSc; Aiara Lobo Gomes^{1,2}, PhD; Simone Mingels^{1,2}, MSc; Andre Dekker^{1,2}, PhD; Johan van Soest^{1,2,3}, PhD

¹Department of Radiation Oncology, GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

²Maastricht Radiation Oncology Institute, Maastricht, The Netherlands

³Brightlands Institute for Smart Society (BISS), Faculty of Science and Engineering, Maastricht University, Heerlen, The Netherlands

Corresponding Author:

Ekaterina Akhmad
Department of Radiation Oncology
GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+
Paul-Henri Spaaklaan 1
Maastricht 6229 GS
The Netherlands
Phone: +31 (0) 43 3885457
Email: ekaterina.akhmad@maastrichtuniversity.nl

Abstract

Background: AI technologies are currently experiencing significant growth in development and use in health care. The AI model lifecycle includes several stages, all of which should be adequately described. Reporting standards are commonly used for this purpose. However, the plain-text structure of these standards prevents transparent machine-readability of this information.

Objective: This systematic review aims to investigate AI model reporting standards from clinical databases and technical documents published since 2016.

Methods: The methodological systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. The inclusion criteria of this systematic review were papers published from 2016 to April 2026 in English and in a peer-reviewed journal with open access. We aimed to include standards for different data types (structured, images, and text-based). The exclusion criteria were specific medical domain, incomplete reports, and reviews. The following libraries have been consulted: Scopus and Web of Science, MEDLINE and Embase, the IEEE Xplore digital library, and the ACM Digital Library. We applied an Automated Systematic Review software tool (Utrecht University) to find relevant papers after records retrieval. Two reviewers (EA, DS) screened relevant papers independently and invited a third reviewer (JvS) to reach a consensus. All included papers were analyzed to extract relevant lifecycle stage, publication type, data type, and reporting questions. The results were presented and synthesized in a consolidated list by grouping questions and ordering them in a logical sequence according to the AI model lifecycle.

Results: The total number of included studies was 25 from clinical and technical databases (22 and 3, respectively). The consolidated list includes 134 questions ordered in 7 groups: “Problem identification,” “Data collection and processing,” “Model development,” “Technical validation,” “Functional validation,” “Operation & Monitoring,” and one additional group with “Statements.” Groups according to the lifecycle stages contain all relevant information for each stage. Both clinical and technical papers reflect information for “Problem identification” (s=5), “Data collection and processing” (s=25), “Model development” (s=13), “Technical validation” (s=14). While “Functional validation” (s=30) was represented mostly in clinical papers and “Operation & Monitoring” (s=18) in technical ones. In the “Statements” (s=29), all general questions were summarized, describing reusability and reproducibility of the model, and highlighting ethical considerations, such as model fairness, explainability, interpretability, and related risk. It was shown that only 10 out of 15 FAIR (findable, accessible, interoperable, reusable) principles were addressed in the consolidated list, with focus on providing rich metadata (F2) and accurate and relevant descriptions (R1).

Conclusions: We furthermore considered the mapping of the consolidated questions with FAIR principles. In future, we aim to incorporate the results of our work into a metadata schema for AI models developed according to FAIR principles.

JMIR AI 2026;5:e92329; doi: [10.2196/92329](https://doi.org/10.2196/92329)

Keywords: artificial intelligence; machine learning; reporting guidelines; medicine; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Rationale

Machine learning and AI models are currently experiencing significant growth in development and use [1], especially in health care. AI products are being increasingly deployed in hospitals worldwide [2]. The AI model lifecycle generally includes several stages, such as development, validation, and deployment [3]. All stages should be adequately described to give enough information when comparing different AI models on clinical validity, or during deployment, use, and monitoring of AI models in clinical practice. For this purpose, reporting standards are commonly used to provide reproducibility and clinical validity of findings in data science studies and specified also for AI models. Furthermore, these guidelines help authors in reporting their research, and are becoming mandatory for publishing in academic journals [4]. However, there is a rising number of standards in recent years based on the community publishing a reporting standard; from clinical to technical domains and from general to very domain specific [5].

Reporting standards mostly use approaches of storing information with plain-text and lack structure. This plain-text approach complicates comparing questions, let alone answers on the specific reporting standard questions. This reduces the comparison of AI models and methods used during the various AI model lifecycle stages. Furthermore, the paper or literature-based format of reporting prevents machine-readability. Therefore, their contents are difficult to meaningfully collect, process, and order for storing purposes in libraries, repositories, or other digital platforms. As a result, it is challenging for potential users or other stakeholders to perform systematic searches for the existence of specific AI models. The common approach to deal with these obstacles in information collection and storage is to apply the FAIR (findable, accessible, interoperable, and reusable) principles [6,7]. This framework was introduced in 2016 to describe the concept of open science and reproducible scientific practice [7]. Although originally developed for datasets, FAIR principles are also applied to AI models, since models can be considered as a number of data-containing components such as features, parameters, executable files [6,8]. In this analogy to datasets, the metadata describes the AI model, where the model itself is the data object. Several systematic reviews of reporting standards in medicine were published that aimed to provide a high-level overview for writing and reviewing research papers [9], to improve information delivery across general medical specialties [10], and to summarize the key

information necessary for reporting based on the maximum agreement between standards [5]. However, we did not come across any comparison between FAIR principles and reporting standards. All stakeholders (AI developers, health care providers, clinicians, and other healthcare professionals) could benefit from using an alignment between reporting standards and FAIR principles for AI models and datasets, as provided information will be shared easily and transparently.

Objectives

This study aimed to investigate AI model reporting standards from clinical databases and technical documents published since 2016, when the FAIR principles were introduced [7]. For that purpose, the methodological systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2020 guidelines [11]. This systematic review aimed to study the following research questions:

1. What is the complete list of themes and questions included in reporting standards?
2. How do different themes correspond with the scope of reporting standards?
3. Which of the FAIR principles are addressed in reporting standards?

Methods

Overview

We conducted a systematic review of published methodological journal papers and technical reports from 2016 to April 2026 to analyze the description methods of AI models in different application fields. The protocol of this review was published before at the Open Science Framework [12]. We used the PRISMA 2020 guidelines [11] to describe the results of the systematic review. The PRISMA 2020 checklist is available in the [Checklist 1](#).

Eligibility Criteria

To be included in our review, a paper had to meet all the following general criteria:

- Period: from 2016 to April 2026
- Published in the English language
- Published in a peer-reviewed journal
- Open-access and/or Institutional Access

Papers that cover all data types are included, such as structured data, medical images, and text. However, we excluded papers from a specific domain (for example, medical images for surgery, biology, medical physics, nuclear

medicine, and radiomics). We also excluded studies that mention only some parts of reports, for example, methods for data collection. We excluded review papers of previously published works if they did not contribute additional meaning to the reporting. For example, if the paper summarizes the previous standards without additional information. However, papers were included if it proposed a new reporting standard based on review and new information from experts.

We conducted an initial search in March 2024 to cover the period from 2016 to March 2024 and a second search in April 2026 for the period from 2024 to April 2026. Therefore, we used the same search queries for all databases and an identical analysis.

Information Source

The general data sources included Scopus and Web of Science databases. For biomedical papers, we searched MEDLINE and Embase; for technical papers, we searched IEEE Xplore and the ACM digital libraries. For MEDLINE and Embase, we used the Ovid search interface (Ovid Technologies). The title, abstract, and keywords of papers and standard items (for example, MeSH terms), if applicable for the database, were included in search queries. The information types included articles, book chapters, books, and conference materials. In addition to the given search, we used data from systematic and narrative reviews published in peer-reviewed journals in the last 5 years [5,9,13] for validation of our search strategy and a hand search of their reference lists of included reporting papers. For the second search, we used exactly the same databases and search queries as for the first one.

Search Strategy

The search query for each database is presented in Table S1 in [Multimedia Appendix 1](#).

Selection Process

We combined the identified papers in one database and then deduplicated them in EndNote (version 21.3, Clarivate) using the multifield strategy [14]. Due to the large number of records retrieved from the reference databases, we used the Automated Systematic Review software (AsReview) to find relevant papers based on their active learning algorithms [15]. The selection process was conducted for 2 searches separately.

For the first search (in March 2024), we used AsReview (version 1.6.2), then we provided prior information for the primary algorithm training, including 5 relevant records and 5 irrelevant records, and reviewed papers in 2 iterations. Firstly, a review process was performed with the default feature extraction technique, such as term frequency-inverse document frequency (TF-IDF) and a classifier (naive Bayes). One author (EA) marked the ranked articles as relevant and irrelevant, which provided continuous training for AsReview. The stopping criterion for the first version was 20 consecutive papers marked as irrelevant. After reaching the stopping criteria, the result table was extracted from AsReview and used for advanced model training on the

next step. Secondly, we implemented a more advanced feature extraction technique (sentence BERT [Bidirectional Encoder Representations from Transformers]) and classifier (fully connected neural network), while using relevant and irrelevant records from the first iteration for model pretraining. As in the previous stage, the reviewer (EA) indicated papers as relevant or irrelevant, with the stopping criterion of 50 consecutive irrelevant papers. After the search update (in April 2026), we screened papers using AsReview (version 3.0.5) and a model ELAS u4, as the AsReview issued an update. We set up the stricter stopping criterion (70 consecutive papers marked as irrelevant).

After the papers' preselection in AsReview, 2 reviewers (EA and DS) screened independently all abstracts marked as relevant to decide whether a paper met the inclusion criteria. Discussions were held to reach a consensus. Then, the same 2 reviewers screened full-text versions of papers included in the previous step and held further discussions for final consensus. A third reviewer (JvS) was invited to resolve disagreements between the 2 reviewers.

Data Collection Process

We developed a template for data extraction in a matrix that includes rows with names of reporting questions and columns for each of the reviewing papers and the corresponding FAIR principles. All included papers were sorted randomly. The reporting topics were iteratively grouped during the review process. Questions were extracted manually from papers using the following steps:

1. Search for a table, a picture, or a structured free text that presents the proposed reporting standard.
2. Extraction of the specific reporting questions from the selected section.
3. Matching the topics with the established lifecycle stage (see below).
4. If a question matched a question already existing in the consolidated list, marked it as "Directly mentioned" (in case of complete similarity) or "Interpreting to be similar."
5. If the question was not stated before in the consolidated list, we added it into the list and marked it as "Directly mentioned."

We based our list of model development stages on the lifecycle published by the US Food and Drug Administration (FDA) [3], which includes the following phases:

- Planning and Design
- Data Collection and Management
- Model Building and Tuning
- Verification and Validation
- Model deployment
- Operation and Monitoring
- Real-world Performance Evaluation

To avoid misclassification of questions to lifecycle phases, the classification of reporting standard questions to a specific phase was based on the goal of the reporting standard or proposed future applications. For example, terms such as "validation" and "implementation" can have different meanings depending on context. We assumed that reporting

standards mainly cover phases from 1 to 5, as most reporting standards attempt to describe the development and validation of models.

Quality and Risk of Bias Assessment

To verify the reliability of our data extraction and analysis, 2 researchers (EA and DS) screened the papers independently. Moreover, we analyzed papers in terms of reported themes and corresponding FAIR principles. The third reviewer (JvS) contributed to reaching a consensus.

Analysis

The reported questions from different papers were used to build a general list while matching them with corresponding stages of the AI model lifecycle. Themes were marked for the specific reviewed paper as “mentioned explicitly,” “interpreted as the same,” and none (not mentioned). We mapped extracted reporting questions with FAIR principles [16] for the AI model itself and its metadata. In our study, the AI model was assumed to be developed, so the data of the AI model will include output coefficients, weights, and other data related to the execution of the AI model. Therefore,

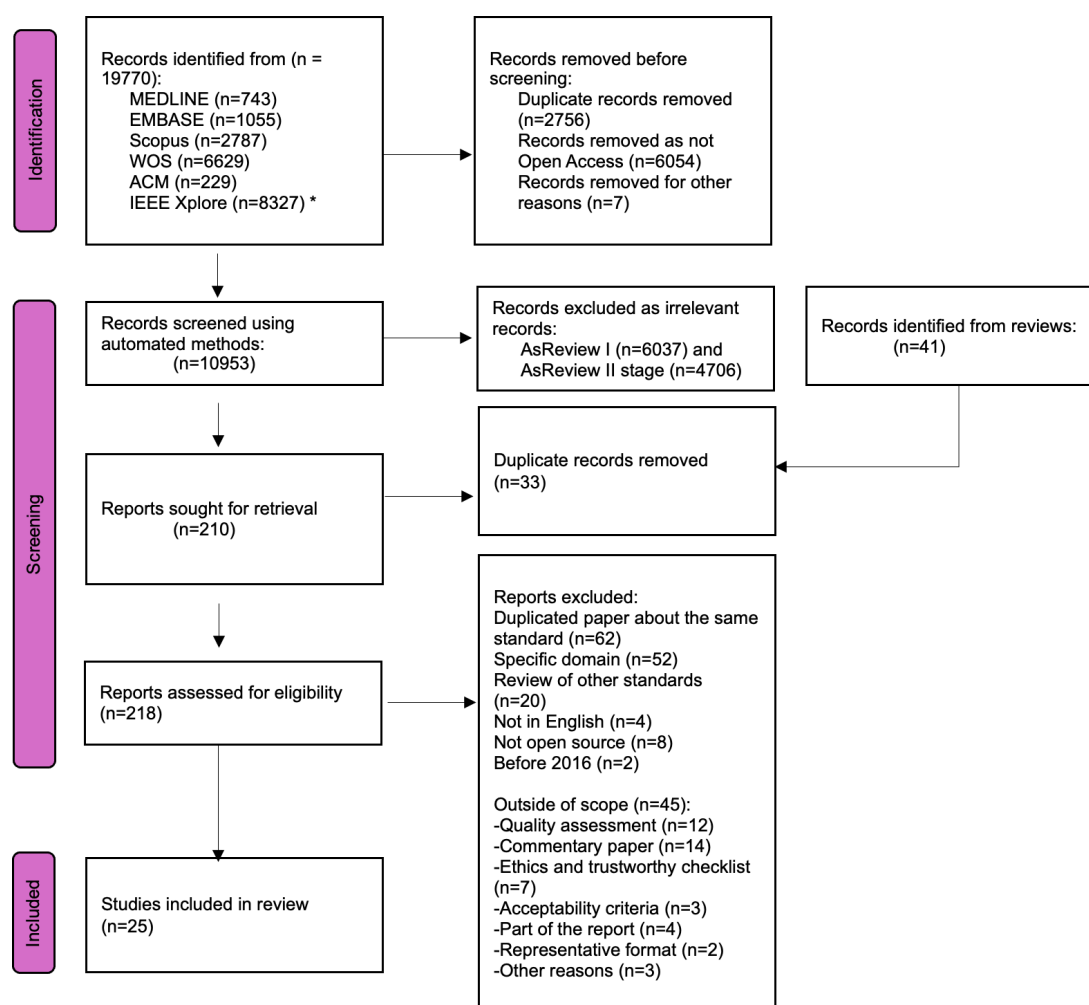
methods regarding the training pipeline and scripts and/or decisions regarding data processing were attributed to the metadata of the AI model.

Results

Study Selection

A total of 25 papers [17-41] met all eligibility criteria and were included in this study. For the presentation of included papers, we used the PRISMA flow diagram [11] for systematic reviews to summarize included papers (Figure 1). We identified 19,770 papers related to reporting standards in peer-reviewed journals. After deduplicating and removing records without open access (including also institutional access), we used AsReview to screen 6166 papers at the first stage and 4027 at the second stage of review, from which we reviewed 218 full-text documents. 193 papers were excluded as they did not meet the inclusion criteria; for example, 52 papers were excluded because of the specific application domain of AI models, such as medical physics, nuclear medicine, or in-game tools.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Flow-diagram. AsReview: Automated Systematic Review software.



We also excluded 4 papers because they were published in other languages, 20 papers because they presented reviews of different standards without new interpretation, 8 papers had no open-access full-text, and 2 papers were published before 2016. Sixty-two records were removed due to the duplication of papers about the same reporting standard. For example, 10 papers refer to the protocol of the standard's development. We also excluded different extensions of the same standard, such as the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines, which has a few extensions or variations: TRIPOD+AI [17], TRIPOD-Cluster [42], TRIPOD for abstract [43], TRIPOD-SRMA (systematic reviews and meta-analysis) [44]. These papers would not add specifically new information to the question list (TRIPOD-Cluster aimed to describe studies based on clustered datasets, TRIPOD for abstracts focused on writing abstracts, TRIPOD-SRMA was an extension for writing systematic reviews). Only TRIPOD+AI and TRIPOD-LLM were included as the more general, which focused specifically on prediction model studies in AI or studies using large language model (LLM). We also excluded previous versions of checklists; for example, Clinical AI Modeling (CLAIM) was first published in [45] with updates in [18]. As we aim to have up-to-date questions from the working group, we included the latest version [18].

In total, 45 papers were identified as outside of scope because of different factors and were excluded from the review. These factors included, for example, whether they provide only checklists of questions for quality, ethical, and/or trustworthy assessment, not specifically for reporting. These checklists commonly include several items to assess transparency, trustworthiness, suitability, and other aspects of AI models, for example, by presenting a diagnostic quality model for integrating AI models into clinical practice [46]. As the purpose of these papers was not to report the AI studies, they were considered out of scope. Additionally, there were commentary papers and papers focusing only on the specific part of the report (eg, methods for data collection); or papers with the discussion of the representation format of the report but without describing the content itself.

Study Characteristics

Among the included records, the number of papers from health care journals was almost 5 times more than from technical journals (22 and 3, respectively). The general characteristics of the papers included are presented in Table 1. Regarding data type, 15 papers covered reporting standards without any specific data type (general); 4 papers focused on medical images, 2 papers covered reporting of prediction models based on structured data and 4 papers were for text-based data.

Table 1. The characteristics of the included papers.

Group of characteristics and description	Number of records
Type of references' source	
Clinical database	22
Technical database	3
Data type	
General	15
Medical images	4
Structured data	2
Text-based data	4
Year of publication	
2016-2018	1
2018-2020	2
2020-2022	7
2022-2024	2
2024-2026 (April)	13

Comparison of Reporting Standards

In total, 25 papers [17-41] were screened to fill in the template. Table 2 presents the characteristics of the included papers. From each paper, we extracted the study characteristics as the evidence that lies in the foundation of the presented question list. We could distinguish 3 main types: consensus-based research (n=14) [17-30], a scientific publication from a biomedical database (n=8) [31-38], and

a scientific publication from a technical database (n=3) [39-41]. The first type was determined based on the materials and methods section in the paper, which stated that Delphi methods or other consensus-building approaches were used. The scientific publications were granted for other records published in peer-reviewed journals, but without mentioning consensus-based approaches.

Table 2. The characteristics of included papers.

	Citation	Data type	Study characteristics	Stage of lifecycle
1.	DECIDE-AI ^a [19]	General	Consensus-based research	Model building and tuning, verification and validation
2.	SPIRIT-AI ^b [20]	General	Consensus-based research	Model deployment (functional validation)
3.	CONSORT-AI ^c [21]	General	Consensus-based research	Model deployment (functional validation)
4.	TRIPOD+AI ^d [17]	Clinical data	Consensus-based research	Model building and tuning, verification and validation
5.	REFORMS ^e (Kapoor et al.) [22]	General	Consensus-based research	Model building and tuning, verification and validation
6.	Luo et al [23]	Clinical data	Consensus-based research	Model building and tuning, verification and validation
7.	Tejani et al [18]	Medical images	Consensus-based research	Model building and tuning, verification and validation
8.	Moassefi et al [24]	Medical images	Consensus-based research	Model building and tuning, verification and validation
9.	CHEERS-AI ^f (Elvidge et al) [25]	General	Consensus-based research	Model deployment (functional validation)
10.	TRIPOD-LLM ^g (Gallifant et al) [26]	LLM ^h	Consensus-based research	Across the whole lifecycle
11.	STARD-AI ⁱ (Sounderajah et al) [27]	General	Consensus-based research	Verification and validation, model deployment (functional validation)
12.	CHART ^j [28]	LLM	Consensus-based research	Verification and validation, model deployment (functional validation)
13.	GAMER ^k (Luo et al) [29]	LLM	Consensus-based research	Model deployment
14.	MedInAI (Bottacin et al) [30]	General	Consensus-based research	Across the whole lifecycle
15.	MI-CLAIM ^l (Norgerot et al) [31]	Medical images	Scientific publication	Model building and tuning, verification and validation
16.	CAIR ^m (Olczak et al) [32]	General	Scientific publication	Model building and tuning, verification and validation, model deployment
17.	Cabitza et al [33]	General	Scientific publication	Model building and tuning, verification and validation
18.	Stevens et al [34]	General	Scientific publication	Model building and tuning, verification and validation
19.	Model facts (Sendak et al) [35]	General	Scientific publication	Model deployment
20.	Kwak, Kim [36]	General	Scientific publication	Across the whole lifecycle
21.	MI-CLEAR-LLM ⁿ (Park et al) [37]	LLM	Scientific publication	Model deployment
22.	RIDGE ^o (Maleki et al) [38]	Medical images	Scientific publication	Model building and tuning, verification and validation
23.	Model card (Mitchell et al) [39]	General	Scientific publication from technical databases	Model deployment
24.	McCradden et al [40]	General	Scientific publication from technical databases	Model building and tuning, verification and validation, model deployment (functional validation)
25.	FactSheets (Arnold et al) [41]	General	Scientific publication from technical databases	Verification and validation, model deployment

^aDECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI.

^bSPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI.

^cCONSORT-AI: Consolidated Standards of Reporting Trials–AI.

^dTRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI.

^eREFORMS: Consensus-based Recommendations for Machine-learning-based Science.

^fCHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI.

^gTRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

^hLLM: large language model.

ⁱSTARD-AI: Standards for Reporting Diagnostic Accuracy–AI.

^jCHART: Chatbot Assessment Reporting Tool.

^kGAMER: Generative Artificial Intelligence Tools in Medical Research.

^lMI-CLAIM: Minimum Information about Clinical AI Modeling.

^mCAIR: Clinical AI Research.

ⁿMI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care.

^oRIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency.

Papers that were intended to guide the reporting aspects for clinical trials of AI models ([Standard Protocol Items: Recommendations for Interventional Trials–AI] SPIRIT-AI [20] and (Consolidated Standards of Reporting Trials–AI) CONSORT-AI [21]) or health economic evaluations ([Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI] CHEERS-AI [25]) were related more to the functional validation that occurs before market certification. This process could be connected with the phase of “Model deployment,” as the AI model at this stage has a scalable and capable format of integration into a system and could be validated in the real clinical practice workflow. Some papers provide only a brief explanation, for

example, Model facts [35] and Model cards [39]; in these cases, the stage was defined according to the questions.

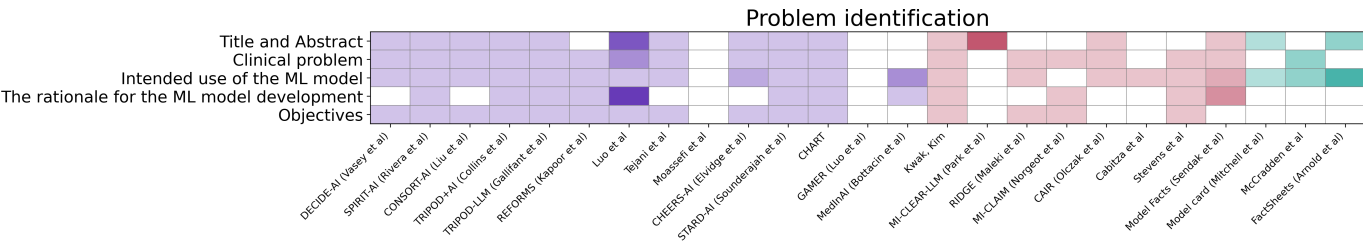
The consolidated list was formed by grouping 134 questions and ordering them in a logical sequence according to the AI model lifecycle. We extracted 7 groups of questions “Problem identification,” “Data collection and processing,” “Model development,” “Technical validation,” “Functional validation,” “Operation & Monitoring,” and one additional group with “Statements.” For cases when one question relates to several questions from the consolidated list, we matched them correspondingly. The complete consolidated list of questions can be found in [Multimedia Appendix 2](#). We will discuss in detail each of the formed groups in the Problem Identification section.

Problem Identification

Figure 2 shows questions that were extracted for the problem identification stage (s=5) that could be reflected in the introduction section of the article, containing also information about the title and objectives of the study. Since problem identification is seen as the most important stage in the (machine learning) ML lifecycle, reporters should address extracted questions about the clinical problem, intended use, and the rationale for the ML model development. This

diagram shows that almost all papers cover these questions to some extent, except the papers by Moassefi et al [24] and Generative Artificial Intelligence Tools in Medical Research (GAMER) [29]. The maximum number of questions for problem identification (14 items) was found in the paper by Luo et al [23], which covered detailed questions about—for example—the prediction goal and the prediction model type (such as classification, regression, and survival prediction).

Figure 2. Questions from the consolidated list that relate to the problem identification. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials–AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; ML: machine learning; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI; STARD-AI: Standards for Reporting Diagnostic Accuracy–AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

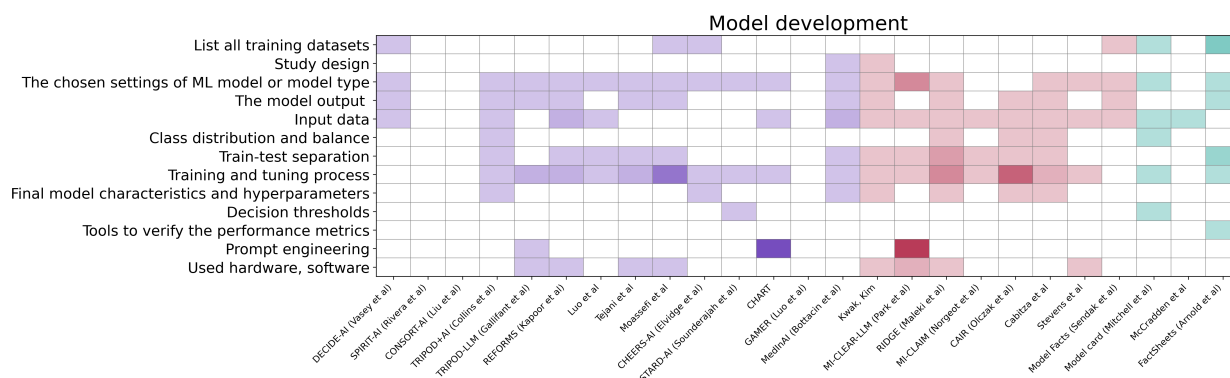


Model Development

Figure 3 represents questions for the model development phase (s=13), ranging from the list of all training datasets to the hardware and packages used. Only SPIRIT-AI and CONSORT-AI do not require this information, as they were aimed at clinical evaluation of the AI model as a final product. The most frequent questions relate to the model type and training and tuning process with different levels of detail. For example, some papers used the general expressions

like “How were the models trained” and “Information about training algorithms,” while Moassefi et al [24] and Olczak et al [32] ask for specific details. We extracted the question “Tools to verify the performance metrics” mentioned only in 1 paper by Mitchel et al [39] and “Decision thresholds” in 2 papers by Arnold et al [41] and STARD-AI [27]. We included the specific description for all questions in the Multimedia Appendix 2.

Figure 3. Questions from the consolidated list that relate to the model development. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials-AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; ML: machine learning; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI; STARD-AI: Standards for Reporting Diagnostic Accuracy-AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

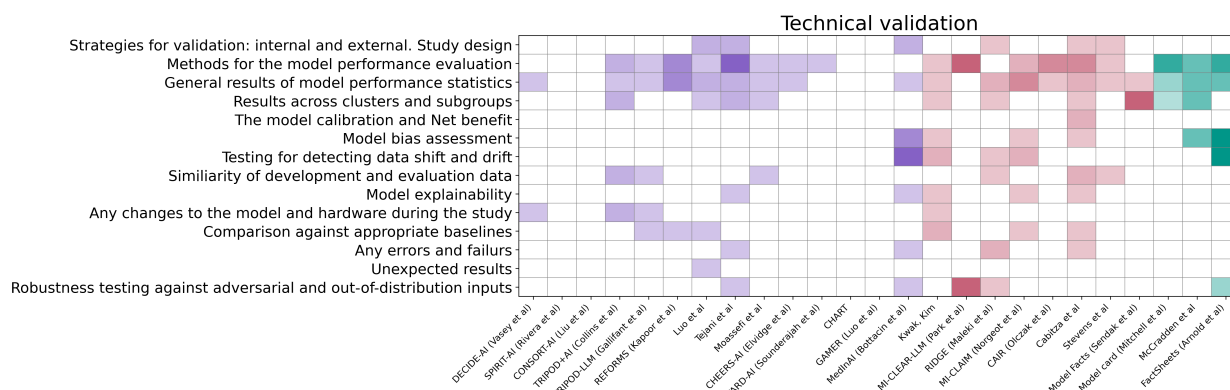


Technical Validation

Figure 4 demonstrates the reporting questions (s=14) for the technical validation of AI models. This testing can be performed on internal or external datasets. The goal of this lifecycle stage is to assess the technical performance of the AI model. The general topics, such as methods (n=17) [17,18,22-27,31-34,36,37,39-41], general results (n=19) [17-19,22-27,30-41], and results across clusters and subgroups (n=10) [17,18,23,24,33,35,36,38-40], were addressed in most

papers. Fewer papers (n=6 [30,31,33,36,40,41] and n=5 [30,31,36,38,41]) cover the remaining questions, such as model bias assessment and testing for data shift and drift detection. Several standards mentioned the question “Comparison against appropriate baselines” aimed to show an analysis of received validation results against appropriate baselines. The baseline values are usually stated from the literature, for example, from validation studies of similar AI models.

Figure 4. Questions from the consolidated list that relate to the technical validation. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials-AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; ML: machine learning; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI; STARD-AI: Standards for Reporting Diagnostic Accuracy-AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

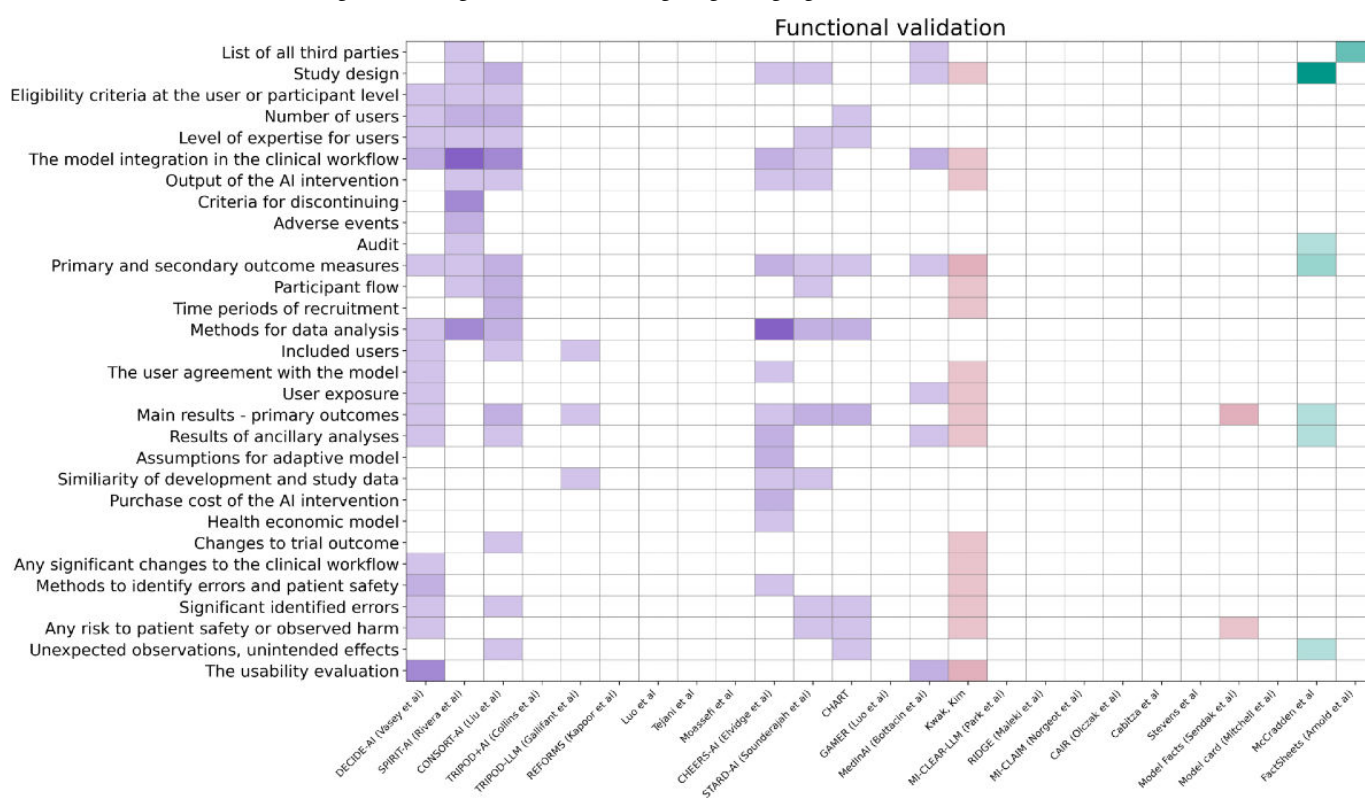


Functional Validation

As shown in Figure 5, questions related to the functional validation (s=30) were primarily mentioned in 4 papers: (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence) DECIDE-AI, SPIRIT-AI, CONSORT-AI, and CHEERS-AI. DECIDE-AI describes the early-stage clinical evaluation of an AI model. Evaluation of AI model real-world effectiveness and safety during clinical trials is reported by standards

for trial protocols (SPIRIT-AI) and trial reports (CONSORT-AI). These standards include a broad spectrum of questions about studying primary outcomes of AI intervention and any observed harm or risks to patient safety. CHEERS-AI is used for reporting economic evaluations of AI-based health interventions and measuring cost-effectiveness. In addition, the work of Kwak and Kim [36] presents a reporting guideline and a checklist across diverse study designs.

Figure 5. Questions from the consolidated list that relate to the functional validation. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials-AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; ML: machine learning; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI; STARD-AI: Standards for Reporting Diagnostic Accuracy-AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

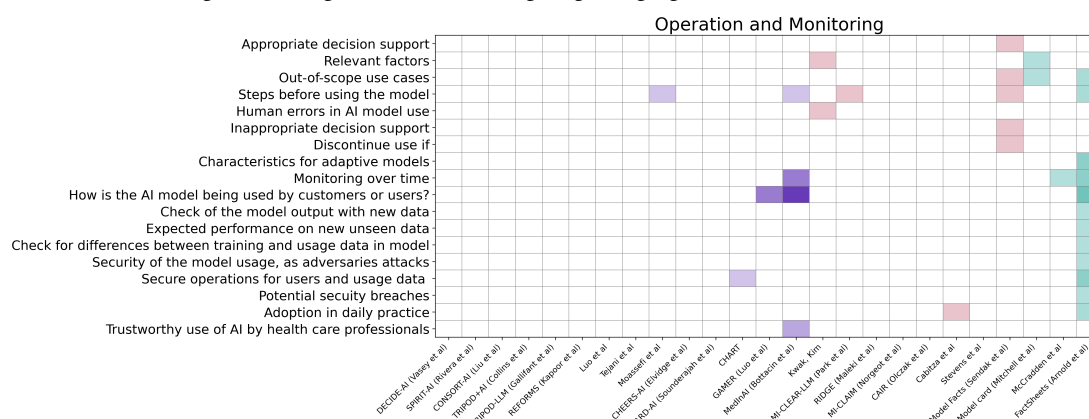


Model Facts consolidate general information about patient safety and efficacy evaluation, which is why we also present it in this section. McCradden et al [40] explores ethical and social justice principles not only for model development, but also for model evaluation, including silent trial evaluation (prospective non-interventional evaluation) and prospective clinical evaluation to study the influence of the AI intervention on clinical decision-making. However, it focuses more on ethical considerations and the importance of diversity in AI clinical trials rather than other aspects of reports.

Figure 6 represents questions for the Operation and Monitoring stage of the AI system (s=18). Arnold et al

[41] (FactSheets) is the most complete reporting standard in this domain, listing generic technical descriptions that are needed for the ML model operation stage. Some questions are addressed by commonly used Model card [39] and Model Facts [35]. MedinAI includes several sections to describe an AI model as service that should be technically approved and effective in real-world health care. In addition, it also provides a separate group of model updating items [30]. CHART [28] and GAMER [29] relate to the use of LLM models.

Figure 6. Questions from the consolidated list that relate to the operation and monitoring. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials–AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; ML: machine learning; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI; STARD-AI: Standards for Reporting Diagnostic Accuracy–AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

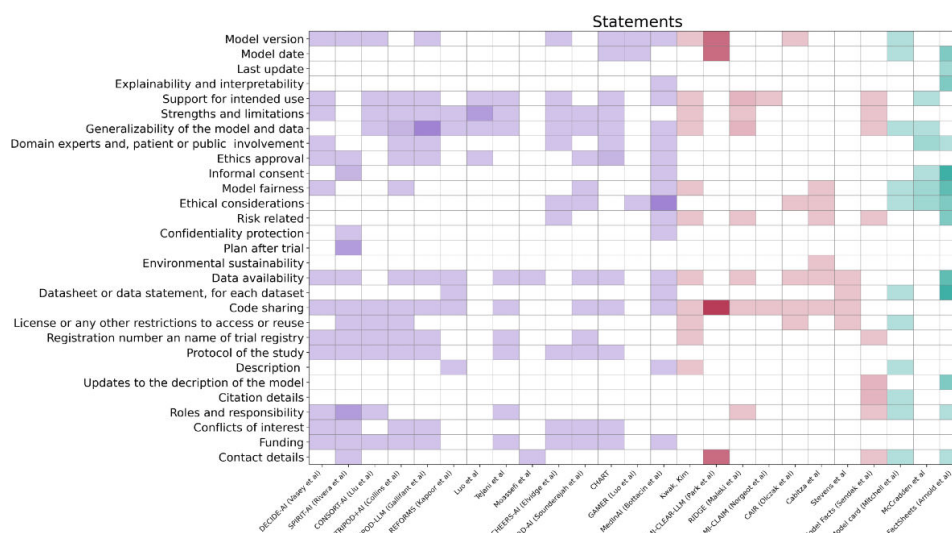


Statements

In the section “Statements” (Figure 7), we summarize all questions (s=29) with general model information, supporting its use in clinical practice. In addition, this section includes topics regarding the reusability and reproducibility of the

model, such as data availability and code sharing, license, and information about protocol registration and ethics approval. Other questions reflect model fairness, explainability and interpretability, related risk and other ethical considerations, that are mentioned at various stages of AI lifecycle.

Figure 7. Questions from the consolidated list that relate to the statements about the model. The color code provides the information about evidence characteristics: purple—consensus-based research, red—scientific publications, green—scientific publication from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials–AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: Generative Artificial Intelligence Tools in Medical Research; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; REFORMS: Consensus-based Recommendations for Machine-learning-based Science; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI; STARD-AI: Standards for Reporting Diagnostic Accuracy–AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.

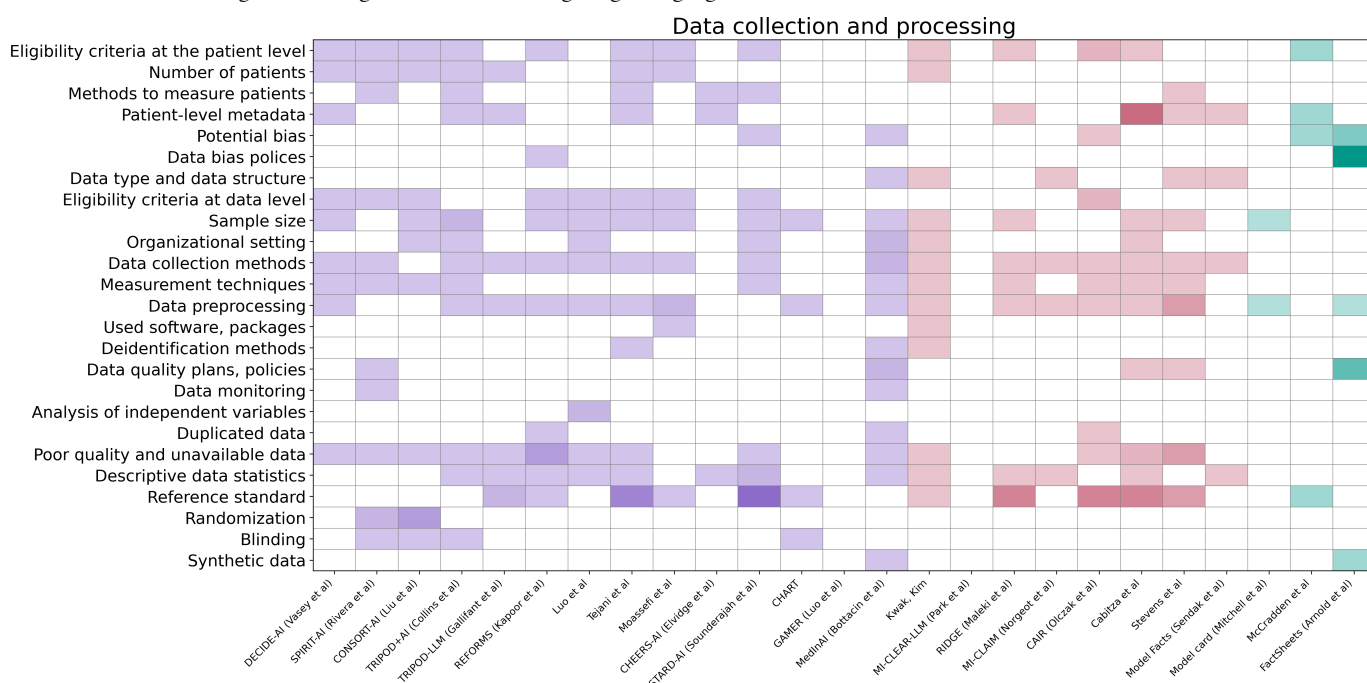


Data Collection and Processing

Finally, steps for data collection and processing were included in the last section of questions (Figure 8). This section contains patient and data description questions (s=25), including data collection and measurement techniques, as well as questions about data preprocessing and analysis, and handling of poor-quality and missing data. Only 3 papers [30,

36, 18] included requirements for deidentification methods, as it was related to medical images which require more proficient methods to deidentify personal information from DICOM (Digital Imaging and Communications in Medicine) tags. We could not distinguish between questions for clinical and technical papers.

Figure 8. Questions from the consolidated list that relate to the data collection and processing. The color code provides the information about evidence characteristics: purple – consensus-based research, red – scientific publications, green – scientific publications from technical databases. The color intensity corresponds to the number of questions addressed in each paper, with the highest intensity for the maximum number [17-27,29-41]. CAIR: Clinical AI Research; CHART: Chatbot Assessment Reporting Tool; CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI; CONSORT-AI: Consolidated Standards of Reporting Trials-AI; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; GAMER: ; MI-CLAIM: Minimum Information About Clinical AI Modeling; MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care; REFORMS: ; RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI; STARD-AI: Standards for Reporting Diagnostic Accuracy-AI; TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI; TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models.



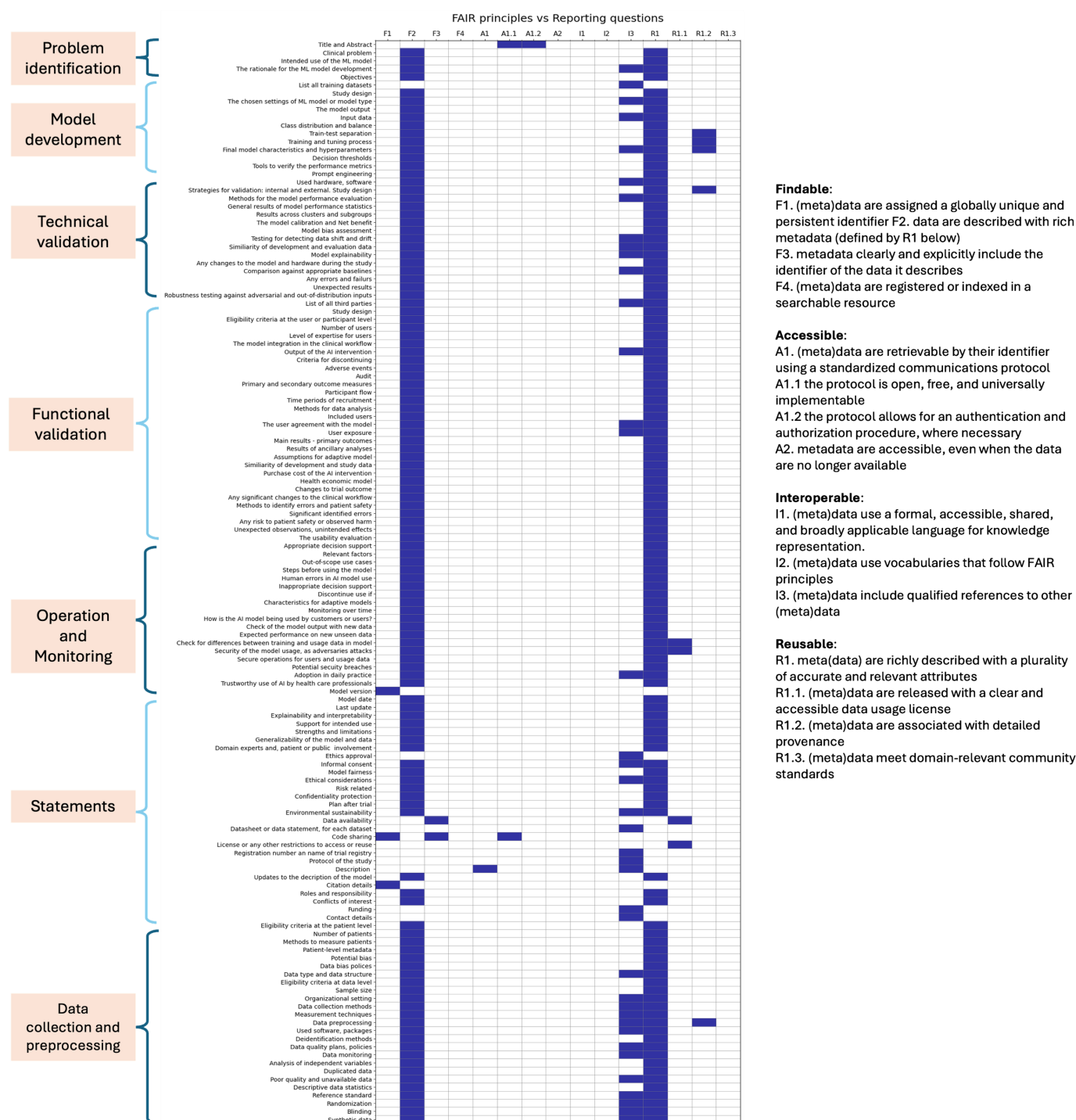
Reporting Standards and FAIR Principles

We matched the FAIR principles with the sections or phases as explained above by using the general FAIR principles published by Wilkinson et al in 2016 [7]. These recommendations cover both metadata and data (in our case, the AI model).

Figure 9 shows the connection between question groups – as discussed in the previous sections – and FAIR principles.

It was shown that only 10 out of 15 FAIR principles were addressed in the consolidated list. Multimedia Appendix 2 includes the whole list of reporting questions together with corresponding FAIR principles. In this section, we are discussing only general findings related to extracted groups of questions.

Figure 9. The mapping of reporting questions to FAIR principles. FAIR: findable, accessible, interoperable, reusable; ML: machine learning; ML: machine learning.



mention it in the list of applied FAIR principles; however, it might potentially relate to vocabularies that follow FAIR principles (I2) and meeting domain-relevant community standards (R1.3).

The group of questions named “General model statements” relates to the principles F2, I3, A2, and R1. With questions about code and dataset sharing, we assume providing the links using secure transfer protocols (A1) that are open, free, and universally implementable (A1.1). Security operations and contact details provide information about access to the model and datasets that could be aligned with the authentication (A1.2). The question about license and restriction on how to access or reuse the model relates to the R1.1 principle. Finally, items about funding, ownership, and contact details could also be associated with R1.2 as additional provenance information.

Discussion

General interpretation

The paper studied published reporting standards for AI models from both clinical and technical perspectives to extract and consolidate the complete list of reporting questions, and assess whether they aligned with FAIR principles. This systematic review analyzed 25 reporting standards for AI models. The results of our review present the complete list of reporting questions related to different stages of the lifecycle, mapped to FAIR principles. Readers could report AI studies at different lifecycle stages using a more comprehensive consolidated reporting guideline collected in this systematic review.

The previous study by Klement et al [47] showed the 37 extracted reporting items grouped into the 5 categories: study details, data description, methodology for model development, model evaluation, and explainability. The authors consolidated items from specific domains (like biology, nuclear medicine, and cardiovascular diseases) and also some not finalized checklists (such as MINIMAR [Minimum Information for Medical AI Reporting], as only a development protocol was published [48]). Since the authors included standards only from clinical sources, the final list does not incorporate details about technical validation, deployment, operation, and monitoring stages, as presented in our review. Another paper was published by Kolbinger et al [5] where the list of reporting topics was combined in 1 set and grouped into 5 categories: clinical rationale, data, model training and validation, critical appraisal, ethics, and reproducibility. We observed the overlap with our extracted list, with more detailed questions included in the model training, and a separate section for 2 types of validation: technical and functional. We also extracted a more extended list of statements that relate to critical appraisal, ethics, and reproducibility groups. The final list was also formed considering that at least 50% of all standards reported this question. We did not use any thresholds for consolidation as we wanted to see how different questions aligned with

lifecycle stages, and all reporting standards served different purposes.

Another approach was used in the paper of Shiferaw et al [49], in which the authors discussed the quality of published reporting standards and incorporated the necessary topics related to transparency, reproducibility, and ethics in medical AI research. The research identified that TRIPOD+AI, DECIDE-AI, APPRAISE-AI (Assessment of Predictive Performance and Bias in Artificial Intelligence Studies), SPIRIT-AI, and CONSORT-AI have the highest score in quality evaluation. We included these papers except for APPRAISE-AI, which matched the exclusion criteria as a paper from a specific domain. In the other review papers, Klontzas et al [50] proposed a consolidated list for medical radiological images that could be used during the review process of studies, while Garbin and Marques [51] highlighted the exact purpose by aligning the reporting standards with the AI model lifecycle stage while not proposing a consolidated list of standards. The stages used are overlapping with ours, except that it has one stage called “Product test” without distinguishing between technical and functional validations. In our opinion, this distinction is important since these validations are performed by different specialists and will produce different validation properties and results. Finally, Ibrahim et al [13] reviewed reporting standards for different purposes, such as clinical trials and diagnostic accuracy studies. This paper does not provide a consolidated list and focuses on clinical implications.

Findings According to the AI Model Lifecycle

As we consolidated the list of questions on what to report on each stage of the AI model lifecycle, we need to discuss its interpretation. The lifecycle typically initiates by defining the problem. Most of the papers ask it in a general way, for example, to describe the clinical background, the intended use of the ML model, and the rationale for the ML model development. Luo et al [23] included a separate section describing the setting of the study and the prediction problem. These questions could provide comprehensive information about the intentions behind the proposed ML model development and propagate into specific inclusion and exclusion criteria for model use cases. Most of the included standards stated the necessity to report this information.

The group “Data collection and management” includes a broad spectrum of questions related to patient and data eligibility criteria, as well as methods for collection and measurement. These questions can be repeated for different stages of the AI lifecycle – for training, technical, and functional validation. Data collection was covered by most included standards and highlighted the importance of these questions for all stages of the lifecycle, which is characterized using datasets. Previously, a datasheet was developed to elicit the information that a dataset might contain [52]. This proposal also matched the key stages of the dataset lifecycle. In our work, we did not compare questions from the consolidated list with the datasheet, as it was not included by our eligibility criteria. Although we can expect a considerable

overlap between items. Arnold et al mentioned that the datasheet found a significant reflection in the Factsheet [41], one of the included papers. The necessity to report datasheet or data statements was also included in papers [22,34,39].

There is no guidance in the literature specific to model development descriptions. However, papers usually describe both development and performance evaluation. The consolidated list results in the combination of main topics covering the general questions for model training, while the specific details could be found in the whole list (see Multimedia Appendix 2).

The technical validation is usually conducted by the AI model developers or providers. This procedure aims to assess the technical performance of the developed model, including model performance, reliability, robustness, and safety. The AI model is usually presented as a prototype or preproduction model before it is ready for deployment. Functional validation is typically performed as clinical trials with AI interventions and aims to assess the real-world performance, safety, usability, and relevance of the AI model in clinical use. Hence, a functional evaluation not only evaluates the AI model, but the system in which the AI model is integrated. This step could be referred to the clinical validation, a part of the AI model clinical evaluation process [53]. Before the real deployment of AI models, it is important to get evidence about the model's safety, effectiveness, and efficiency. For this stage, the AI model should be presented as a system in an executable and ready-to-use format, so we matched this process with a deployment stage. Technical and clinical validations are crucial steps of the AI model lifecycle, and the results of this review presented the comprehensive list of reporting topics.

The section with statements could be divided into several subgroups of questions: additional information on data and model fairness; ethical considerations with protocol and registration, description of collecting the informal consent and confidential protections; contact and general details on providers, funding, and responsibilities; question to provide accessibility and reusability of the model: as code and datasets sharing, datasheet; general and technical details about model operations and monitoring. These questions correspond to the parts of critical appraisal, ethics, and reproducibility detected in [5].

Findings Related to FAIR Principles

FAIR principles were first introduced by Wilkinson [7]. The Research Data Alliance (RDA) Data Maturity Model working group presented indicators for the FAIR assessment [16], based on the development of FAIR assessment metrics for research data [54]. The separate research group worked on the FAIR indicators for research software [55]. Furthermore, (FAIR principles for machine learning) FAIR4ML was proposed for software with machine learning [56].

Our results showed that only 9 out of 15 FAIR principles were addressed in the consolidated list. However, some papers discuss FAIR-related requirements in the description of reports. For example, the first version of the CLAIM

guideline required defining the data elements [45]. Based on this reporting standard, common data elements are used by the radiological community to describe used data: a unique identifier, human-readable definition, the possible values for a particular data element, links to other ontologies (such as [Systematized Nomenclature of Medicine – Clinical Terms] SNOMED CT, [Radiology lexicon] RadLex), as well as versioning data [57,58]. These questions might meet domain-relevant community standards (R1.3) and use vocabularies that follow FAIR principles (I2). However, these questions were excluded from the recent version of the CLAIM guideline [18].

The most common questions throughout all included papers, related to FAIR principles, were about code and data availability. While it is clear for the datasets sharing, the question about code sharing was presented in different ways. Some standards require authors to deposit all code used for modeling and/or data analysis into a publicly accessible repository [18], including analytical code [17], code used to train and evaluate the model to produce all results reported in the paper [22], and access to the AI model itself [32]. Other authors mention just reproducibility [34] or state whether and how the AI intervention and/or its code can be accessed [20, 21]. All in all, from this review, we see little overlap between FAIR principles and reporting standards for AI models. Our suggestion would be to apply FAIR principles (where the data is the actual model, and metadata are descriptions of the model) to AI algorithms.

Limitations

Our systematic review has some limitations related to the evidence included. The included papers contain a few eligible scientific publications from technical databases. These papers extended the consolidated list of reporting questions with detailed information; for example, Arnold et al [41] published an extensive list of questions related to the “Statements” group. These questions might be less useful for some ML models, but we aimed to include all extracted information in the consolidated list to present the necessary aspects for reporting along the lifecycle. However, the list combined by different lifecycle stages cannot be seen as a replacement for the already existing reporting standards that are required for drafting papers and review processes.

We aligned the list with FAIR principles proposed by Wilkinson et al in 2016 [7]. In the field, there are also other known standards, such as FAIR principles for research software (FAIR4RS) [59]. However, this list does not fully include requirements for metadata, mentioning only its findability and accessibility. The specific work was done for datasets by RDA [16], aiming to adopt the general FAIR principles in the FAIR data maturity model. Another work is FAIR4ML models [56]. Both FAIR4RS as well as FAIR4ML cover principles for metadata and data (datasets or AI models) in an extended way. However, papers were published in the open-source repositories without peer review. Therefore, we decided to base our comparison on the initial FAIR principles.

During our review process, 2 authors (EA and DS) read full-text papers after the screening, including the third expert (JvS) to reach consensus, while only 1 author (EA) performed the screening in AsReview. The search was also performed by 1 author (EA); however, our search strategy was developed by 3 of the authors (EA, DS, JvS), and it was reviewed and adjusted by a librarian from the host institution. The extraction of information was performed by 1 author (EA) and presented for discussion at several group meetings and at a conference [60]. Because some of the review steps were not conducted dually and independently, we introduced some risk of error in the consolidated list of questions. Furthermore, we used the stopping criteria of 50 papers for the AsReview screening, which is enough for collecting the most relevant papers. Using this probabilistic screening may have resulted in missing a few publications. However, since our aim was not to perform a quantitative systematic review, the omission of a few publications would not influence our conclusion.

We did not evaluate the quality of reporting standards as it was done in Klement et al [47]. Some authors applied 2 following checklists: AGREE II instrument (Appraisal of Guidelines for Research and Evaluation II) [61] and RIGHT (Reporting Items for Practice Guidelines in Health care) [62], which are aiming to evaluate publications and preparing papers. We did not specify quality assessment as our main goal. Firstly, some papers were extracted from nonclinical databases where AGREE II and RIGHT were not applicable because they are designed to assess and report clinical practice guidelines. Secondly, previously published reviews [47,49] performed this quality assessment according to these 2 checklists for TRIPOD+AI, DECIDE-AI, CONSORT-AI, SPIRIT-AI, MI-CLAIM (Minimum Information About Clinical AI Modeling), STARD. Nevertheless, we used 3 evidence levels and removed low-quality papers according to the eligibility criteria.

In addition, we limited publication to the English language as it is a general publication language for standards. Furthermore, these English language papers should be open access and/or Institutional Access, which excluded technical documents (like ISO [International Organization for Standardization] standards) and regulatory papers such as the AI Act [63]. Lastly, we included only general papers regardless of specific domain (such as radiomics, biology, or pathology). Although we might miss some papers with additional questions, these likely only add details of already included questions.

Future Implications

The extracted and consolidated list of reporting questions for each AI model lifecycle stage offers significant benefits to developers, deployers, potential users, and other stakeholders. Firstly, the consolidated list summarizes all questions from

different published reporting standards. It allows for an overview of all aspects that should be reported at different stages of the AI lifecycle. Therefore, it provides transparency of lifecycle stages and limits possible challenges in later stages [51]. This aspect opens the perspective of developing the tooling to provide necessary information at the certain stage, as it was proposed by Brereton et al [64]. Future work would be to identify correspondence between our consolidated list and requirements from technical (eg, ISO standards) and regulatory (eg, the Medical Device Regulation regulation, AI Act) documents. For example, the AI Act proposes the minimal content to describe general characteristics of the AI system [63]. By using the consolidated list as AI model metadata, developers and deployers can disseminate AI models more easily than with a plain-text description [65]. Potential users could meet their needs in choosing AI models for their specific clinical tasks. However, it should be noted that the value of sharing should outweigh the effort required to create metadata. For this purpose, LLM agents could help us to create structured metadata from published papers with an AI model description. The future implications of our results should also meet the criteria for reproducibility and reusability. For that purpose, we included research questions related to FAIR principles.

The importance of adopting FAIR principles for AI has been raised before [66]. To address one of the basic aspects of the FAIR principles, metadata should be machine-readable. The Mobilized Computable Biomedical Knowledge community already proposed to use RDF and Semantic Web standards [65]. Furthermore, Adhikari et al [67] proposed to use the Web Ontology Language for explainable AI solutions by providing rich, machine-readable metadata for AI system descriptions (types of input data, model type, and task type) and parameters of use cases. We hypothesize that incorporation of FAIR principles into the consolidated list from this work will help to develop a comprehensive “digital object” description and machine actionable AI model, to speed-up the turnaround time from data analysis and AI development towards clinical practice.

Conclusion

Our study provides the complete consolidated list of reporting questions for AI models in health care. We extracted 134 questions along the lifecycle stages as problem identification, model development, technical and functional validation, operation and monitoring, and general statements. Results showed that only 9 out of 15 general FAIR principles were addressed in the consolidated list. This consolidated list could be used to create AI model metadata to emphasize the sharing of AI models. Incorporating FAIR principles in AI models' metadata may improve their reproducibility and reusability.

Acknowledgments

Authors are grateful to Dr Marieke Schor for her support in creating an optimal search strategy. During the preparation of this manuscript, authors used an AI tool for grammar checks. Authors did not use Gen AI in preparation for this paper. One AI tool (Grammarly) was used only for grammar check and refinement (AI Activity N1 from ChatGPT).

Funding

This study was financially supported by the NWO Perspectief PersOn project (project no. P21-03), and European Union's Horizon Europe research and innovation programme under grant agreement No 101136262 (project name: BETTER).

Authors' Contributions

Conceptualization: AD (lead), JvS (equal)

Data curation: EA (lead), DS (supporting)

Formal analysis: EA (lead), DS (equal)

Funding acquisition: AD (lead), JvS (supporting)

Investigation: EA (lead), DS (equal), JvS (supporting)

Methodology: EA (lead), DS (supporting), ALG (supporting), SM (supporting), JvS (supporting)

Project administration: JvS

Resources: AD (lead), JvS (equal)

Supervision: AD (lead), JvS (equal), ALG (equal)

Validation: JvS (lead), ALG (equal)

Visualization: EA

Writing – original draft: EA (lead)

Writing – review & editing: DS (supporting), ALG (supporting), SM (supporting), AD (equal), JvS (lead)

Conflicts of Interest

JvS and AD are shareholders and employees receiving salary of Medical Data Works B.V.; although company activities are unrelated to this manuscript. No mitigation methods were implemented during this study.

Multimedia Appendix 1

Search strategy.

[\[DOCX File \(Microsoft Word File\), 24 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

The consolidated table with reporting questions.

[\[XLSX File \(Microsoft Excel File\), 137 KB-Multimedia Appendix 2\]](#)

Checklist 1

PRISMA 2020 checklist.

[\[DOCX File \(Microsoft Word File\), 284 KB-Checklist 1\]](#)

References

1. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: data, validation, communication challenges. *Med Phys*. Oct 2018;45(10):e834-e840. [doi: [10.1002/mp.12811](https://doi.org/10.1002/mp.12811)] [Medline: [30144098](https://pubmed.ncbi.nlm.nih.gov/30144098/)]
2. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten M. Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022. *Eur Radiol*. Jan 2024;34(1):348-354. [doi: [10.1007/s00330-023-09991-5](https://doi.org/10.1007/s00330-023-09991-5)] [Medline: [37515632](https://pubmed.ncbi.nlm.nih.gov/37515632/)]
3. The AI lifecycle expanded to illustrate per-phase considerations. FDA. URL: <https://www.fda.gov/media/180310/download?attachment> [Accessed 2026-09-23]
4. What is a reporting guideline? Equator Network. URL: <https://www.equator-network.org/about-us/what-is-a-reporting-guideline/> [Accessed 2025-08-08]
5. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)*. Apr 11, 2024;4(1):71. [doi: [10.1038/s43856-024-00492-0](https://doi.org/10.1038/s43856-024-00492-0)] [Medline: [38605106](https://pubmed.ncbi.nlm.nih.gov/38605106/)]
6. Ravi N, Chaturvedi P, Huerta EA, et al. FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. *Sci Data*. Nov 10, 2022;9(1):657. [doi: [10.1038/s41597-022-01712-9](https://doi.org/10.1038/s41597-022-01712-9)] [Medline: [36357431](https://pubmed.ncbi.nlm.nih.gov/36357431/)]
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. Mar 15, 2016;3(1):160018. [doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)] [Medline: [26978244](https://pubmed.ncbi.nlm.nih.gov/26978244/)]

8. Shiferaw KB, Balaur I, Welter D, Waltemath D, Zeleke AA. CALIFRAME: a proposed method of calibrating reporting guidelines with FAIR principles to foster reproducibility of AI research in medicine. *JAMIA Open*. Dec 2024;7(4):ooae105. [doi: [10.1093/jamiaopen/ooae105](https://doi.org/10.1093/jamiaopen/ooae105)] [Medline: [39430802](#)]
9. Klontzas ME, Gatti AA, Tejani AS, Kahn CE. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell*. May 2023;5(3):e230055. [doi: [10.1148/ryai.230055](https://doi.org/10.1148/ryai.230055)] [Medline: [37293341](#)]
10. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol*. Jul 2021;49(5):470-476. [doi: [10.1111/ceo.13943](https://doi.org/10.1111/ceo.13943)] [Medline: [33956386](#)]
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Mar 29, 2021;372:n71. [doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](#)]
12. Akhmad E. Data for MIE 2025. OSF. Apr 15, 2026. URL: <https://osf.io/fmgza/> [Accessed 2026-10-01]
13. Ibrahim H, Liu X, Rivera SC, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*. Jan 6, 2021;22(1):11. [doi: [10.1186/s13063-020-04951-6](https://doi.org/10.1186/s13063-020-04951-6)] [Medline: [33407780](#)]
14. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. *jmla*. Sep 12, 2016;104(3):240-243. [doi: [10.5195/jmla.2016.24](https://doi.org/10.5195/jmla.2016.24)]
15. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell*. Feb 2021;3(2):125-133. [doi: [10.1038/s42256-020-00287-7](https://doi.org/10.1038/s42256-020-00287-7)]
16. FAIR Data maturity model. Specification and guidelines. Zenodo. Jun 25, 2020. URL: <https://zenodo.org/records/3909563> [Accessed 2026-09-23]
17. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. Apr 16, 2024;385:e078378. [doi: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378)] [Medline: [38626948](#)]
18. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. Jul 2024;6(4):e240300. [doi: [10.1148/ryai.240300](https://doi.org/10.1148/ryai.240300)] [Medline: [38809149](#)]
19. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. May 18, 2022;377:e070904. [doi: [10.1136/bmj-2022-070904](https://doi.org/10.1136/bmj-2022-070904)] [Medline: [35584845](#)]
20. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. Sep 2020;26(9):1351-1363. [doi: [10.1038/s41591-020-1037-7](https://doi.org/10.1038/s41591-020-1037-7)] [Medline: [32908284](#)]
21. Liu X, Rivera SC, 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. *BMJ*. Sep 9, 2020;370:m3164. [doi: [10.1136/bmj.m3164](https://doi.org/10.1136/bmj.m3164)] [Medline: [32909959](#)]
22. Kapoor S, Cantrell EM, Peng K, et al. REFORMS: consensus-based recommendations for machine-learning-based science. *Sci Adv*. May 3, 2024;10(18):eadk3452. [doi: [10.1126/sciadv.adk3452](https://doi.org/10.1126/sciadv.adk3452)] [Medline: [38691601](#)]
23. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. Dec 16, 2016;18(12):e323. [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](#)]
24. Moassefi M, Singh Y, Conte GM, et al. Checklist for reproducibility of deep learning in medical imaging. *J Imaging Inform Med*. Aug 2024;37(4):1664-1673. [doi: [10.1007/s10278-024-01065-2](https://doi.org/10.1007/s10278-024-01065-2)] [Medline: [38483694](#)]
25. Elvidge J, Hawksworth C, Avşar TS, et al. Consolidated health economic evaluation reporting standards for interventions that use artificial intelligence (CHEERS-AI). *Value Health*. Sep 2024;27(9):1196-1205. [doi: [10.1016/j.jval.2024.05.006](https://doi.org/10.1016/j.jval.2024.05.006)] [Medline: [38795956](#)]
26. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. Jan 2025;31(1):60-69. [doi: [10.1038/s41591-024-03425-5](https://doi.org/10.1038/s41591-024-03425-5)] [Medline: [39779929](#)]
27. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. Oct 2025;31(10):3283-3289. [doi: [10.1038/s41591-025-03953-8](https://doi.org/10.1038/s41591-025-03953-8)] [Medline: [40954311](#)]
28. CHART Collaborative. Reporting guideline for chatbot health advice studies: Chatbot Assessment Reporting Tool (CHART) Statement. *Ann Fam Med*. Sep 2025;23(5):389-398. [doi: [10.1370/afm.250386](https://doi.org/10.1370/afm.250386)]
29. Luo X, Tham YC, Giuffrè M, et al. Reporting guideline for the use of generative artificial intelligence tools in medical research: the GAMER statement. *BMJ Evid Based Med*. Dec 1, 2025;30(6):390-400. [doi: [10.1136/bmjebm-2025-113825](https://doi.org/10.1136/bmjebm-2025-113825)] [Medline: [40360239](#)]
30. Bottacin WE, de Souza TT, Melchioris AC, Reis WCT. Explanation and elaboration of MedinAI: guidelines for reporting artificial intelligence studies in medicines, pharmacotherapy, and pharmaceutical services. *Int J Clin Pharm*. Aug 2025;47(4):957-969. [doi: [10.1007/s11096-025-01906-2](https://doi.org/10.1007/s11096-025-01906-2)] [Medline: [40249526](#)]

31. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. Sep 2020;26(9):1320-1324. [doi: [10.1038/s41591-020-1041-y](https://doi.org/10.1038/s41591-020-1041-y)] [Medline: [32908275](#)]
32. Olczak J, Pavlopoulos J, Prijs 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*. Oct 2021;92(5):513-525. [doi: [10.1080/17453674.2021.1918389](https://doi.org/10.1080/17453674.2021.1918389)] [Medline: [33988081](#)]
33. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform*. Sep 2021;153:104510. [doi: [10.1016/j.ijmedinf.2021.104510](https://doi.org/10.1016/j.ijmedinf.2021.104510)] [Medline: [34108105](#)]
34. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes*. Oct 2020;13(10):e006556. [doi: [10.1161/CIRCOUTCOMES.120.006556](https://doi.org/10.1161/CIRCOUTCOMES.120.006556)] [Medline: [33079589](#)]
35. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. 2020;3(1):41. [doi: [10.1038/s41746-020-0253-3](https://doi.org/10.1038/s41746-020-0253-3)] [Medline: [32219182](#)]
36. Kwak SG, Kim J. Comprehensive reporting guidelines and checklist for studies developing and utilizing artificial intelligence models. *Korean J Anesthesiol*. Jun 2025;78(3):199-214. [doi: [10.4097/kja.25075](https://doi.org/10.4097/kja.25075)] [Medline: [40468627](#)]
37. Park SH, Suh CH, Lee JH, et al. Minimum reporting items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM): 2025 updates. *Korean J Radiol*. Dec 2025;26(12):1123-1132. [doi: [10.3348/kjr.2025.1522](https://doi.org/10.3348/kjr.2025.1522)] [Medline: [41199132](#)]
38. Maleki F, Moy L, Forghani R, et al. RIDGE: reproducibility, integrity, dependability, generalizability, and efficiency assessment of medical image segmentation models. *J Digit Imaging Inform med*. 2024;38(4):2524-2536. [doi: [10.1007/s10278-024-01282-9](https://doi.org/10.1007/s10278-024-01282-9)]
39. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2019:220-229. [doi: [10.1145/3287560.3287596](https://doi.org/10.1145/3287560.3287596)]
40. Mccradden M, Odusi O, Joshi S, et al. What's fair is... fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning. In: *FAccT '23: The 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2023:1505-1519. [doi: [10.1145/3593013.3594096](https://doi.org/10.1145/3593013.3594096)]
41. Arnold M, Bellamy RKE, Hind M, et al. FactSheets: increasing trust in AI services through supplier's declarations of conformity. *arXiv*. Preprint posted online on Feb 7, 2019. URL: <http://arxiv.org/abs/1808.07261> [Accessed 2024-09-23] [doi: [10.1147/JRD.2019.2942288](https://doi.org/10.1147/JRD.2019.2942288)]
42. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ*. Feb 7, 2023;380:e071058. [doi: [10.1136/bmj-2022-071058](https://doi.org/10.1136/bmj-2022-071058)] [Medline: [36750236](#)]
43. Heus P, Reitsma JB, Collins GS, et al. Transparent reporting of multivariable prediction models in journal and conference abstracts: TRIPOD for abstracts. *Ann Intern Med*. Jun 2, 2020;173(1). [doi: [10.7326/M20-0193](https://doi.org/10.7326/M20-0193)] [Medline: [32479165](#)]
44. Snell KIE, Levis B, Damen JAA, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ*. May 3, 2023;381:e073538. [doi: [10.1136/bmj-2022-073538](https://doi.org/10.1136/bmj-2022-073538)] [Medline: [37137496](#)]
45. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. Mar 2020;2(2):e200029. [doi: [10.1148/ryai.2020200029](https://doi.org/10.1148/ryai.2020200029)] [Medline: [33937821](#)]
46. Lennerz JK, Salgado R, Kim GE, et al. Diagnostic quality model (DQM): an integrated framework for the assessment of diagnostic quality when using AI/ML. *Clin Chem Lab Med*. Jan 25, 2023;61(4):544-557. [doi: [10.1515/ccim-2022-1151](https://doi.org/10.1515/ccim-2022-1151)]
47. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res*. Aug 31, 2023;25:e48763. [doi: [10.2196/48763](https://doi.org/10.2196/48763)] [Medline: [37651179](#)]
48. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. Dec 9, 2020;27(12):2011-2015. [doi: [10.1093/jamia/ocaa088](https://doi.org/10.1093/jamia/ocaa088)] [Medline: [32594179](#)]
49. Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review. *JAMIA Open*. Feb 2025;8(1):ooae155. [doi: [10.1093/jamiaopen/ooae155](https://doi.org/10.1093/jamiaopen/ooae155)] [Medline: [39759773](#)]
50. Klontzas ME, Gatti AA, Tejani AS, Kahn CEJ. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell*. May 2023;5(3):e230055. [doi: [10.1148/ryai.230055](https://doi.org/10.1148/ryai.230055)] [Medline: [37293341](#)]

51. Garbin C, Marques O. Assessing methods and tools to improve reporting, increase transparency, and reduce failures in machine learning applications in health care. *Radiol Artif Intell*. Mar 2022;4(2):e210127. [doi: [10.1148/ryai.210127](https://doi.org/10.1148/ryai.210127)] [Medline: [35391771](https://pubmed.ncbi.nlm.nih.gov/35391771/)]
52. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM*. Dec 2021;64(12):86-92. [doi: [10.1145/3458723](https://doi.org/10.1145/3458723)]
53. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digit Med*. 2020;3(1):55. [doi: [10.1038/s41746-020-0260-4](https://doi.org/10.1038/s41746-020-0260-4)] [Medline: [32337371](https://pubmed.ncbi.nlm.nih.gov/32337371/)]
54. Devaraju A, Huber R, Mokrane M, et al. FAIRsFAIR Data object assessment metrics. Zenodo. Preprint posted online on Apr 14, 2022. [doi: [10.5281/ZENODO.3934401](https://doi.org/10.5281/ZENODO.3934401)]
55. Hong NPC, Katz DS, Barker M, et al. FAIR Principles for research software (FAIR4RS principles). Zenodo. Preprint posted online on Mar 24, 2022. URL: <https://zenodo.org/records/6623556> [Accessed 2026-09-23] [doi: [10.15497/RDA00065](https://doi.org/10.15497/RDA00065)]
56. Limani F, Tofik L, Latif A, Tochtermann K. FAIR for machine learning model: principles and assessment metrics. Zenodo; Sep 24, 2024. URL: <https://zenodo.org/records/13835105> [Accessed 2025-01-31]
57. Kohli M, Alkasab T, Wang K, et al. Bending the artificial intelligence curve for radiology: informatics tools from ACR and RSNA. *J Am Coll Radiol*. Oct 2019;16(10):1464-1470. [doi: [10.1016/j.jacr.2019.06.009](https://doi.org/10.1016/j.jacr.2019.06.009)] [Medline: [31319078](https://pubmed.ncbi.nlm.nih.gov/31319078/)]
58. RDES126 - Acute aortic syndrome. RadElement. URL: <https://radelement.org/home/sets/set/RDES126> [Accessed 2025-04-01]
59. Barker M, Chue Hong NP, Katz DS, et al. Introducing the FAIR principles for research software. *Sci Data*. Oct 14, 2022;9(1):622. [doi: [10.1038/s41597-022-01710-x](https://doi.org/10.1038/s41597-022-01710-x)] [Medline: [36241754](https://pubmed.ncbi.nlm.nih.gov/36241754/)]
60. Akhmad E, Slob D, Lobo Gomes A, Dekker A, van Soest J. What to report? a systematic review of medical AI reporting guidelines: preliminary results. *Stud Health Technol Inform*. May 15, 2025;327:390-391. [doi: [10.3233/SHTI250351](https://doi.org/10.3233/SHTI250351)] [Medline: [40380462](https://pubmed.ncbi.nlm.nih.gov/40380462/)]
61. Brouwers MC, Kho ME, Browman GP, et al. AGREE II: Advancing guideline development, reporting and evaluation in health care. *Can Med Assoc J*. Dec 14, 2010;182(18):E839-E842. [doi: [10.1503/cmaj.090449](https://doi.org/10.1503/cmaj.090449)]
62. Chen Y, Yang K, Marušić A, et al. A reporting tool for practice guidelines in health care: the RIGHT statement. *Ann Intern Med*. Jan 17, 2017;166(2):128-132. [doi: [10.7326/M16-1565](https://doi.org/10.7326/M16-1565)] [Medline: [27893062](https://pubmed.ncbi.nlm.nih.gov/27893062/)]
63. Annex IV. AI Act. URL: <https://artificialintelligenceact.eu/annex/4/> [Accessed 2026-09-29]
64. Brereton TA, Malik MM, Rost LM, et al. AImedReport: a prototype tool to facilitate research reporting and translation of artificial intelligence technologies in health care. medRxiv. Preprint posted online on Mar 29, 2024. URL: <https://www.medrxiv.org/content/10.1101/2024.01.16.24301358v2> [Accessed 2026-09-23] [doi: [10.1016/j.mcpdig.2024.03.008](https://doi.org/10.1016/j.mcpdig.2024.03.008)]
65. Alper BS, Flynn A, Bray BE, et al. Categorizing metadata to help mobilize computable biomedical knowledge. *Learn Health Syst*. May 9, 2022;6(1):e10271. [doi: [10.1002/lrh2.10271](https://doi.org/10.1002/lrh2.10271)] [Medline: [35036552](https://pubmed.ncbi.nlm.nih.gov/35036552/)]
66. Huerta EA, Blaiszik B, Brinson LC, et al. FAIR for AI: an interdisciplinary and international community building perspective. *Sci Data*. Jul 26, 2023;10(1):487. [doi: [10.1038/s41597-023-02298-6](https://doi.org/10.1038/s41597-023-02298-6)] [Medline: [37495591](https://pubmed.ncbi.nlm.nih.gov/37495591/)]
67. Adhikari A, Wenink E, van der Waa J, Bouter C, Tolios I, Raaijmakers S. Towards FAIR Explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. In: PETRA '22: The 15th International Conference on Pervasive Technologies Related to Assistive Environments. Association for Computing Machinery; 2022:562-568. [doi: [10.1145/3529190.3535693](https://doi.org/10.1145/3529190.3535693)]

Abbreviations

AGREE II: Appraisal of Guidelines for Research and Evaluation II

AsReview: Automated Systematic Review software

BERT: Bidirectional Encoder Representations from Transformers

CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI

CLAIM: Clinical AI Modeling

CONSORT-AI: Consolidated Standards of Reporting Trials-AI

DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI

DICOM: Digital Imaging and Communications in Medicine

FAIR: findable, accessible, interoperable, reusable

FAIR4ML: FAIR principles for machine learning

FAIR4RS: FAIR principles for research software

FDA: Food and Drug Administration

GAMER: Generative Artificial Intelligence Tools in Medical Research

ISO: International Organization for Standardization

LLM: large language model

MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modeling

MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health care

ML: machine learning

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RadLex: radiology lexicon

RDA: Research Data Alliance

REFORMS: Consensus-based Recommendations for Machine-learning-based Science

RIGHT: Appraisal of Guidelines for Research and Evaluation II

SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–AI

STARD: Standards for Reporting Diagnostic Accuracy

TF-IDF: term frequency-inverse document frequency

TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI

TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for studies using Large Language Models

TRIPOD-SRMA: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Systematic Reviews and Meta Analysis

Edited by Ivan Steenstra; peer-reviewed by Milton Campoverde-Molina, Sadhasivam Mohanadas; submitted 28.Jan.2026; final revised version received 06.Jul.2026; accepted 06.Jul.2026; published 09.Oct.2026

Please cite as:

Akhmad E, Slob D, Gomes AL, Mingels S, Dekker A, Soest J

AI Model Reporting Standards in Health Care: Systematic Review and Consolidation

JMIR AI 2026;5:e92329

URL: <https://ai.jmir.org/2026/1/e92329>

doi: [10.2196/92329](https://doi.org/10.2196/92329)

© Ekaterina Akhmad, Daniël Slob, Aiara Lobo Gomes, Simone Mingels, Andre Dekker, Johan van Soest. Originally published in JMIR AI (<https://ai.jmir.org>), 09.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.