Review

Evaluation Frameworks for Clinical AI Incorporating Validation Strategies, Real-World Applicability, and Ethical Principles: Scoping Review

Diana Carolina López Medina[1*], MSc, MD, PhD; Aida Oliveros-Navarro[1*], MSc, MD; Nataly Moreno Angel[2*], MSc, MD; Andrés Camilo Herrera- Arellano[3*], MD; Marcela Henao-Pérez[1*], MD, PhD

[1]Research Grupo Infettare, School of Medicine, Universidad Cooperativa Colombia, Medellín, Antioquia, Colombia
[2]Internal Medicine Residency Program, School of Medicine, Universidad Cooperativa de Colombia, Medellín, Colombia
[3]Medicine Program, School of Medicine, Universidad Cooperativa de Colombia, Medellín, Colombia
*all authors contributed equally

Corresponding Author:

Diana Carolina López Medina, MSc, MD, PhD
Research Grupo Infettare
School of Medicine, Universidad Cooperativa Colombia
Calle 50, No 40-74, Bloque A
Medellín, Antioquia 050012
Colombia
Phone: 57 3108406678
Email: diana.lopezme@campusucc.edu.co

Abstract

Background: AI shows substantial potential in health care; however, the absence of standardized evaluation frameworks limits its safe and effective clinical implementation because of inconsistent validation requirements and fragmented ethical principles. Existing guidelines vary in structure, methodological rigor, and ethical integration, creating uncertainty.

Objective: This study aimed to systematically map, characterize, and critically analyze existing evaluation frameworks for clinical AI, focusing on three core dimensions: methodological rigor, validation strategies (internal validation, including reporting of technical and clinical performance; external validation, including real-world applicability), and alignment with the United Nations Educational, Scientific and Cultural Organization (UNESCO) AI ethical considerations.

Methods: A scoping review was conducted following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Six databases (PubMed, Embase, BVS, EBSCOhost, ProQuest, and Sage) and the Enhancing the Quality and Transparency of Health Research Network were searched without language or date restrictions up to February 2026. Eligible documents included peer-reviewed papers, gray literature, and organizational guidelines describing evaluation or reporting frameworks for clinical AI. Editorials, commentaries, and conference abstracts lacking a clearly defined evaluative framework or clinical applicability were excluded. Two reviewers independently screened records and extracted data. Data were extracted across three domains: (1) general characteristics, (2) methodological rigor and validation parameters, and (3) ethical integration and were synthesized using a dot plot–based gap map. Ethical adherence was assessed using a 10-domain UNESCO-based scoring matrix. No formal risk-of-bias assessment was conducted, consistent with scoping review methodology.

Results: From 3363 records, 46 frameworks met the inclusion criteria. Mapping revealed a rapidly expanding but fragmented landscape. Most frameworks targeted investigational use (88%), with limited focus on clinical applicability. Frameworks varied in structure, methodology, and scope, with a predominance of reporting guidelines and few validated tools. Most (63%) were developed through multi-institutional collaborations, and 32.6% incorporated transdisciplinary participation. Only 31.8% reported technical metrics (commonly area under the curve, sensitivity, and specificity), and 15.9% provided clinical indicators (eg, predictive values or calibration). Only 11.4% achieved methodological rigor, incorporating validation aligned with intended use, while most relied on partial validation strategies, highlighting a gap between model development and clinical evaluation. Ethical integration was heterogeneous: only 5 frameworks achieved high compliance (≥80%), whereas 4 scored <10%. The most frequently addressed UNESCO principles were awareness and education (71.1%) and transparency

and explainability (70%), while human oversight (24.4%) and adaptive governance (33.3%) were least represented. Findings indicate a misalignment between framework design, validation requirements, and clinical implementation.

Conclusions: Evaluation frameworks for clinical AI remain heterogeneous and oriented toward investigational contexts. Critical gaps persist in methodological rigor, validation aligned with intended use, and fragmented ethical coverage. These findings highlight the need for standardized, robust, and ethically grounded frameworks to enable safe, reliable, and scalable integration of AI into clinical practice.

Trial Registration: PROSPERO CRD420251019640; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251019640

J Med Internet Res 2026;28:e78168; doi: 10.2196/78168

Keywords: artificial intelligence; clinical decision support systems; biomedical technology assessment; bioethics; validation studies as topic; scoping reviews; machine learning; evidence-based practice

Introduction

AI has evolved substantially since its conceptualization in the mid-20th century, when early research explored the possibility of machines emulating human cognitive functions. Driven by advances in computational power, algorithm development, and access to large-scale datasets, AI has expanded from theoretical domains to widespread applications, including health care. Current uses include medical image analysis, disease risk prediction, drug discovery, and treatment personalization [1-3].

In recent years, the number of AI applications proposed for health care has increased rapidly, with hundreds of models developed for diagnostic support, risk prediction, and workflow optimization. However, only a limited proportion of these models have undergone rigorous external validation or prospective clinical evaluation, raising concerns about their readiness for real-world clinical implementation [4-6].

Despite these advances, several challenges continue to limit the integration of AI into clinical practice [7]. These include variability in data quality, limited external validation, lack of transparency in algorithm development, and insufficient assessment of clinical impact in real-world settings [8]. Ensuring accuracy, safety, and reproducibility, while addressing ethical and regulatory requirements, remains essential. The absence of standardized validation procedures and reporting practices represents a major barrier to reliable clinical adoption.

In response, multiple guidelines and evaluation frameworks have been developed to improve the quality, transparency, and clinical applicability of AI models [9]. This lack of standardized evaluation approaches has prompted increasing concern among clinicians, regulators, and researchers regarding the reproducibility, safety, and transparency of AI-based clinical tools [10].

Notable initiatives include the TRIPOD+AI (Transparent Reporting of Multivariable Prediction Models with Artificial Intelligence) [11], the CONSORT-AI (Consolidated Standards of Reporting Trials for Artificial Intelligence) [12], and the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) [13]. More recently, additional initiatives such as STARD-AI (Standards for Reporting Diagnostic Accuracy) [14] and PRISMA-AI (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Artificial Intelligence) [15] have further expanded reporting standards for diagnostic accuracy studies and systematic reviews involving AI technologies. These frameworks provide structured approaches for assessing data quality, methodological rigor, performance metrics, and clinical utility [16].

However, substantial heterogeneity persists across these initiatives, particularly in methodological scope, validation strategies, and the integration of ethical considerations. This variability makes it difficult for researchers and health care systems to identify appropriate approaches for evaluating AI systems intended for clinical use [17,18].

The objective of this scoping review was to systematically map, characterize, and critically analyze existing evaluation and reporting frameworks for clinical AI. Specifically, this review aimed to examine how these frameworks conceptualize and operationalize: (1) methodological rigor and validation strategies, including internal and external validation; (2) claims related to clinical applicability and real-world use; and (3) the integration of ethical principles, with particular emphasis on alignment with the 10 ethical domains proposed by the United Nations Educational, Scientific and Cultural Organization (UNESCO).

Methods

Protocol and Registration

We conducted a scoping review incorporating a quantitative documentary analysis to examine evaluation frameworks for AI models in clinical practice. The methodology was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [19], the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 for Abstracts Checklist [20], and was registered in PROSPERO (CRD420251019640).

The process involved the following two main stages: (1) a systematic search and selection of information sources and (2) data extraction and comparative analysis across predefined thematic domains. These domains encompassed (1) the general characteristics of each AI evaluation framework, (2) validation aspects (eg, performance metrics and generalizability), and (3) the integration of ethical principles.

Eligibility Criteria

All searches covered literature up to February 28, 2026, with no start date limit and no language or publication status restrictions applied.

We included documents that proposed or described an evaluation or reporting framework for clinical AI, encompassing the following source types:

- Peer-reviewed scientific papers: publications in scientific journals presenting frameworks or guidelines for AI evaluation.
- Gray literature documents: documents such as white papers, technical reports, or policy guidelines not formally published in peer-reviewed journals.
- Official websites of relevant organizations: frameworks or directives available on official websites of relevant health, research, or regulatory organizations.
- Introductory series or conceptual papers without real-world application or empirical evaluation.
- Comparative or descriptive analyses of existing AI checklists.
- Documents from out-of-scope domains, including expert conference proceedings
- Letters to the editor, commentaries, or editorial responses lacking original data
- Other reasons

Publications were excluded if they did not present a tangible AI evaluation framework or were outside the scope of our review. Specifically, we excluded methodological or development protocols without validation results or empirical testing of AI models; introductory series or conceptual papers without real-world application or empirical evaluation; comparative or descriptive analyses of existing AI checklists; documents from out-of-scope domains, including expert conference proceedings; letters to the editor, commentaries, or editorial responses lacking original data; and other reasons. Duplicate frameworks reported in multiple sources were counted only once.

Information Sources

Comprehensive searches were performed in six electronic databases: PubMed (MEDLINE), BVS, Embase, EBSCOhost (including CINAHL and others), Sage Journals, and ProQuest Central. The process was conducted independently by DCLM and MH-P.

In addition, we searched pertinent gray literature sources. The EQUATOR (Enhancing the Quality and Transparency of Health Research) Network was used to identify relevant reporting guidelines for AI, and we screened the official websites of major organizations and consortia involved in clinical AI for any published frameworks [21]. This search was conducted by NMA and DCLM.

Search

The search strategy combined controlled vocabulary terms (eg, MeSH and Emtree headings) with free-text keywords, ensuring comprehensive coverage of the following three concept areas: (1) AI and machine learning in health care, (2) model validation and performance evaluation, and (3) ethical principles and frameworks. An example PubMed search strategy is provided below:

("Artificial Intelligence"[Mesh] OR "Artificial Intelligence"[tiab] OR "Machine Learning"[Mesh] OR "Machine Learning"[tiab] OR "Deep Learning"[tiab] OR "Neural Networks, Computer"[Mesh] OR "Neural Network*"[tiab] OR "Natural Language Processing"[Mesh] OR "Natural Language Processing"[tiab] OR "Generative AI"[tiab] OR "Explainable AI"[tiab]) AND ("Validation Studies as Topic"[Mesh] OR "Validation"[tiab] OR "External Validation"[tiab] OR "Model Validation"[tiab] OR "Generalizability"[tiab] OR "Calibration"[tiab] OR "Performance Metrics"[tiab] OR "Reproducibility of Results"[Mesh] OR "Overfitting"[tiab]) AND ("Decision Support Systems, Clinical"[Mesh] OR "Clinical Decision Support"[tiab] OR "Clinical Decision-Making"[Mesh] OR "Predictive Models"[Mesh] OR "Prediction Model*"[tiab] OR "Prognostic Model*"[tiab] OR "Risk Prediction"[tiab] OR "Diagnostic Accuracy"[tiab] OR "Computer-Assisted Diagnosis"[Mesh]) AND ("Ethics, Medical"[Mesh] OR "Bioethics"[Mesh] OR "Ethic*"[tiab] OR "Equity"[tiab] OR "Transparency"[tiab] OR "Accountability"[tiab] OR "Informed Consent"[tiab] OR "Patient Autonomy"[tiab] OR "Privacy"[tiab] OR "Confidentiality"[tiab] OR "Responsible AI"[tiab] OR "Evaluation Framework*"[tiab] OR "Reporting Guideline*"[tiab] OR "CONSORT-AI"[tiab] OR "SPIRIT-AI"[tiab] OR "TRIPOD-AI"[tiab] OR "STARD-AI"[tiab] OR "CLAIM"[tiab]).

Equivalent search queries were adapted for each database using the respective controlled vocabulary and syntax (refer to Table S1 in Multimedia Appendix 1 for the full search strategies).

Selection of Sources of Evidence

The search strategies were reviewed by the study coauthors. All references retrieved were exported into Microsoft Excel, which was used for reference management, deduplication, and subsequent screening steps. First, 2 reviewers (ACHA and NMA) independently screened the titles and abstracts of all unique records against the eligibility criteria. Next, full-text papers or documents were obtained and assessed for inclusion by the same 2 reviewers working in duplicate. Any discrepancies or uncertainties in study selection were resolved through discussion; if consensus could not be reached, a third reviewer served as an adjudicator.

Data Charting Process

A standardized data charting form was used to extract relevant information from each included source, structured according to the 3 key domains of interest. For each included framework, data were charted and organized into comparative tables corresponding to general framework characteristics, validation and performance evaluation aspects, and ethical considerations.

Two reviewers collaboratively performed the data extraction, with cross-verification by the broader team to ensure consistency. Any discrepancies were resolved through

discussion, and when consensus could not be reached, a third reviewer adjudicated (DCLM).

General Characteristics

General characteristics extracted and compared across frameworks included development team composition, disciplinary scope, methodological approach, structural domains, extensions, and stated purpose. Development team composition was categorized as unitary, binary, ternary, or quaternary, including the involvement of one, two, three, or four or more institutions, respectively. Disciplinary scope was defined as interdisciplinary if the team included members from different academic fields, and transdisciplinary if it also incorporated nonacademic stakeholders. Structural domains were grouped into IMRD (Introduction, Methods, Results, and Discussion) and non-IMRD formats. Methodological approach, stated purpose, and extensions were recorded exactly as reported in each evaluation framework.

Validation and Performance Aspects

We extracted details to understand each framework's orientation toward clinical use and its recommended validation stringency. Each framework's intended purpose or primary use case was noted and categorized as either research-focused (primarily for studies developing or reporting AI models) or clinical utility-focused (intended to guide evaluation for real-world implementation). For clinically oriented frameworks, we further noted the specified application domain (eg, diagnostic vs prognostic models) and any criteria used to define "clinical utility" [9].

To appraise methodological rigor, we collected information on the following three key validation parameters defined a priori: technical performance, clinical performance, and generalizability [22]. Technical performance (internal validity) measured using metrics such as sensitivity, specificity, area under the curve (AUC) or receiver operating characteristic (ROC), and free-response ROC curve (FROC). ROC and FROC were classified as internal validation parameters, as they are primarily influenced by threshold effects rather than disease prevalence [23,24]. Clinical performance (external validity), assessed using predictive values (positive and negative) and calibration accuracy in AI models. These metrics are influenced by disease prevalence and therefore better reflect real-world performance [22]. Generalizability, defined as the framework's ability to account for limitations related to uncontrolled overfitting. Overfitting occurs when an algorithm is trained and tested on the same dataset used for development or when internal data splits are used instead of independent or external datasets for validation [25].

- Technical performance (internal validity): measured using metrics such as sensitivity, specificity, AUC or ROC, and FROC. ROC and FROC were classified as internal validation parameters, as they are primarily influenced by threshold effects rather than disease prevalence [23,24].

For this review, evaluation frameworks were classified a priori into three levels of methodological rigor with respect to their assessment of clinical utility and generalizability:

- Complete evaluation: defined as frameworks that require validation using independent external datasets distinct from those used for model training, while also considering the epidemiological study design (eg, diagnostic cohort studies or case-control designs). This approach ensures both internal validity (the model performs adequately in its development sample) and external validity (the model retains performance across real-world clinical settings). Frameworks in this category provide the strongest evidence of clinical reliability, as they minimize overfitting and explicitly account for study design quality.
- Partial evaluation: defined as frameworks that acknowledge the importance of external validation datasets but do not assess the methodological quality or epidemiological design of those datasets. Although this category represents a step toward generalizability, it does not guarantee that external data are sufficiently robust or representative. Consequently, such frameworks provide only limited evidence to support clinical implementation, as risks of selection bias and prevalence-related distortions remain.
- Inadequate evaluation: defined as frameworks that consider AI models to be generalizable even when validation is based solely on internal datasets, typically through data splitting or cross-validation within the same population. This approach fails to ensure external validity and is highly susceptible to overfitting, as reported performance primarily reflects adaptation to the original dataset. Frameworks in this category should not be regarded as sufficient to support clinical applicability; rather, they are relevant only for the developmental phase of model assessment.

Synthesis of Results

We synthesized the charted data through a descriptive comparative approach, emphasizing both the distribution of characteristics across frameworks and the identification of patterns or gaps in current practice. Extracted information was tabulated and compared across the three thematic domains to facilitate a side-by-side analysis of frameworks (eg, allowing readers to contrast different frameworks' requirements for validation or ethical compliance).

A dot plot was used as a graph to present the gap map, following the approach suggested by Nyanchoka et al [26]. The evaluation of gaps was performed across 3 predefined domains. For each domain, frameworks were assessed according to specific criteria, and scores were assigned to indicate the presence or absence of gaps.

In Domain 1 "general characteristics," frameworks were classified as having no gap (1 point) when they incorporated quaternary teams and a transdisciplinary scope. Frameworks that only included an interdisciplinary scope or other team compositions were considered to have a gap (0 points).

In Domain 2 "validation and performance aspects," frameworks received a score of no gap (1 point) when they addressed clinical utility and provided comprehensive validation (complete generalizability) and appropriate performance considerations. Frameworks that did not explicitly include these elements were classified as having a gap (0 points).

In Domain 3 "ethics," frameworks were considered to have no gap (1 point) when they explicitly incorporated direct and indirect principles and did not present missing ethical components (direct + incorrect and no missing components). Frameworks lacking these elements were categorized as having a gap (0 points).

This scoring approach allowed the identification of gaps across domains and facilitated the graphical representation of the distribution of strengths and weaknesses among the evaluated frameworks.

We did not perform a formal quantitative meta-analysis given the qualitative nature of the data; however, we carried out simple quantitative aggregations (counts and percentages) for certain features to summarize prevailing trends (reported in the Results section).

Protocol Deviations

Deviations from the initial protocol included updating the literature search at a later date than originally planned to ensure the currency of the evidence base, adding a gap map to enhance the presentation of results, incorporating PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) reporting elements to improve transparency of the search process, refining the approach to results synthesis, and expanding the number of databases searched. These modifications were made to strengthen the completeness, transparency, and interpretability of the review and did not change its primary objective.

Ethical Considerations

The study was approved by the Research Ethics Committee of the Universidad Cooperativa de Colombia (protocol INV3688). The ethical evaluation of the included frameworks followed the UNESCO ethical principles for AI [27]. An ethical scoring matrix (Table S2 in Multimedia Appendix 1 [13-15,17,28-69]) was used to assess each framework across the 10 UNESCO domains, with predefined criteria indicating whether each principle was explicitly addressed (fully covered with clear guidance), partially aligned (mentioned or implied but not fully detailed), or absent. Each principle (eg, transparency, fairness, privacy, human oversight, and accountability) was evaluated individually using a semi-quantitative scoring system. Domain-level scores were then calculated to compare the ethical coverage across frameworks. Two reviewers conducted the assessment independently, resolving discrepancies through consensus. This structured matrix-based approach ensured a systematic and reproducible analysis of the ethical strengths and gaps within current AI evaluation frameworks used in clinical practice.

Results

Study Selection

A total of 3363 records were identified through database and gray literature searches. After removal of duplicates, 391 records were screened at the full-text level. Of these, 345 were excluded for not meeting the eligibility criteria, primarily because they did not propose or describe an evaluation framework for clinical AI or lacked evaluative content. Ultimately, 46 distinct evaluation frameworks were included in the scoping review. Figure 1 presents the PRISMA flow diagram detailing the number of records identified, screened, excluded, and finally included in the review. Table 1 summarizes all included papers and their main characteristics, while the complete list of excluded studies with corresponding reasons is available in Multimedia Appendix 2.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Adapted from Page et al [70].

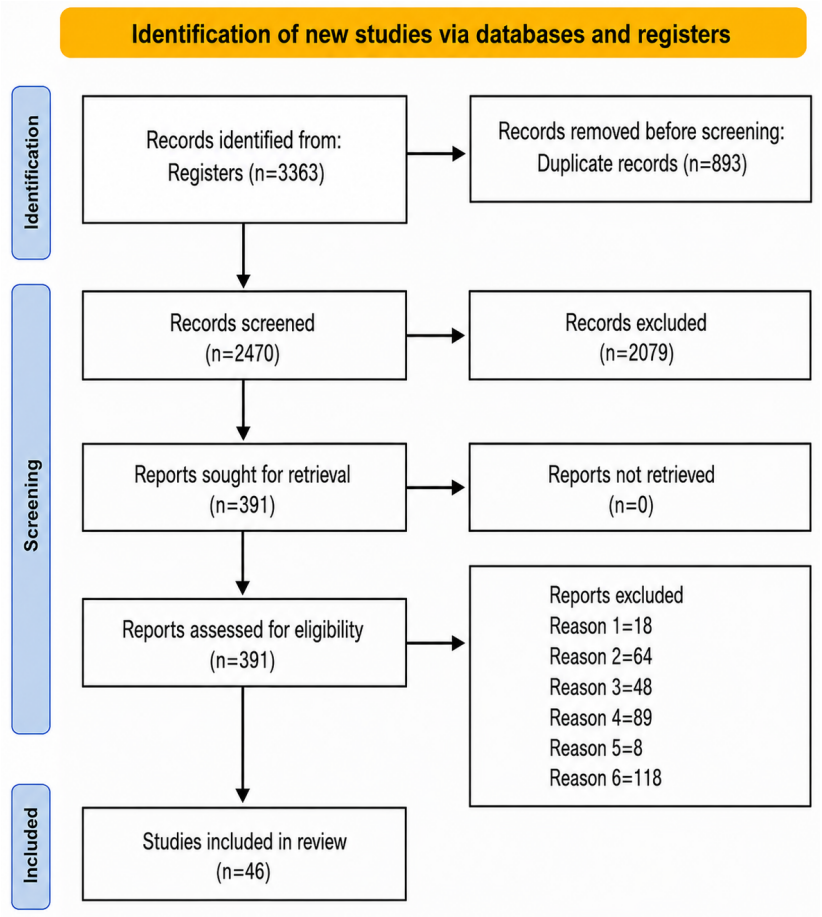


Table 1. Summary of the included documents: collaboration type, disciplinary scope, development methodology, structural domains, and stated purpose of the identified AI evaluation frameworks.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Luo et al, 2016 [28]	Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research	Quaternary	Interdisciplinary	Delphi method (iterative consensus process via email with 11 experts from 3 institutions across 3 continents).	IMRD[a]	Not applicable	To establish minimum reporting items and procedural steps ensuring valid application, transparent reporting, and reproducibility of machine learning predictive models in biomedical research.
Floridi et al, 2018 [29]	AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	Quaternary	Transdisciplinary	Consensus-based ethical synthesis and policy Delphi integrating comparative analysis of existing AI ethics principles and collaborative drafting of 20 action-oriented recommendations.	Non-IMRD	Not applicable	To formulate a comprehensive ethical framework and actionable policy roadmap for building a "Good AI Society" that aligns technological innovation with human dignity, social justice, and accountability.
Reps et al, 2018 [30]	Standardized Framework for Generating and Evaluating Patient-Level Prediction Models Using Observational Healthcare Data	Ternary	Interdisciplinary	Framework design and proof-of-concept implementation integrating best-practice guidelines with the OMOP[b] Common Data Model; implemented via open-source R packages (R Core Team), (Patient Level Prediction) within the OHDSI[c] network.	IMRD	Not applicable	To provide a standardized, transparent, and reproducible framework for developing, validating, and sharing patient-level predictive models across observational health care databases using a common data model.
Cruz Rivera et al, 2020 [31]	Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI[d] Extension	Quaternary	Transdisciplinary	Multistage consensus-based development: literature review, expert consultation, generation of 26 candidate items, 2-round Delphi (103 participants), international consensus meeting (31 stakeholders), checklist pilot (34 participants), and final EQUATOR[e]-registered extension.	IMRD	EQUATOR—Official reporting guideline extension of SPIRIT 2013 (SPIRIT-AI)	To establish AI-specific protocol reporting standards ensuring transparency, reproducibility, safety, and completeness in clinical trial protocols evaluating AI interventions.
Hernandez-Boussard et al, 2020 [32]	MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care	Unitary	Interdisciplinary	Conceptual proposal based on expert opinion and synthesis of existing reporting standards (MIAME[f], CONSORT[g], SPIRIT[h], PRISMA[i], STROBE[j], and TRIPOD[k]); no Delphi or consensus process.	Non-IMRD	Not applicable	To propose minimum reporting standards for AI and ML[l] models in health care, focusing on transparency of training data, model design, population characteristics, and validation to improve reproducibility and mitigate bias.
Liu et al, 2020 [33]	Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI[m] Extension	Quaternary	Transdisciplinary	Staged consensus process: literature review, candidate item generation, 2-round Delphi survey (103 participants), international consensus meeting (31 participants), pilot testing (34 participants), and development using EQUATOR[e] reporting-guideline methodology.	IMRD	CONSORT-AI (EQUATOR)	To develop a reporting guideline extension (CONSORT-AI) specifying additional items required for transparent, complete, and reproducible reporting of randomized clinical trials evaluating AI-based interventions.
Mongan et al, 2020 [34]	Checklist for Artificial Intelligence in Medical Imaging (CLAIM)	Binary	Interdisciplinary	Consensus-based guideline development integrating and adapting existing reporting standards (STARD[n], STROBE, CONSORT, and EQUATOR) to AI in medical imaging.	IMRD	Not applicable	To establish a standardized reporting framework for authors and reviewers to ensure transparency, reproducibility, and methodological rigor in studies applying AI to medical imaging.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Norgeot et al, 2020 [35]	Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) Checklist	Quaternary	Interdisciplinary	Consensus-driven documentation framework: design of a 6-part checklist defining minimal reporting standards for clinical AI modeling; developed collaboratively and released with a public GitHub repository for community feedback.	Non-IMRD	Not applicable	To establish a minimum set of information required for transparent, fair, and reproducible reporting of clinical AI modeling studies.
Sengupta et al, 2020 [36]	Proposed Requirements for Cardiovascular Imaging–Related Machine Learning Evaluation (PRIME) Checklist	Quaternary	Interdisciplinary	Consensus-based methodological synthesis: iterative development of a 7-domain framework through expert review and synthesis of best practices in cardiovascular ML[l] model design, validation, and reporting.	Non-IMRD	Not applicable	To establish a standardized set of requirements and reporting guidelines for developing, validating, and communicating machine learning models in cardiovascular imaging, ensuring reproducibility, transparency, and methodological consistency.
Stevens et al, 2020 [37]	Recommendations for Reporting Machine Learning Analyses in Clinical Research	Ternary	Interdisciplinary	Consensus-based methodological synthesis: development of structured recommendations integrating existing reporting standards (STROBE and TRIPOD) with machine learning–specific considerations for clinical research.	Non-IMRD	Not applicable	To propose structured recommendations and reporting elements for transparent, reproducible, and interpretable reporting of machine learning analyses in clinical research.
Young et al, 2020 [38]	Artificial Intelligence in Dermatology: A Primer	Binary	Interdisciplinary	Narrative review and conceptual synthesis: integrative appraisal of AI applications in dermatology, analyzing performance, equity, generalizability, and interpretability across existing studies.	Non-IMRD	Not applicable	To provide a comprehensive overview and conceptual framework for understanding the capabilities, limitations, and ethical challenges of AI in dermatology, with proposed metrics for reporting model performance.
Cabitza and Campagner, 2021 [39]	The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies	Unitary	Interdisciplinary	Checklist development based on CRISP-DM[o] methodology: synthesis of prior frameworks (MINIMAR[p], CONSORT-AI, SPIRIT-AI, MI-CLAIM[q], WHO[r] or ITU[s] ML4H[t], PROBAST, and TRIPOD) into a 30-item tool organized in six phases: problem understanding, data understanding, data preparation, modeling, validation, and deployment.	Non-IMRD	Not applicable	To introduce a comprehensive 30-item checklist to assist authors and reviewers in evaluating the methodological soundness, reproducibility, and quality of medical AI studies, aligning with international reporting standards.
Ji et al, 2021 [40]	Evaluation Framework for Successful Artificial Intelligence–Enabled Clinical Decision Support Systems: Mixed Methods Study	Ternary	Interdisciplinary	Mixed methods design combining Delphi process, cognitive interviews, and structural equation modeling development and psychometric validation of a 28-item measurement instrument assessing AI-CDSS[u] success across six latent variables (system, information, and service quality; perceived ease of use; perceived benefit; user acceptance).	IMRD	Not applicable	To develop and validate a comprehensive evaluation framework identifying key determinants of success for AI-enabled clinical decision support systems, focusing on user acceptance as the central construct.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Olczak et al, 2021 [41]	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	Quaternary	Interdisciplinary	Consensus-driven guideline development: synthesis of key methodological, statistical, and ethical principles in medical AI, culminating in the CAIR[v] checklist and reporting recommendations for clinicians and AI researchers.	IMRD	Not applicable	To propose a comprehensive reporting and evaluation checklist (CAIR) for presenting and interpreting medical AI studies, enhancing communication between engineers, clinicians, and health care stakeholders.
Schwendicke et al, 2021 [42]	Artificial Intelligence in Dental Research: Checklist for Authors, Reviewers, and Readers	Quaternary	Transdisciplinary	Consensus-based e-Delphi process following CREDES[w] guidelines, integrating prior frameworks (CLAIM[x], STARD, TRIPOD, CONSORT-AI, SPIRIT-AI, RECORD)[y]; resulted in a 31-item checklist for authors, reviewers, and readers.	IMRD	Not applicable	To provide a 31-item consensus checklist for improving the planning, conduct, and reporting of AI studies in dental research, enhancing reproducibility, transparency, and methodological rigor.
Bazoukis et al, 2022 [43]	The inclusion of augmented intelligence in medicine: A framework for successful implementation	Quaternary	Transdisciplinary	Conceptual synthesis and policy framework development integrating ethical, regulatory, and practical dimensions of AI adoption into a unified framework addressing reliability, oversight, liability, equity, patient rights, and cybersecurity.	Non-IMRD	Not applicable	To propose an integrated regulatory and operational framework to guide developers, clinicians, researchers, and regulators in the safe and equitable implementation of augmented intelligence in clinical practice.
Daneshjou et al, 2022 [44]	CheckList for Evaluation of Image-Based AI Reports in Dermatology (CLEAR Derm): Consensus Guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group	Quaternary	Transdisciplinary	Two-round virtual consensus process: systematic literature review (PubMed 2008-2021) followed by expert panel consensus (19 ISIC[z] members) using Delphi methodology to develop a 25-item checklist grouped into four domains (data, technique, technical assessment, and application).	IMRD	Yes – aligns with EQUATOR or PRISMA-related initiatives (referencing STARD-AI, CONSORT-AI, SPIRIT-AI, and DECIDE-AI[aa])	To establish a consensus-based, dermatology-specific checklist (CLEAR Derm) for evaluating image-based AI studies, ensuring fairness, reliability, transparency, and ethical clinical translation.
Fusar-Poli et al, 2022 [45]	Development and validation of the Clinical Artificial Intelligence Research (CAIR) checklist for reporting clinical AI studies: a multi-phase international consensus study	Quaternary	Transdisciplinary	Multiphase international Delphi consensus: four-stage process (literature review, Delphi rounds with >150 experts from 26 countries, pilot testing, and validation) to develop a 22-item checklist for transparent and standardized reporting of clinical AI research.	IMRD	Yes – aligned with EQUATOR or PRISMA-related initiatives; complementary to CONSORT-AI, SPIRIT-AI, and TRIPOD-AI[ab]	To create a validated, consensus-based reporting guideline (CAIR) ensuring transparency, reproducibility, and clinical relevance in studies applying AI to health care research and practice.
Kwong et al, 2021 [46]	Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework	Quaternary	Interdisciplinary	PRISMA-guided systematic literature review and expert item synthesis resulting in a 26-item checklist mapped a TRIPOD.	IMRD	PRISMA followed; no specific PRISMA extension reported	To develop a urology-specific, standardized reporting framework for ML studies to improve transparency, reproducibility, comparability, and clinical uptake.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Lu et al, 2022 [47]	Assessment of Adherence to Reporting Guidelines by Commonly Used Clinical Prediction Models From a Single Vendor: A Systematic Review	Unitary	Interdisciplinary	Systematic review (PRISMA-based): MEDLINE search (Nov 2020 to Dec 2020) identifying 15 model reporting guidelines; synthesis of 220 unique items and cross-sectional evaluation of 12 deployed AI models from Epic Systems using expert adjudication.	IMRD	PRISMA followed; no specific PRISMA extension reported	To evaluate the overlap among existing AI model reporting guidelines and assess how well the documentation of widely deployed models adheres to these standards, identifying reporting gaps in reliability and fairness.
Shen et al, 2022 [48]	An Ethics Checklist for Digital Health Research in Psychiatry: Viewpoint	Ternary	Transdisciplinary	Stakeholder-informed consensus workshop: interdisciplinary and stakeholder meeting (May 2020) supported by NIH[ac] Bioethics Supplement; qualitative synthesis leading to a 20-item ethics checklist across six domains (informed consent, equity and access, privacy and partnerships, regulation and law, return of results, and duty to warn or report).	Non-IMRD	Not applicable	To develop a 20-question ethics checklist addressing procedural safeguards and ethical, legal, and social implications (ELSI) in digital psychiatry and deep phenotyping research.
Vasey et al, 2022 [49]	Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI	Quaternary	Transdisciplinary	Two-round modified Delphi + virtual consensus meeting conducted according to EQUATOR Network standards; 151 experts from 18 countries and 20 stakeholder groups participated; guideline registered on EQUATOR and Open Science Framework.	IMRD	Yes – official EQUATOR or PRISMA-related extension; complements CONSORT-AI, SPIRIT-AI, TRIPOD-AI, and STARD-AI	To provide a multistakeholder, consensus-based reporting guideline (DECIDE-AI) for the early clinical evaluation of AI-based decision support systems, focusing on clinical utility, safety, human factors, and readiness for large-scale trials.
Abdulazeem et al, 2023 [50]	A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data	Binary	Interdisciplinary	Systematic review (PRISMA-based): registered on PROSPERO (CRD42021264582); databases: Cochrane, PubMed, Web of Science, Elsevier, BioRxiv, ACM, IEEE; screening with Rayyan (Rayyan Systems Inc., Cambridge, MA, USA); risk of bias via PROBAST and methodological appraisal via CHARMS[ad].	IMRD	PRISMA (2020) followed; no extension reported	To identify clinical conditions targeted by ML models trained or validated using real-world primary health care (PHC) data, and to map the methodological characteristics, validation approaches, and performance measures of these models.
Cacciamani et al, 2023 [15]	PRISMA-AI: Reporting guidelines for systematic reviews and meta-analyses on AI in healthcare	Quaternary	Transdisciplinary	Multiphase development process: literature review, Delphi survey among multidisciplinary experts, consensus meeting, piloting, and creation of the PRISMA-AI[ae] checklist and explanation or elaboration document; guideline registered with EQUATOR and ClinicalTrials.gov (NCT05382455).	Non-IMRD	Yes–official PRISMA (PRISMA-AI)	To establish PRISMA-AI, a reporting guideline extension for systematic reviews and meta-analyses addressing AI-based interventions in health care, ensuring transparency, reproducibility, and clinical applicability.
Debray et al, 2023 [51]	Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster[af]): explanation and elaboration	Quaternary	Interdisciplinary	Delphi consensus and methodological synthesis: iterative expert consensus based on prior TRIPOD, TRIPOD-E&E[ag] (2015), and empirical evaluations of clustered datasets (individual participant data meta-analyses and EHR[ah]-based studies); resulted in a 19-item checklist.	IMRD	Yes–TRIPOD-Cluster builds upon the original TRIPOD and aligns with PRISMA-IPD, EQUATOR, and PROBAST frameworks	To provide guidance for transparent reporting of multivariable prediction models developed or validated using clustered data, addressing heterogeneity, generalizability, and bias in individual participant data meta-analyses and multicenter or EHR studies.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Elvidge et al, 2023 [52]	Consolidated Health Economic Evaluation Reporting Standards for Interventions that use Artificial Intelligence (CHEERS-AI)	Quaternary	Transdisciplinary	Multiphase development: (1) initiation (steering group + longlist), (2) three-round Delphi study with 58-31 experts, (3) consensus meeting, (4) patient involvement via EURORDIS Digital Advisory Group, (5) piloting of items on 9 economic evaluations, and (6) final steering group ratification.	IMRD	Not PRISMA – this is an official CHEERS extension (CHEERS-AI), aligned with EQUATOR, not PRISMA.	To develop a reporting guideline extension ensuring transparent, complete, and reproducible reporting of economic evaluations of AI-based health interventions, adding AI-specific items to CHEERS-2022.
Klement and El Emam, 2023 [53]	Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation	Binary	Interdisciplinary	Scoping-review–informed consolidation of reporting guidelines: broad literature search (192 papers), screening against predefined criteria, quality appraisal using a 9-item checklist, extraction of reporting items from 17 high-quality guidelines; followed by external expert review (JMIR AI editorial board) and initial validation on 6 ML studies.	IMRD	PRISMA followed (for the ML guideline search); no specific PRISMA extension reported	To produce a single consolidated checklist of reporting items for prognostic and diagnostic ML modeling studies (in-silico and shadow-mode), improving transparency, reproducibility, and methodological clarity.
Kwong et al, 2023 [54]	APPRAISE-AI[ai] Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support	Quaternary	Interdisciplinary	Framework development informed by literature review + expert panel refinement; includes reliability testing (interrater and intrarater ICCs[aj]), construct validation (correlation with expert scores, citation rates, QUADAS-2[ak], and TRIPOD), and application to a published systematic review of sepsis prediction models.	IMRD	Not applicable	To develop and validate APPRAISE-AI, a quantitative scoring tool that assesses the methodological rigor, reporting quality, and robustness of clinical AI studies across six domains (clinical relevance, data quality, methodological conduct, robustness, reporting quality, reproducibility).
Murphy et al, 2023 [55]	A guide to optometrists for appraising and using artificial intelligence in clinical practice	Binary	Interdisciplinary	Narrative review with expert-informed synthesis: conceptual explanation of AI fundamentals for optometry, followed by a practical checklist addressing regulatory approval, intended use, clinical applicability, population fit, performance metrics, and explainability.	Non-IMRD	Not applicable	To provide optometrists with a practical, clinically oriented checklist for appraising whether AI systems are safe, appropriate, and suitable for use in routine optometric practice.
Collins et al, 2024 [56]	TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods	Quaternary	Transdisciplinary	Multiphase consensus process following EQUATOR guidance: literature review, item generation, modified Delphi (2 rounds, 170-200 participants); patient or public involvement meeting; online consensus meeting (28 participants); checklist refinement and expansion.	IMRD	TRIPOD family (EQUATOR): TRIPOD-Cluster, TRIPOD-SRMA, TRIPOD-LLM[al], TRIPOD-AI (TRIPOD + AI). Not a PRISMA extension.	To provide an updated, harmonized reporting guideline (TRIPOD + AI) for transparent, complete, and accurate reporting of prediction model studies using regression or machine learning methods.
Elfer et al, 2024 [57]	Reproducible Reporting of the Collection and Evaluation of Annotations for Artificial Intelligence Models	Quaternary	Interdisciplinary	Expert-informed operationalization of a prior workflow (Wahab et al[57,71]) into a reporting framework + checklist, demonstrated through application to an annotation project (HTT[am]). Includes iterative refinement, piloting, expert review, and quality assessment of dataset construction processes. No Delphi or formal consensus process.	IMRD	CLEARR-AI (not a PRISMA, TRIPOD, CONSORT, SPIRIT, STARD, DECIDE, or CHEERS extension. Conceptually aligned with EQUATOR AI extensions but not an official extension.)	To develop and demonstrate CLEARR-AI, a reporting framework and checklist for transparent, reproducible documentation of medical image annotation datasets used in AI model development, validation, and evaluation.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
El Emam et al, 2024 [58]	Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS)	Binary	Interdisciplinary	Editorial synthesis describing CREMLS[an] checklist (developed previously via structured literature review and quality appraisal), illustrating item application through published examples; not a new framework development study.	Non-IMRD	CREMLS (independent checklist)	To present and formalize JMIR Publications' adoption of the CREMLS checklist for improving completeness, methodological transparency, and reproducibility in diagnostic and prognostic ML studies.
Guni A et al, 2024 [59]	Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): Protocol for a Qualitative	Quaternary	Transdisciplinary	Three-stage development protocol: (1) project organization; (2) item generation (mapping review, meta-research study, international scoping survey, PPIE[ao] focus group); (3) modified Delphi process (multiple online rounds and consensus meeting); piloting and drafting of QUADAS-AI tool and explanation or elaboration document.	IMRD	EQUATOR: QUADAS family extension (in development). Not PRISMA.	To develop QUADAS-AI, an AI-specific quality assessment tool for systematic reviews evaluating the diagnostic accuracy of AI systems; addressing biases, applicability, and methodological challenges unique to AI-based diagnostic studies.
Kapoor et al, 2024 [60]	REFORMS: Consensus-based Recommendations for Machine-learning-based Science	Quaternary	Interdisciplinary	Consensus-based development: extensive literature review + multiple rounds of internal expert revision + virtual consensus discussions among 19 researchers; no Delphi, no PPIE, no stakeholder outreach.	Non-IMRD	Independent checklist. NOT a PRISMA extension; NOT EQUATOR-registered; NOT: TRIPOD, CONSORT, SPIRIT, STARD, CHEERS	To establish cross-disciplinary recommendations and a 32-item checklist to improve rigor, reproducibility, transparency, and error detection in machine-learning–based science.
Labkoff et al, 2024 [61]	Toward a responsible future: recommendations for AI-enabled clinical decision support	Quaternary	Transdisciplinary	Large-scale consensus process: four introductory webinars (ethics or religion, patient perspectives, regulatory context, and risk management), a 2-d in-person workshop with >200 stakeholders, breakout groups, iterative qualitative synthesis, and 4-month iterative Delphi-like consensus process among authors.	IMRD	Independent consensus framework; not PRISMA or EQUATOR	To produce consensus-based, cross-stakeholder recommendations for trustworthy, safe, effective, and equitable development, validation, implementation, certification, monitoring, and lifecycle governance of AI-enabled clinical decision support (AI-CDS) systems.
Masters and Salcedo, 2024 [62]	A checklist for reporting, reading, and evaluating Artificial Intelligence Technology Enhanced Learning (AITEL) research in medical education	Binary	Interdisciplinary	Narrative, expert-informed development: multisource review of ISO[ap] standards, FDA[aq] guidance, EU AI[ar] Act, educational frameworks; iterative refinement; workshop feedback (AMEE 2022 [62]); internal expert review; no Delphi and no consensus panel.	Non-IMRD	Not applicable	To propose a structured reporting checklist for describing, evaluating, and understanding AI-based Technology-Enhanced Learning (AITEL) systems in medical education, improving transparency, reproducibility, and ethical oversight.
Ning et al, 2024 [17]	Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist	Quaternary	Interdisciplinary	Systematic scoping review (193 papers) using PRISMA-ScR[as], followed by thematic synthesis of ethical principles, generation of a 10-principle ethics checklist (TREGAI[at]); iterative expert internal review; no Delphi or consensus process; checklist maintained as a live online document.	IMRD	PRISMA-ScR (used for the review); TREGAI itself is not a PRISMA or EQUATOR extension	To provide a systematic ethical assessment framework and propose the TREGAI checklist to reinforce transparency, accountability, equity, and responsible practice in generative AI health care research.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Ray et al, 2024 [63]	Decoding skin cancer classification: perspectives, insights, and advances through researchers' lens	Binary	Interdisciplinary	Narrative, experience-based framework derivation: introduction of 2 practical checklists based on prior educational literature, teaching experience, and iterative refinement within clinical simulation programs; no Delphi or formal guideline development.	Non-IMRD	Not applicable	To introduce 2 practical checklists for faculty preparation and scenario design in AI-enhanced medical simulation, supporting safe, transparent, and intentional incorporation of AI tools in clinical education.
Tejani et al, 2024 [13]	Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update	Quaternary	Interdisciplinary	Formal Delphi consensus process: renewal of EQUATOR registration, recruitment of 73 international experts, 2-round Delphi (n=72), steering committee review, restructuring of items, and creation of CLAIM 2024 checklist.	IMRD	EQUATOR-registered reporting guideline (CLAIM 2024) — NOT PRISMA	To revise and formalize CLAIM into the 2024 EQUATOR-registered guideline, improving transparency, reproducibility, and completeness of AI medical imaging research through a structured, expert-derived 44-item reporting checklist.
Uribe et al, 2024 [64]	Integrating Generative AI in Dental Education: A Scoping Review of Current Practices and recommendations	Binary	Interdisciplinary	Scoping review following JBI[au] methodology; protocol registered on OSF[av]; multisource search (university websites, search engines, and email outreach).	IMRD	PRISMA-ScR (used for reporting the review)	To identify and summarize existing institutional guidelines on generative AI use in dental education; does not create or validate an evaluation framework.
Warren et al, 2024 [65]	An Introductory Guide to Artificial Intelligence in Interventional Radiology: Part 2: Implementation Considerations and Harms	Ternary	Interdisciplinary	Narrative expert-informed framework development, derived from regulatory documents (WHO[aw], FDA, and IMDRF)[ax], prior IR AI frameworks, safety science, and authors' clinical experience; iterative internal refinement; no Delphi, no consensus meeting, no systematic review.	IMRD	Independent implementation framework. Not a PRISMA extension; not EQUATOR-registered	To provide a risk-based framework and an 11-item checklist to guide safe, structured, and context-aware implementation and evaluation of AI tools in interventional radiology practice.
Kalaycıoğlu et al, 2025 [66]	Evaluating the sample size requirements of tree-based ensemble machine learning techniques for clinical risk prediction	Ternary	Interdisciplinary	Extensive simulation study comparing sample size requirements for ensemble MLTs[ay] (bagging, RF[az], and boosting) vs logistic regression; uses real datasets + multiple DGMs[ba]; evaluates performance metrics (MAPE[bb], C-statistic, calibration, and Brier score).	IMRD	Not applicable	To evaluate whether existing sample size formulas for logistic regression apply to tree-based MLTs and to determine sample size requirements for development and external validation of MLT-based clinical risk prediction models.
Pan American Health Organization (PAHO), 2025 [67]	AI prompt design for public health: Using generative AI responsibly	Unitary	Interdisciplinary	Narrative, evidence-informed guidance manual; structured chapters; practical prompt templates; includes a Prompt Review Checklist but no evaluation framework, no Delphi, no consensus process.	Non-IMRD	Not applicable	To provide guidance for responsible prompt design in generative AI applications for public health, including templates, examples, quality control considerations, and a prompt review checklist.
Sounderajah et al, 2025 [14]	STARD-AI Steering Committee. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence	Quaternary	Transdisciplinary	Multistakeholder development: systematic review, expert survey (80 experts), PPIE focus group, candidate item generation, modified Delphi (2 rounds, >240 participants), preconsensus prioritization, international consensus meeting, and final Steering Committee refinement.	IMRD	EQUATOR — Official reporting guideline extension of STARD 2015 (STARD-AI)	To provide a minimum essential reporting guideline for diagnostic accuracy studies evaluating AI-based tests, improving transparency, reducing bias, enhancing reproducibility, and enabling clinical, regulatory, and policy decision-making.

Author, year, and reference	Title	Development team	Disciplinary scope	Method	General domains	Extension	Purpose
Tuygunov et al, 2025 [68]	The Transformative Role of Artificial Intelligence in Dentistry: A Comprehensive Overview Part 2: The Promise and Perils, and the International Dental Federation Communique	Quaternary	Interdisciplinary	Narrative concise review synthesizing recent literature and summarizing the FDI White Paper; no Delphi, no consensus, no systematic review.	Non-IMRD	Not applicable	To provide an updated overview of AI in dentistry, including educational uses, patient communication, integration challenges, ethical considerations, and a summary of the FDI AI Communiqué.
Wang et al, 2025 [69]	A practical guide for nephrologist peer reviewers: evaluating artificial intelligence and machine learning research in nephrology	Quaternary	Interdisciplinary	Narrative expert synthesis integrating established guidelines (TRIPOD-AI, TRIPOD-LLM, STARD, STROBE, and CRISP-DM) into an applied framework for peer reviewers; includes an 8-step evaluation schema but no Delphi or consensus or systematic review.	IMRD	TRIPOD-AI; TRIPOD-LLM (EQUATOR extensions used as foundational tools; no new extension created)	To provide a structured, guideline-based evaluation framework enabling nephrologist peer reviewers to appraise AI or machine learning studies rigorously, focusing on validation, dataset quality, bias, interpretability, and real-world applicability.

[a]IMRD: Introduction, Methods, Results, and Discussion.
[b]OMOP: Observational Medical Outcomes Partnership.
[c]OHDSI: Observational Health Data Sciences and Informatics.
[d]SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence.
[e]EQUATOR: Enhancing the Quality and Transparency of Health Research.
[f]MIAME: Minimum Information About a Microarray Experiment.
[g]CONSORT: Consolidated Standards of Reporting Trials.
[h]SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials.
[i]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
[j]STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.
[k]TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.
[l]ML: machine learning.
[m]CONSORT-AI: Consolidated Standards of Reporting Trials for Artificial Intelligence.
[n]STARD: Standards for Reporting Diagnostic Accuracy Studies.
[o]CRISP-DM: Cross-Industry Standard Process for Data Mining.
[p]MINIMAR: MINimum Information for Medical AI Reporting.
[q]MI-CLAIM: Minimum Information About Clinical Artificial Intelligence Modeling.
[r]WHO: World Health Organization.
[s]ITU: International Telecommunication Union.
[t]ML4H: Machine Learning for Health
[u]AI-CDSS: Artificial Intelligence–Clinical Decision Support System.
[v]CAIR: Clinical AI Research.
[w]CREDES: Conducting and REporting DElphi Studies.
[x]CLAIM: Checklist for Artificial Intelligence in Medical Imaging.
[y]RECORD: Reporting of studies Conducted using Observational Routinely-collected health Data.
[z]ISIC: International Skin Imaging Collaboration.
[aa]DECIDE-AI: Developmental and Evaluation Checklist for Decision Support Systems Driven by Artificial Intelligence.
[ab]TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence

[ac]NIH: National Institutes of Health.
[ad]CHARMS: Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies.
[ae]PRISMA-AI: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Artificial Intelligence.
[af]TRIPOD-Cluster: Transparent Reporting of Multivariable Prediction Models Developed or Validated Using Clustered Data.
[ag]TRIPOD-E&E: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Explanation and Elaboration.
[ah]EHR: electronic health record.
[ai]APPRAISE-AI: Assessment of Predictive Performance and Bias in Artificial Intelligence Studies.
[aj]ICC: intraclass correlation coefficient.
[ak]QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2.
[al]TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Models.
[am]HTT: High-Throughput Truthing project
[an]CREMLS: Consolidated Reporting of Machine Learning Studies.
[ao]PPIE: Patient and Public Involvement and Engagement.
[ap]ISO: International Organization for Standardization.
[aq]FDA: Food and Drug Administration.
[ar]EU AI: European Union Artificial Intelligence Act.
[as]PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
[at]TREGAI: Transparent Reporting of Ethics for Generative Artificial Intelligence.
[au]JBI: Joanna Briggs Institute.
[av]OSF: Open Science Framework.
[aw]WHO: World Health Organization.
[ax]IMDRF: International Medical Device Regulators Forum.
[ay]MLT: Machine Learning Technique.
[az]RF: random forest.
[ba]DGM: data-generating mechanism.
[bb]MAPE: mean absolute prediction error.

Domain 1: General Characteristics

The included publications, spanning from 2016 to 2025, illustrate a substantial expansion of methodological and reporting initiatives aimed at strengthening the evaluation of AI systems in health care. A majority of frameworks (63%, 29/46) were developed through quaternary, multi-institutional collaborations, and 32.6% (15/46) involved transdisciplinary participation, including regulators, industry partners, policymakers, or patient representatives.

With respect to methodological development, more than half of the frameworks were created using formal consensus processes, such as Delphi rounds, modified Delphi methods, consensus meetings, or iterative multistakeholder refinement. Among these, several represent EQUATOR-endorsed extensions (eg, CONSORT-AI, SPIRIT-AI [Standard Protocol Items – Recommendations for Interventional Trials – Artificial Intelligence Extension], Developmental and Evaluation Checklist for Decision Support Systems Driven by Artificial Intelligence [DECIDE-AI], CLAIM 2024, TRIPOD-Cluster [TRIPOD Extension for Clustered Data], TRIPOD + AI, CHEERS-AI [Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence], and STARD-AI). An additional 13% (6/46) were derived from systematic or scoping reviews. Only 2 frameworks (4.3%), APPRAISE-AI (Assessment of Predictive Performance and Bias in Artificial Intelligence Studies) and the AI-CDSS success framework, underwent empirical validation procedures, including reliability testing, construct validity assessment, or psychometric modeling.

Structural formats also varied: 65% (30/46) of the documents followed an IMRD-compliant structure. Across all frameworks, recurring domains included dataset description, model development procedures, performance metrics, internal and external validation strategies, reproducibility requirements, and overall reporting transparency. Several frameworks were specialty-specific, addressing fields such as dermatology, dentistry, cardiology, urology, or interventional radiology.

Importantly, 52.2% (24/46) of the documents corresponded to official extensions of established reporting guidelines (PRISMA or EQUATOR) or to independently implemented frameworks aligned with these standards.

Taken together, the included documents reveal a rapidly expanding and increasingly structured landscape of AI evaluation frameworks, characterized by high levels of interdisciplinarity, growing methodological convergence, and a clear movement toward formalized reporting standards designed to enhance transparency, reproducibility, and clinical applicability in AI-based health care research.

Domain 2: Validation and Performance Aspects

Two documents were not included in Table 2 because they addressed exclusively ethical considerations and did not provide evaluative components related to technical performance, clinical effectiveness, or clinical purpose.

Table 2. Characterization of the utility of AI evaluation frameworks applied in clinical practice.

Author, Year, and Reference	Objective	Technical performance	Clinical effectiveness	Generalizability	Clinical purpose
Luo et al, 2016 [28]	Investigative	It does not specify the parameters	It does not specify the parameters	Complete	Prediction
Floridi et al, 2018 [29]	Investigative	Not reported	Not reported	Not reported	Not reported
Reps et al, 2018 [30]	Clinical utility	AUC[a]	Sensitivity, specificity	Incomplete	Prediction
Cruz Rivera et al, 2020 [31]	Investigative	Not reported	Not reported	Not reported	Diagnostic
Hernandez-Boussard et al, 2020 [32]	Clinical utility	Not reported	Not reported	Complete	Diagnostic and prediction
Liu et al, 2020 [33]	Investigative	Not reported	Not reported	Not reported	Diagnostic
Mongan et al, 2020 [34]	Investigative	Sensitivity, specificity, AUC, PPV[b], NPV[c], and calibration accuracy	Not reported	Complete	Prediction
Norgeot et al, 2020 [35]	Investigative	F_1-scores, Dice coefficient, or AUC	Sensitivity, specificity, PPV, NPV, NNT[d], and AUC	Complete	Not reported
Sengupta et al, 2020 [36]	Investigative	AUC	Not reported	Incomplete	Not reported
Stevens et al, 2020 [37]	Investigative	AUC	Not reported	Incomplete	Prediction
Young et al, 2020 [38]	Investigative	Not reported	Not reported	Not reported	Not reported
Cabitza and Campagner, 2021 [39]	Investigative	Sensitivity, specificity, and AUC	PPV, NPV[c], and accuracy	Not reported	Not reported
Ji et al, 2021 [40]	Clinical utility	Not reported	Not reported	Not reported	Not reported
Olczak et al, 2021 [41]	Clinical utility	Exactitude, sensitivity, specificity, AUC, others (F_1-score or Dice score)	Calibration accuracy, PPV and NPV, among others	Wrong	Diagnostic and Prediction

Author, Year, and Reference	Objective	Technical performance	Clinical effectiveness	Generalizability	Clinical purpose
Schwendicke et al, 2021 [42]	Investigative	Sensitivity, specificity, and AUC	Not reported	Wrong	Not reported
Bazoukis et al, 2022 [43]	Investigative	Not reported	Not reported	Not reported	Not reported
Daneshjou et al, 2022 [44]	Investigative	Accuracy and free-response ROC[e]	Not reported	Incomplete	Diagnostic
Fusar-Poli et al, 2022 [45]	Investigative	Not reported	Not reported	Not reported	Diagnostic and prediction
Kwong et al, 2021 [46]	Investigative	Sensitivity, PPV, and AUC	Not reported	Complete	Diagnostic and prediction
Lu et al, 2022 [47]	Investigative	Not reported	Not reported	Not reported	Not reported
Vasey et al, 2022 [49]	Investigative	Not reported	Not reported	Not reported	Not reported
Abdulazeem et al, 2023 [50]	Investigative	Not reported	AUROC[f], sensitivity, specificity, predictive values, accuracy	Not reported	Diagnostic and prediction
Cacciamani et al, 2023 [15]	Investigative	Not reported	Not reported	Not reported	Not reported
Debray et al, 2023 [51]	Investigative	Not reported	Not reported	Wrong	Prediction
Elvidge et al, 2023 [52]	Investigative	Not reported	Not reported	Not reported	Not reported
Klement et al, 2023 [53]	Investigative	Not reported	Not reported	Not reported	Diagnostic and prediction
Kwong et al, 2023 [54]	Investigative	Not reported	Not reported	Not reported	Prediction
Murphy et al, 2023 [55]	Investigative	AUC	Sensitivity and specificity	Not reported	Not reported
Collins et al, 2024 [56]	Investigative	Not reported	Not reported	Not reported	Diagnostic and prediction
Elfer et al, 2024 [57]	Investigative	Not reported	Not reported	Not reported	Diagnostic
El Emam et al, 2024 [58]	Investigative	Not reported	Not reported	Not reported	Diagnostic and prediction
Guni et al, 2024 [59]	Investigative	Not reported	Not reported	Not reported	Not reported
Kapoor et al, 2024 [60]	Investigative	Not reported	Not reported	Not reported	Diagnostic and prediction
Labkoff et al, 2024 [61]	Investigative	Not reported	Not reported	Not reported	Not reported
Masters and Salcedo, 2024 [62]	Investigative	Not reported	Not reported	Not reported	Not reported
Ray et al, 2024 [63]	Investigative	AUC	Sensitivity, specificity, accuracy, and precision	Not reported	Diagnostic
Tejani et al, 2024 [13]	Investigative	Sensitivity	Not reported	Incomplete	Diagnostic and prediction
Uribe et al, 2024 [64]	Investigative	Not reported	Not reported	Not reported	Not reported
Warren et al, 2024 [65]	Investigative	Not reported	Not reported	Not reported	Not reported
Kalaycioglu et al, 2025 [66]	Investigative	Not reported	HF 30-d mortality; AMI[g] in-hospital death	Not reported	Prediction
PAHO[h] 2025 [67]	Investigative	Not reported	Not reported	Not reported	Not reported
Sounderajah et al, 2025 [14]	Investigative	Not reported	Not reported	Not reported	Diagnostic
Tuygunov et al, 2025 [68]	Not reported	Not reported	Not reported	Not reported	Not reported
Wang et al, 2025 [69]	Investigative	Not reported	Not reported	Not reported	Not reported

[a]AUC: area under the curve.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]NNT: number needed to treat.
[e]ROC: receiver operating characteristic.
[f]AUROC: area under the receiver operating characteristic curve.
[g]AMI: acute myocardial infarction.
[h]PAHO: Pan American Health Organization.

The evaluation frameworks revealed substantial heterogeneity in how studies articulate their objectives, define clinical purpose, and report technical or clinical performance parameters. Most frameworks were investigational in nature (approximately 88%, 39/44), whereas only 4 frameworks explicitly stated a focus on assessing clinical utility (eg, Reps et al [30], Hernandez-Boussard et al [32], Ji et al [40], and Olczak et al [41]). This distribution indicates that

most published frameworks remain exploratory rather than implementation-oriented.

Across studies, technical performance reporting was inconsistent. Only 31.8% (14/44) provided at least 1 performance metric. The most frequently reported indicators were AUC or ROC values, present in 12/44 studies, followed by sensitivity and specificity. Less commonly, frameworks used positive predictive values and negative predictive values, accuracy, or segmentation metrics such as the F_1-score or Dice coefficient (eg, Norgeot et al [35] and Olczak et al [41]). A small subset incorporated measures of calibration accuracy, number needed to treat, or task-specific metrics (eg, 30-day mortality in Kalaycıoğlu et al [66]).

Reporting of clinical effectiveness metrics was even more limited. Only 7 (15.9%) studies explicitly referenced indicators of clinical performance, most commonly predictive values, accuracy, or calibration. The overwhelming majority did not report any clinical effectiveness parameters, suggesting that most frameworks remain focused on model development and internal performance rather than real-world clinical applicability.

When classifying the rigor of evaluation (technical performance), only 5 (11.4%) frameworks were categorized as providing a complete evaluation, integrating internal and external validation components (eg, Luo et al [28], Hernandez-Boussard et al [32], Mongan et al [34], Norgeot et al [35], Kwong et al 2021, and Debray et al [51]). An additional 5 (11.4%) studies were judged as incomplete, and 3 studies were rated as incorrect (“wrong”) due to methodological inconsistencies or misalignment between reported metrics and intended clinical purpose.

Regarding clinical purpose, approximately one-third (9/44) of the frameworks focused on diagnostic applications, while 7 of 44 addressed prediction. Another subset (7/44) incorporated both diagnostic and prognostic applications.

Domain 3: Ethical Alignment With UNESCO Principles

Analysis of the ethical dimensions of the included frameworks revealed substantial variability in their alignment with the 10 UNESCO AI ethical principles. Overall compliance was modest, with only a small group of documents achieving high scores (≥80%), including works by Floridi et al [29], Shen et al [48], Masters and Salcedo [62], Ning et al [17], and the PAHO [67], each demonstrating broad and explicit ethical integration. In contrast, four documents (Luo et al [28], Stevens et al [37], Abdulazeem et al [50], and Debray et al [51]) showed no or minimal ethical reporting, scoring below 10%. Across principles, the highest levels of adherence were observed for awareness and education (P7, 71.1%), transparency and explainability (P9, 70%), and proportionality and safety (P2, 63.3%), indicating a stronger emphasis on communicability, system clarity, and risk mitigation. Conversely, human oversight (P4, 24.4%) and adaptive, multistakeholder governance (P8, 33.3%) were the least addressed, reflecting limited incorporation of human-in-the-loop safeguards and participatory governance structures. Only a minority of frameworks exhibited balanced coverage across all principles, revealing a fragmented ethical landscape where most documents address isolated ethical components rather than offering comprehensive, principle-wide guidance (Figure 2 [13-15,17,28-69]).

Figure 2. Percentage of compliance with United Nations Educational, Scientific and Cultural Organization ethical principles across AI evaluation frameworks [13-15,17,28-69]. P1: equity and nondiscrimination; P2: proportionality and safety; P3: sustainability; P4: human oversight and decision-making; P5: right to privacy and data protection; P6: responsibility and accountability; P7: awareness and education; P8: adaptive and multistakeholder governance and collaboration; P9: transparency and explainability; P10: safety and security; 0: no report; 1: indirectly; 2: directly; PAHO: Pan American Health Organization; UNESCO: United Nations Educational, Scientific and Cultural Organization.

Author, Year [Ref.]	UNESCO PRINCIPE										Total score	Percentage score (%)
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10		
Luo W., 2016 [29]	0	0	0	0	0	0	0	0	0	0	0	0
Floridi L., 2018 [30]	2	2	2	1	1	2	1	2	2	1	16	80
Reps J., 2018 [31]	1	1	1	1	1	0	2	0	2	0	9	45
Cruz-Rivera S., 2020 [32]	1	2	0	2	0	1	1	0	0	0	7	35
Hernandez-Boussard T., 2020 [33]	0	2	1	0	1	0	2	0	1	1	8	40
Liu X., 2020 [34]	1	0	1	1	1	1	1	1	0	2	9	45
Mongan J., 2020 [35]	1	2	1	0	0	1	2	0	1	1	9	45
Norgeot B., 2020 [36]	1	2	1	0	0	1	2	0	1	1	9	45
Sengupta PP., 2020 [37]	1	1	1	0	1	1	1	0	1	1	8	40
Stevens L., 2020 [38]	0	0	0	0	0	0	0	0	0	0	0	0
Young A., 2020 [39]	2	2	2	0	2	1	1	0	1	0	11	55
Cabitza, F. 2021 [40]	0	2	2	0	2	1	1	0	2	1	11	55
Ji M., 2021 [41]	1	1	1	0	0	1	2	0	2	1	9	45
Olczak, J. 2021 [42]	0	0	0	0	2	0	2	2	0	2	8	40
Schwendicke F., 2021 [43]	0	2	2	0	0	2	2	0	2	0	10	50
Bazoukis G., 2022 [44]	2	2	2	0	1	2	1	1	2	0	13	65
Daneshjou, R. 2022 [45]	0	0	2	0	2	0	2	2	2	2	12	60
Fusar-Poli P., 2022 [46]	2	2	2	2	2	0	0	2	0	0	12	60
Kwong J., 2022 [47]	1	1	2	0	1	1	2	1	2	1	12	60
Lu J., 2022 [48]	0	1	1	0	0	1	2	0	2	1	8	40
Shen FX., 2022 [49]	2	2	2	2	2	2	2	2	0	0	16	80
Vasey, B. 2022 [50]	2	0	2	2	2	2	2	2	0	0	14	70
Abdulaz eem H., 2023 [51]	0	0	0	0	0	0	1	0	1	0	2	10
Cacciamani GE.,2023 [52]	1	1	1	0	1	1	2	0	2	1	10	50
Debray T., 2023 [53]	0	0	0	0	0	0	0	0	1	0	1	5
Elvidge, J. 2023 [54]	2	2	2	2	0	2	0	2	0	0	12	60
Klement, W. 2023 [55]	1	1	1	0	0	1	2	1	2	0	9	45
Kwong J., 2023 [56]	1	2	1	0	1	1	2	0	2	1	11	55
Murphy T., 2023 [57]	2	2	1	2	1	1	1	0	2	0	12	60
Collins, G. 2024 [58]	1	1	2	1	0	1	2	1	2	0	11	55
Elfer K., 2024 [59]	0	0	0	0	0	1	2	0	2	0	5	25
El Emam K., 2024 [60]	1	1	1	0	1	1	2	2	2	0	11	55
Guni A., 2024 [61]	0	0	0	0	0	0	2	0	2	1	5	25
Kapoor, S. 2024 [62]	1	2	1	0	1	1	2	0	2	1	11	55
Labkoff S.,2024 [63]	2	2	2	0	1	2	1	1	2	0	13	65
Masters K. y Salcedo D., 2024 [64]	1	2	1	2	2	1	2	1	2	2	16	80
Ning Y., 2024 [17]	2	2	2	1	2	2	1	2	2	2	18	90
Ray A., 2024 [65]	0	1	0	0	0	0	2	0	1	0	4	20
Tejani AS., 2024 [66]	1	2	1	0	2	1	2	0	2	1	12	60
Uribe S.E. 2024 [67]	1	1	0	0	0	2	1	0	2	0	7	35
Warren B., 2024 [68]	2	2	1	0	1	1	1	1	2	2	13	65
Kalaycioglu O., 2025 [69]	0	0	0	0	0	0	2	0	1	0	3	15
PAHO*, 2025 [70]	2	2	2	1	2	2	1	2	2	2	18	90
Souderajah V., 2025 [14]	2	1	0	2	1	1	1	2	2	1	13	65
Tuygunov N., 2025 [71]	2	2	2	0	1	2	1	0	2	2	14	70
Wang Y., 2025 [72]	0	1	1	0	0	0	2	0	2	1	7	35
Total	**45**	**57**	**50**	**22**	**38**	**44**	**64**	**30**	**63**	**32**	-	-
Percentage of compliance (%)	**50**	**63,3**	**55,6**	**24,4**	**42,2**	**48,9**	**71,1**	**33,3**	**70**	**35,6**	-	-

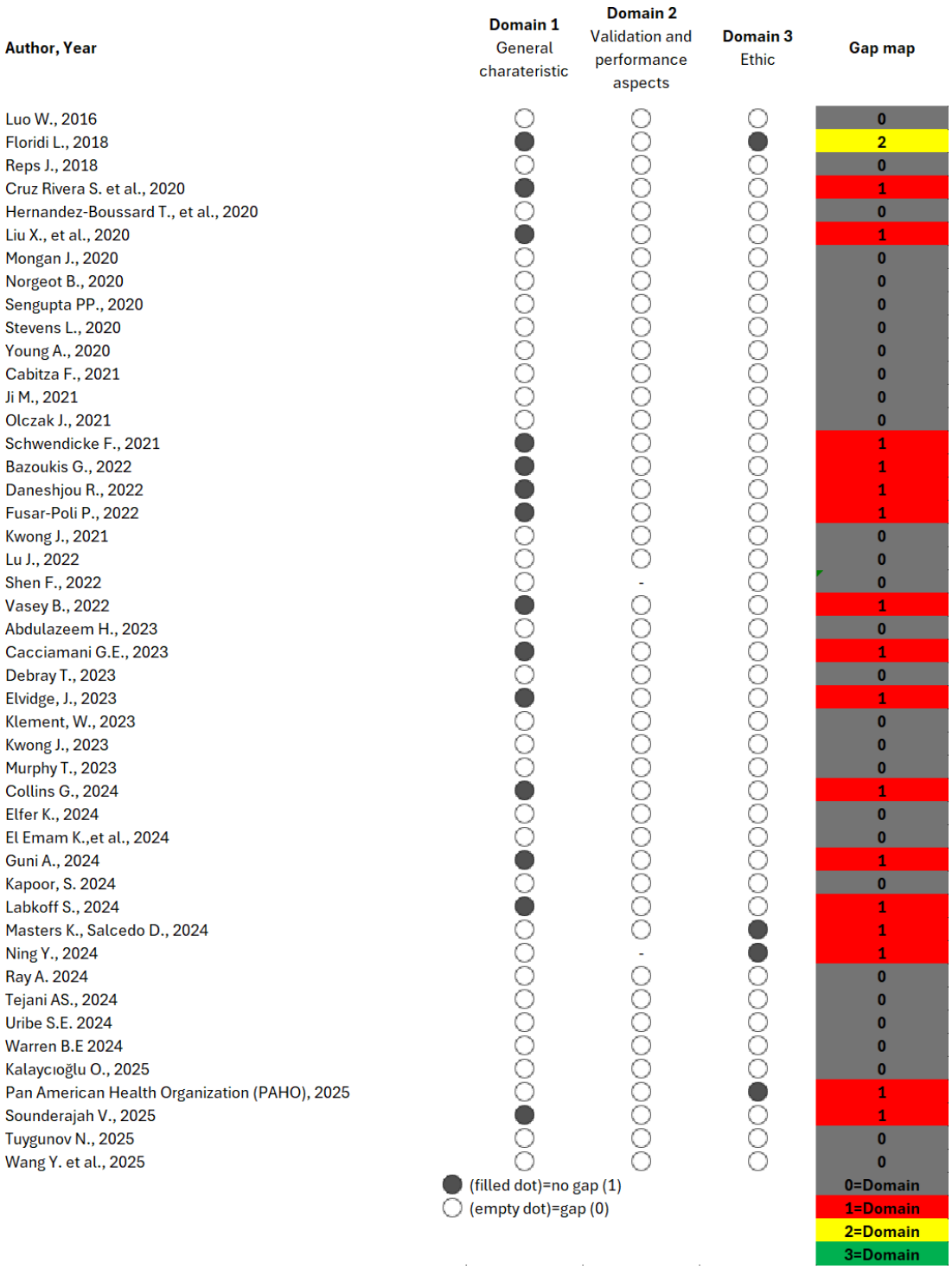
P1= equity and non-discrimination. P2= proportionality and safety. P3= sustainability. P4= human oversight and decision-making. P5= right to privacy and data protection. P6= responsibility and accountability. P7= awareness and education. P8= adaptive and multi-stakeholder governance and collaboration. P9= transparency and explainability. P10= safety and security.

0=no report, 1=indirectly, 2=directly.
*PAHO= Pan American Health Organization

The gap map presented in Figure 3 [13-15,17,28-69] shows that most existing AI evaluation frameworks primarily emphasize methodological and technical aspects of model performance. In contrast, comparatively less attention is given to clinical applicability and ethical considerations. The distribution of data points across the 3 evaluated domains suggests that these frameworks tend to prioritize reporting transparency and reproducibility, while aspects related to clinical utility, external validation, and ethical safeguards for real-world implementation remain insufficiently addressed.

Across the three domains evaluated (general characteristics, validations and performance aspects, and ethics), most frameworks presented substantial gaps. Specifically, 63% (29/46) of frameworks showed gaps across all 3 domains. In contrast, 34.8% (16/46) of the frameworks fully complied with the criteria in at least 1 domain. Among these, 3 frameworks met the criteria for the ethics domain (Floridi et al [29], Masters and Salcedo [62], and Ning et al [17]), while 13 complied with the general characteristics domain. Only 2.2% (1/46) of the frameworks met the high-quality criteria in 2 domains (general characteristics and ethics). Importantly, none of the evaluated frameworks met the high-quality criteria for the validation and performance aspects domain, and no framework fulfilled the criteria across all 3 domains simultaneously.

Figure 3. Gap map of evaluation frameworks for clinical AI across three domains: general characteristics, validation and performance, and ethical integration. Dot plot representation indicates the presence (1) or absence (0) of gaps according to predefined criteria [13-15,17,28-69].



Discussion

This scoping review identified a fragmented landscape of evaluation frameworks for clinical AI, characterized by heterogeneous methodological approaches, limited integration of external validation aligned with intended clinical use, and inconsistent incorporation of ethical principles. These findings directly address the study objectives by highlighting gaps in methodological rigor, real-world applicability, and ethical alignment. Ethical integration was inconsistently addressed, with most frameworks covering only a limited subset of UNESCO ethical domains.

These findings can be better understood in the context of the rapid expansion of methodological and reporting frameworks for clinical AI [6]. This pattern is consistent with the orientation of several influential frameworks included in this review, many of which were developed primarily to improve transparency, reproducibility, and completeness of reporting, while providing more limited coverage of implementation-related issues [72]. Collectively, this proliferation of frameworks reflects a broad consensus regarding the need to promote transparency, rigor, and reproducibility in medical AI research [29,34,49]. Minimum Information for Medical AI Reporting (MINIMAR), for example, was proposed to improve transparent reporting of the design, development, evaluation, and validation of medical AI models, with explicit emphasis on replication, external validation, and the identification of potential biases and unintended consequences [32]. Likewise, Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) was designed to enable assessment of clinical impact, fairness, and bias while facilitating replication of the technical design process of clinical AI studies [35]. In the same direction, Klement and El Emam [53] showed that existing high-quality guidelines did not individually provide complete coverage, supporting the fragmented landscape identified in our gap map. However, our findings suggest that this expansion has occurred without corresponding consolidation of validation standards or clinical applicability.

The predominance of technical and reporting-oriented domains over broader clinical dimensions in our review also aligns with previous literature showing that strong model performance does not necessarily translate into clinical benefit. Lu et al [47] found inconsistent adherence to reporting recommendations in deployed clinical prediction models, particularly for reliability-related items such as external validation, uncertainty, monitoring, fairness, and transparency. This broader concern is also reflected in DECIDE-AI, which notes that strong performance in preclinical studies has not yet been matched by high-quality evidence of improved clinician performance or patient outcomes in clinical settings [49]. Taken together, these findings suggest that current evaluation of clinical AI remains more mature at the level of methodological description than at the level of demonstrated clinical usefulness [46,48].

A related implication is that methodological rigor in clinical AI should not be interpreted narrowly as internal validation or reporting compliance. SPIRIT-AI and CONSORT-AI both emphasize that AI interventions should be described in relation to their intended use, intended users, integration into the clinical pathway, required expertise, handling of input data, output interpretation, and the way outputs contribute to downstream clinical decision-making [31,33]. CONSORT-AI also requires reporting of performance-error analysis and highlights the importance of documenting access to the intervention or its code, while SPIRIT-AI recommends explicit planning for performance-error identification and description of implementation requirements in the trial setting [31,33]. Together, these frameworks support the view that robust evaluation of clinical AI requires attention not only to methodological transparency but also to usability, context of use, and conditions for safe implementation.

Our finding that real-world applicability was less consistently represented is also supported by frameworks explicitly designed to address translational stages of evaluation [73,74]. DECIDE-AI was developed as a stage-specific reporting guideline for early, small-scale, live clinical evaluation and focuses on proof of clinical utility, safety, human factors, and preparation for larger-scale studies [49]. Its rationale explicitly refers to the gap between mathematical performance and clinical utility and to the need to address challenges such as dataset shift, user variability, and implementation in live clinical settings [49]. More recently, Labkoff et al [61] argued that responsible AI-enabled clinical decision support requires a more comprehensive framework spanning model training, explainability, validation, certification, monitoring, continuous evaluation, privacy, fairness, and regulatory oversight. These proposals are consistent with the pattern observed in our review: implementation-relevant domains are increasingly recognized in the literature, but they remain unevenly operationalized across available frameworks.

The limited number of transdisciplinary studies in our review is also noteworthy. Some of the more mature frameworks were developed through broad multistakeholder processes. DECIDE-AI was produced through an international consensus process involving multiple stakeholder groups [49] and Labkoff et al [61] similarly reported a consensus process including clinicians, software developers, academics, ethicists, attorneys, policy experts, scientists, and patients. However, our mapping suggests that such breadth of stakeholder participation is not yet typical across the field. This matters because clinical AI is not only a technical intervention, but also a clinical, organizational, ethical, and regulatory one. Frameworks developed from narrower perspectives may therefore be less able to capture the full conditions needed for trustworthy implementation [75].

The ethical findings in this review are likewise consistent with the literature included in the scoping review. Floridi et al [29] synthesized major AI ethics initiatives around beneficence, nonmaleficence, autonomy, justice, and explicability, arguing that explicability combines intelligibility and accountability and is necessary to make the other principles actionable. Their discussion of justice also extends

beyond discrimination to shared benefit, equal access, and protection of social structures such as health care systems [29]. In more applied health care literature, Shen et al [48] proposed an ethics checklist structured around informed consent, equity, diversity and access, privacy and partnerships, regulation and law, return of results, and duty to warn and report [48]. Ning et al [17] similarly proposed the TREGAI (Transparent Reporting of Ethics for Generative Artificial Intelligence) checklist for generative AI in health care, derived from a scoping review and intended to operationalize ethical evaluation within practical guidance. These studies support our interpretation that ethical concerns are clearly present in the literature, but are not yet incorporated with comparable depth or consistency across clinical AI evaluation frameworks. This limitation is particularly relevant for clinical implementation, as insufficient ethical integration may undermine patient safety, public trust, and regulatory acceptance of AI systems in health care.

A similar pattern can be seen in more implementation-oriented proposals. Bazoukis et al [43] describe a regulatory framework for augmented intelligence in medicine that includes accountability, liability, equity and inclusion, transparency, explainability, education, patient engagement, cybersecurity and privacy, ethics and fairness, as well as safety and postmarket surveillance. Labkoff et al [61] also emphasize transparency, documentation, training, validation, certification, monitoring, safety reporting, fairness, and privacy as necessary elements of responsible AI-enabled clinical decision support. In our review, however, these ethical- and governance-related elements were not uniformly represented across frameworks. This suggests that the current landscape remains fragmented: ethical principles are increasingly acknowledged, but their operational integration into evaluation standards is still incomplete.

Taken together, the three domains examined in this scoping review reveal a recurrent pattern: methodological rigor, clinical applicability, and ethical integration do not consistently converge within individual frameworks. Many frameworks are strong in reporting transparency or technical assessment, but are less explicit regarding external validation, implementation conditions, real-world use, or comprehensive ethical governance. This fragmentation appears to be a structural feature of the current evaluative landscape rather than a weakness of any single framework. It likely reflects the parallel evolution of technical reporting, clinical evaluation, and ethical guidance in AI, which have not yet been fully harmonized within a single evaluative model [29,31-35,43,49,53,61].

Limitations

As with any scoping review, this study has inherent limitations. First, as a scoping review, its purpose was to map and characterize the available literature rather than to compare the superiority or effectiveness of individual frameworks [19,76]. Therefore, the findings should be interpreted as an overview of patterns, emphases, and gaps in the field, not as evidence that one framework performs better than another. Second, the included publications were heterogeneous in scope, purpose, and intended stage of use, ranging from reporting guidelines and consensus statements to conceptual ethics papers and implementation-oriented proposals [29,31,33,34,49,56]. Although this heterogeneity is informative for evidence mapping, it limits direct comparability across sources [15,34,40,45,46].

Third, the gap map required interpretive judgment when classifying whether a framework addressed specific methodological, clinical, or ethical domains, particularly when such issues were mentioned implicitly rather than operationalized explicitly [37,50,51]. For that reason, the categorizations presented here should be understood as a structured synthesis of the literature rather than a definitive ranking. Finally, many of the included publications propose how clinical AI should be evaluated, but do not prospectively test the effect of those frameworks in routine clinical settings [40,54,77]. Accordingly, this review is stronger in identifying what the literature currently prioritizes and omits than in determining how well these frameworks perform in practice [49,61].

Conclusions

In conclusion, this scoping review indicates that current evaluation and reporting frameworks for clinical AI are strongest in technical and reporting-oriented domains, but less consistently address broader validation strategies, real-world applicability, and comprehensive ethical integration. This interpretation is consistent with the included literature, which has emphasized transparency, reproducibility, intended use, validation, fairness, and safety, while also recognizing the persistent gap between algorithm development and responsible implementation in clinical settings.

The broader implication is that advancing clinical AI will require a shift from fragmented, reporting-oriented approaches toward integrated evaluation frameworks that align methodological rigor, real-world validation, and ethical governance. Future frameworks should move beyond checklist-based reporting to support implementation, regulatory decision-making, and continuous monitoring in clinical settings. Without such integration, the translation of AI innovations into routine health care practice will remain limited.

Acknowledgments

As this study is based on previously published and publicly accessible information, it did not require approval from an ethics committee. However, principles of transparency and methodological rigor were ensured throughout the selection and analysis of the reviewed documents. The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: translation

and preparation of press releases and outreach materials. The GAI tools used were ChatGPT-4.5 (OpenAI) and DALL·E 2 (OpenAI). Responsibility for the final paper lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration was submitted by the authors under collective responsibility. The authors used ChatGPT (OpenAI, San Francisco, CA, USA) to translate and improve the wording of the paper, and DALL·E 2 (OpenAI, San Francisco, CA, USA) was used to generate images included in the paper.

Funding

This study was supported by the Universidad Cooperativa de Colombia through the institutional research project (Project INV3688). The funder had no involvement in the study design, data collection, data analysis, interpretation of the findings, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated during this study, including the data extraction table used for the scoping review, are available from the corresponding author upon reasonable request. All primary sources analyzed are publicly available and are cited in the manuscript and supplementary materials. All data are available in Multimedia Appendices 1 and 2.

Authors' Contributions

DCLM and NMA conceptualized the study. DCLM developed the methodology. DCLM, MH-P, ACH-A, and NMA contributed to the software. Validation was performed by MH-P, DCLM, and AON. Formal analysis was conducted by MH-P, DCLM, AON, and NMA. MH-P, DCLM, AON, and ACH-A contributed to the investigation. DCLM and AON provided resources. Data curation was carried out by MH-P, DCLM, and AON. DCLM and NMA prepared the original draft of the paper. MH-P and DCLM reviewed and edited the paper. Visualization was undertaken by MH-P and DCLM. DCLM and AON supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional methodological materials and extracted data supporting the scoping review.
[DOCX File (Microsoft Word File), 87 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Excluded documents.
[DOCX File (Microsoft Word File), 3218 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-ScR checklist.
[PDF File (Adobe File), 661 KB-Checklist 1]

Checklist 2

PRISMA 2020 Abstract checklist.
[PDF File (Adobe File), 296 KB-Checklist 2]

REFERENCES

1. Avila-Tomás JF, Mayer-Pujadas MA, Quesada-Varela VJ. La inteligencia artificial y sus aplicaciones en medicina I: introducción antecedentes a la IA y robótica [Article in Spanish]. Atención Primaria. Dec 2020;52(10):778-784. [doi: 10.1016/j.aprim.2020.04.013]
2. Ávila-Tomás JF, Mayer-Pujadas MA, Quesada-Varela VJ. La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas [Article in Spanish]. Atención Primaria. Jan 2021;53(1):81-88. [doi: 10.1016/j.aprim.2020.04.014]
3. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. Feb 27, 2020;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Jan 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
5. Kapa S. The role of artificial intelligence in the medical field. J Comput Commun. 2023;11(11):1-16. [doi: 10.4236/jcc.2023.1111001]
6. 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] [Medline: 38605106]

7. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res. Jul 7, 2020;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
8. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. Mar 25, 2020;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
9. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
10. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Oct 29, 2019;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
11. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024:q902. [doi: 10.1136/bmj.q902]
12. 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] [Medline: 32909959]
13. 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] [Medline: 38809149]
14. 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] [Medline: 40954311]
15. Cacciamani GE, Chu TN, Sanford DI, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med. Jan 2023;29(1):14-15. [doi: 10.1038/s41591-022-02139-w] [Medline: 36646804]
16. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. 1st ed. Basic Books; 2019. ISBN: 978-1-5416-4464-9
17. Ning Y, Teixayavong S, Shang Y, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health. Nov 2024;6(11):e848-e856. [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
18. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025. URL: https://www.who.int/publications/i/item/9789240084759 [Accessed 2026-06-17]
19. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
20. 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] [Medline: 33782057]
21. Homepage. EQUATOR Network. URL: https://tinyurl.com/49dt83rh [Accessed 2025-11-27]
22. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. Korean J Radiol. Mar 2021;22(3):442-453. [doi: 10.3348/kjr.2021.0048] [Medline: 33629545]
23. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. Oct 1978;8(4):283-298. [doi: 10.1016/s0001-2998(78)80014-2] [Medline: 112681]
24. Choi JS, Han BK, Ko ES, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. Korean J Radiol. May 2019;20(5):749-758. [doi: 10.3348/kjr.2018.0530] [Medline: 30993926]
25. Mutasa S, Sun S, Ha R. Understanding artificial intelligence based radiology studies: What is overfitting? Clin Imaging. Sep 2020;65:96-99. [doi: 10.1016/j.clinimag.2020.04.025] [Medline: 32387803]
26. Nyanchoka L, Tudur-Smith C, Thu VN, Iversen V, Tricco AC, Porcher R. A scoping review describes methods used to identify, prioritize and display gaps in health research. J Clin Epidemiol. May 2019;109:99-110. [doi: 10.1016/j.jclinepi.2019.01.005] [Medline: 30708176]
27. Ethics of Artificial Intelligence. UNESCO; 2025. URL: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [Accessed 2025-11-27]
28. 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] [Medline: 27986644]
29. Floridi L, Cowls J, Beltrametti M, et al. AI4People—an ethical framework for a Good AI Society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr). 2018;28(4):689-707. [doi: 10.1007/s11023-018-9482-5] [Medline: 30930541]

30. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. Aug 1, 2018;25(8):969-975. [doi: 10.1093/jamia/ocy032] [Medline: 29718407]
31. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. Oct 2020;2(10):e549-e560. [doi: 10.1016/S2589-7500(20)30219-3] [Medline: 33328049]
32. 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] [Medline: 32594179]
33. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. Oct 2020;2(10):e537-e548. [doi: 10.1016/S2589-7500(20)30218-1] [Medline: 33328048]
34. 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] [Medline: 33937821]
35. 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] [Medline: 32908275]
36. Sengupta PP, Shrestha S, Berthon B, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): a checklist. JACC Cardiovasc Imaging. Sep 2020;13(9):2017-2035. [doi: 10.1016/j.jcmg.2020.07.015]
37. 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] [Medline: 33079589]
38. Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: a primer. J Invest Dermatol. Aug 2020;140(8):1504-1512. [doi: 10.1016/j.jid.2020.02.026] [Medline: 32229141]
39. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics. Int J Med Inform. Sep 2021;153:104510. [doi: 10.1016/j.ijmedinf.2021.104510]
40. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation framework for successful artificial intelligence-enabled clinical decision support systems: mixed methods study. J Med Internet Res. Jun 2, 2021;23(6):e25929. [doi: 10.2196/25929] [Medline: 34076581]
41. 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] [Medline: 33988081]
42. Schwendicke F, Singh T, Lee JH, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent. Apr 2021;107:103610. [doi: 10.1016/j.jdent.2021.103610]
43. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA. The inclusion of augmented intelligence in medicine: a framework for successful implementation. Cell Rep Med. Jan 18, 2022;3(1):100485. [doi: 10.1016/j.xcrm.2021.100485] [Medline: 35106506]
44. Daneshjou R, Barata C, Betz-Stablein B, et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm Consensus Guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group. JAMA Dermatol. Jan 1, 2022;158(1):90-96. [doi: 10.1001/jamadermatol.2021.4915] [Medline: 34851366]
45. Fusar-Poli P, Manchia M, Koutsouleris N, et al. Ethical considerations for precision psychiatry: a roadmap for research and clinical practice. Eur Neuropsychopharmacol. Oct 2022;63:17-34. [doi: 10.1016/j.euroneuro.2022.08.001] [Medline: 36041245]
46. Kwong JCC, McLoughlin LC, Haider M, et al. Standardized reporting of machine learning applications in urology: the STREAM-URO framework. Eur Urol Focus. Jul 2021;7(4):672-682. [doi: 10.1016/j.euf.2021.07.004] [Medline: 34362709]
47. Lu JH, Callahan A, Patel BS, et al. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: a systematic review. JAMA Netw Open. Aug 1, 2022;5(8):e2227779. [doi: 10.1001/jamanetworkopen.2022.27779] [Medline: 35984654]
48. Shen FX, Silverman BC, Monette P, Kimble S, Rauch SL, Baker JT. An ethics checklist for digital health research in psychiatry: viewpoint. J Med Internet Res. Feb 9, 2022;24(2):e31146. [doi: 10.2196/31146] [Medline: 35138261]
49. 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] [Medline: 35584845]

50. Abdulazeem H, Whitelaw S, Schauberger G, Klug SJ. A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. PLoS One. 2023;18(9):e0274276. [doi: 10.1371/journal.pone.0274276] [Medline: 37682909]
51. 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] [Medline: 36750236]
52. 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] [Medline: 38795956]
53. 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] [Medline: 37651179]
54. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open. Sep 5, 2023;6(9):e2335377. [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]
55. Murphy TI, Armitage JA, van Wijngaarden P, Abel LA, Douglass AG. A guide to optometrists for appraising and using artificial intelligence in clinical practice. Clin Exp Optom. Aug 2023;106(6):569-579. [doi: 10.1080/08164622.2023.2197578] [Medline: 37078176]
56. 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] [Medline: 38626948]
57. Elfer K, Gardecki E, Garcia V, et al. Reproducible reporting of the collection and evaluation of annotations for artificial intelligence models. Mod Pathol. Apr 2024;37(4):100439. [doi: 10.1016/j.modpat.2024.100439] [Medline: 38286221]
58. El Emam K, Leung TI, Malin B, Klement W, Eysenbach G. Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS). J Med Internet Res. May 2, 2024;26:e52508. [doi: 10.2196/52508] [Medline: 38696776]
59. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): protocol for a qualitative study. JMIR Res Protoc. Sep 18, 2024;13:e58202. [doi: 10.2196/58202] [Medline: 39293047]
60. 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] [Medline: 38691601]
61. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. Journal of the American Medical Informatics Association. Nov 1, 2024;31(11):2730-2739. [doi: 10.1093/jamia/ocae209]
62. Masters K, Salcedo D. A checklist for reporting, reading and evaluating Artificial Intelligence Technology Enhanced Learning (AITEL) research in medical education. Med Teach. Sep 2024;46(9):1175-1179. [doi: 10.1080/0142159X.2023.2298756]
63. Ray A, Sarkar S, Schwenker F, Sarkar R. Decoding skin cancer classification: perspectives, insights, and advances through researchers' lens. Sci Rep. Dec 18, 2024;14(1):30542. [doi: 10.1038/s41598-024-81961-3] [Medline: 39695157]
64. Uribe SE, Maldupa I, Schwendicke F. Integrating generative AI in dental education: a scoping review of current practices and recommendations. Eur J Dental Education. May 2025;29(2):341-355. [doi: 10.1111/eje.13074]
65. Warren BE, Bilbily A, Gichoya JW, et al. An introductory guide to artificial intelligence in interventional radiology: part 2: implementation considerations and harms. Can Assoc Radiol J. Aug 2024;75(3):568-574. [doi: 10.1177/08465371241236377] [Medline: 38445517]
66. Kalaycıoğlu O, Pavlou M, Akhanlı SE, de Belder MA, Ambler G, Omar RZ. Evaluating the sample size requirements of tree-based ensemble machine learning techniques for clinical risk prediction. Stat Methods Med Res. Jul 2025;34(7):1356-1372. [doi: 10.1177/09622802251338983] [Medline: 40368385]
67. Pan American Health Organization (PAHO). AI prompt design for public health: Using generative AI responsibly - OPS/OMS. Organización Panamericana de la Salud. Pan American Health Organization. Washington: Pan American Health Organization; 2025. URL: https://www.paho.org/es/documentos/ai-prompt-design-public-health-using-generative-ai-responsibly [Accessed 2025-11-28]
68. Tuygunov N, Samaranayake L, Khurshid Z, et al. The transformative role of artificial intelligence in dentistry: a comprehensive overview part 2: the promise and perils, and the International Dental Federation communique. Int Dent J. Apr 2025;75(2):397-404. [doi: 10.1016/j.identj.2025.02.006] [Medline: 40011130]

69. Wang Y, Cheungpasitporn W, Ali H, et al. A practical guide for nephrologist peer reviewers: evaluating artificial intelligence and machine learning research in nephrology. Ren Fail. Dec 31, 2025;47(1). [doi: 10.1080/0886022X.2025.2513002]
70. Page MJ, McKenzie JE, Bossuyt PM, et al. Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. Revista Española de Cardiología. Sep 2021;74(9):790-799. [doi: 10.1016/j.recesp.2021.06.016]
71. Wahab N, Miligy IM, Dodd K, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. J Pathol Clin Res. Mar 2022;8(2):116-128. [doi: 10.1002/cjp2.256] [Medline: 35014198]
72. 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. Dec 26, 2024;8(1):ae155. [doi: 10.1093/jamiaopen/ooae155]
73. Arshi B, Cowley LE, Rijnhart E, Reeve K, Smits LJ, Wynants L. External validation, impact assessment and clinical utilization of clinical prediction models: a prospective cohort study. J Clin Epidemiol. Oct 2025;186:111902. [doi: 10.1016/j.jclinepi.2025.111902] [Medline: 40675230]
74. Wainstein M, Flanagan E, Johnson DW, Shrapnel S. Systematic review of externally validated machine learning models for predicting acute kidney injury in general hospital patients. Front Nephrol. 2023;3:1220214. [doi: 10.3389/fneph.2023.1220214] [Medline: 37675372]
75. Kotter E, D'Antonoli TA, Cuocolo R, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. Insights Imaging. Feb 13, 2025;16(1):33. [doi: 10.1186/s13244-025-01905-x] [Medline: 39948192]
76. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol. Jul 2016;75:40-46. [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
77. Topol EJ, Verghese A. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books; 2019. ISBN: 978-1-5416-4463-2

Abbreviations

APPRAISE-AI: Assessment of Predictive Performance and Bias in Artificial Intelligence Studies
AUC: area under the curve
CHEERS-AI: Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence
CLAIM: Checklist for Artificial Intelligence in Medical Imaging
CONSORT-AI: Consolidated Standards of Reporting Trials for Artificial Intelligence
DECIDE-AI: Developmental and Evaluation Checklist for Decision Support Systems Driven by Artificial Intelligence
EQUATOR: Enhancing the Quality and Transparency of Health Research
FROC: free-response receiver operating characteristic
IMRD: Introduction, Methods, Results, and Discussion
MI-CLAIM: Minimum Information about Clinical Artificial Intelligence Modeling
MINIMAR: Minimum Information for Medical AI Reporting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-AI: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Artificial Intelligence
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ROC: receiver operating characteristic
SPIRIT-AI: Standard Protocol Items – Recommendations for Interventional Trials – Artificial Intelligence Extension
STARD-AI: Standards for Reporting of Diagnostic Accuracy – Artificial Intelligence
TREGAI: Transparent Reporting of Ethics for Generative Artificial Intelligence
TRIPOD+AI: Transparent Reporting of Multivariable Prediction Models with Artificial Intelligence
TRIPOD-Cluster: TRIPOD Extension for Clustered Data
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by Stefano Brini; peer-reviewed by Charlotte Ahmadu, Priyanshu Sharma; submitted 27.May.2025; final revised version received 30.Mar.2026; accepted 31.Mar.2026; published 22.Jul.2026

Please cite as:
López Medina DC, Oliveros-Navarro A, Moreno Angel N, Herrera- Arellano AC, Henao-Pérez M

Evaluation Frameworks for Clinical AI Incorporating Validation Strategies, Real-World Applicability, and Ethical Principles: Scoping Review
J Med Internet Res 2026;28:e78168
URL: https://www.jmir.org/2026/1/e78168
doi: 10.2196/78168