

Clinician-Facing Artificial Intelligence-Based Clinical Decision Support Systems Evaluated in Real-World Health Care Settings: A PubMed/MEDLINE-Based, PMC-Weighted Scoping Review and Evidence Map

Pınar Özel ^{1,*} and Ali Olamat ²

¹ Department of Software Development, Faculty of Computer and Information Technologies, Cappadocia University, Mustafapaşa Campus, Ürgüp, Nevşehir 50420, Türkiye

² Medical Engineering Department, Faculty of Engineering, Al-Ahliyya Amman University, Amman 19111, Jordan

* Correspondence: pinar.kiziloglu@kapadokya.edu.tr

Received: 12 July 2026; Accepted: 27 August 2026; Published: 15 October 2026

How to cite: Özel, P.; Olamat, A. Clinician-Facing Artificial Intelligence-Based Clinical Decision Support Systems Evaluated in Real-World Health Care Settings: A PubMed/MEDLINE-Based, PMC-Weighted Scoping Review and Evidence Map. *TIJFS* **2026**, *8* (1), 5. DOI: [10.37978/tijfs.v08i01.005](https://doi.org/10.37978/tijfs.v08i01.005)

Academic Editors: Muhammad Shoaib Akhtar



© 2026 by the TIJFS. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Background: Clinician-facing artificial intelligence-based clinical decision support systems (AI-CDSSs) are increasingly evaluated with real-world data, but evidence on implementation, human factors, responsible AI, and patient benefit remains unclear. The objective of this research is to map applications, technical characteristics, evaluation approaches, integration, outcomes, and implementation gaps of clinician-facing AI-CDSSs in real-world health care settings. **Methods:** PubMed/MEDLINE was searched on 15 June 2026 using the same strategy for each publication year from 2020 to 2026, changing only the date range. After duplicate removal and rule-based prescreening, available PMC XML and outside-PMC reports were evaluated against clinical, patient-oriented, practical, and translational criteria. AI-assisted screening and charting received targeted human review and TA200/FT150/DX120 quality control; full-corpus dual-human screening was not performed. **Results:** Of 14,097 canonical annual records, 14,083 unique records were screened; 5140 reports were sought and 3767 were considered for eligibility. The review included 1181 reports, mapping to 1180 candidate report families in the minimal duplicate-report check. Final stages were clinical validation, 741 (62.7%); prospective/user evaluation, 270 (22.9%); and implementation/impact evaluation, 170 (14.4%). Separate internal, external, multicenter and temporal validation extraction signals were detected in 707 (59.9%), 441 (37.3%), 312 (26.4%) and 27 (2.3%) reports, respectively. Safety, fairness/equity, explainability/interpretability and economic/resource evaluation signals were detected in 98 (8.3%), 61 (5.2%), 155 (13.1%) and 182 (15.4%) reports, respectively. **Conclusions:** Evidence remains dominated by validation rather than prospective, implementation, and impact studies. More comprehensive evaluation and reporting of temporal validation, technical integration, safety, fairness, explainability, economics, and post-deployment governance across settings, populations, devices, and clinical workflows are needed to demonstrate accountable clinical value beyond technical functionality.

Keywords: artificial intelligence; clinician-AI interaction; clinical decision support; implementation science; machine learning; real-world evidence; responsible AI

1. Introduction

Clinical decision support refers to computerized functions that provide patient-specific assessments, alerts, predictions, or suggestions at relevant points in care. Artificial intelligence-based clinical decision support systems (AI-CDSSs) generate these outputs with methods such as machine learning, deep learning, natural language processing, computer vision, probabilistic models, and generative AI. Because performance can change across populations, measurement processes, workflows, and institutions, clinician-facing AI-CDSS evaluation must address calibration, transportability, usability, safety, and the conditions in which clinicians encounter and act on model outputs [1,2].

Clinical decision support systems for clinicians (CDSSs) enable a variety of activities from diagnosis and identification to prognosis and risk stratification, treatment planning, triage, monitoring, prescribing, and workflow management. The potential future benefit is in improved consistency in pattern recognition, early identification of clinical decline, and reduced workload. However, functionality or prediction accuracy is not sufficient on its own to prove clinical value. Different types of diagnostic aid, prioritization algorithms, risk predictions, and treatment suggestions impact health care delivery differently and thus need their own evidence regarding error, threshold, oversight, accountability, and downstream effects [1,2].

The distinction between performance and impact is central to medical AI. DECIDE-AI is a reporting guideline for early-stage clinical evaluation of AI decision support; it emphasizes safety, human factors, and performance in the intended clinical context [1]. The same translational problem has been described as a chasm between model performance and clinical impact [2]. A systematic review and a scoping review of randomized AI studies similarly found that favorable technical or task outcomes did not consistently translate into patient benefit [3,4].

Evidence on implementation is therefore critical. The systematic review shows that measures like adoption, diffusion, implementation costs, and sustainability are underreported in comparison to technical performances [5]. For example, the implementation of AI in the field of medical imaging could improve efficiency, but its pooled effects are heterogeneous and significantly dependent on the environment of workflow [6]. Implementation is hence defined as more of a sociotechnical process than an all-or-nothing phenomenon, including such aspects as timing and location of use, output presentation, user training, disagreement management, and influence on decision-making without extra burden.

Trust and human factors are vital when it comes to clinician-facing technology. Features like transparency, reliability, education, usability, engagement of all the stakeholders, fit within workflows, and accountability form the basis of trust [7]. While the feature of explainability could lead to appropriate use of the technology, most research regarding explainable AI tends to be limited in its methodology or disconnected from clinical practice [8]. Even though the explanation may be scientifically valid, it could still be ineffective if it is not provided at the right time, if it is difficult to understand, or if it creates cognitive load. Hence, evaluation must take into account the interaction of the four factors—system, user, task, and context.

Evidence on responsible-AI is still inconsistent. As highlighted by the aforementioned SWOT analysis, issues of privacy, transparency, safety, governance, and human supervision have been cited as unresolved translational challenges in clinical AI [9]. Evidence from the empirical literature shows how algorithmic bias can create inequities in the distribution of resources and health care itself [10,11]. In low- and middle-income health care systems, infrastructure and capacity limitations influence implementation [12]. Monitoring after deployment is nascent even though the clinical performance of algorithms can change because of dataset, calibration, process, or software shifts.

Ethical and regulatory considerations have become increasingly lifecycle-centric for technologies. The publications by the World Health Organization highlight ethical, human rights, risk–benefit analysis, governance, transparency, and evaluation considerations [13,14]. The U.S. Food and Drug Administration published its final guidance for predetermined change-control plans for software functions enabled by AI for devices in August 2025 and its final guidance for clinical decision-support software in January 2026; information about lifecycle and performance in the real world is used for evaluation and monitoring [15–17]. Publications from the International Medical Device Regulators Forum include principles for the clinical evaluation of software as a medical device and good machine-learning practice [18,19]. Regulation (EU) 2026/1744, the Digital Omnibus on AI, entered into force on 27 July 2026 and modified Regulation (EU) 2024/1689 [20].

Reporting and appraisal guidelines have progressed together. Guidelines such as CONSORT-AI and SPIRIT-AI have extended reporting for clinical trials and protocols to include AI-related inputs and outputs, human–AI interactions, error detection, and integration [21,22]. TRIPOD+AI and PROBAST+AI help in reporting and appraising the prediction model built via regression or machine learning [23,24]. STARD-AI and CLAIM have guidelines regarding diagnostic tests and medical imaging respectively, and FUTURE-AI promotes lifecycle-based health care AI [25–28].

Even with the broadening set of guidance, there continues to be a fragmentation of evidence across different specialties, technology types, and translational steps. Studies might rely on patient data without considering routine use; describe a physician interface without evaluating usage; or describe positive discrimination without evaluating calibration or workflow implications and patient outcome. Reviews incorporating applications to patients, administration, autonomous diagnosis, and decision support for physicians will blur the differences. Thus, an evidence map focused on Clinical validation needs to distinguish Clinical validation from Prospective/user evaluation and Implementation/impact evaluation, and include considerations of missing and unreliable information.

In reaction to that, a report-level, reliability-oriented evidence map of clinician-facing AI-CDSSs evaluated in the setting of real-world health care was conducted. A report was considered to be any publication-level document that had an independent bibliographic citation, regardless of the number of reports that referred to the same primary study. The key novelty of the project is that it provides an auditable distinction between Clinical validation, Prospective/user evaluation, and Implementation/impact evaluation of the technology. Additionally, the evidence map mapped out such aspects as clinical use cases, technical features, validation methods, integration signals, outcome, safety, equity, explainability,

economics, governance, and postimplementation monitoring of the technology, with explicit constraints on the inferential process due to search bias and extraction reliability.

2. Methods

2.1. Review Design and Reporting Framework

In accordance with the PRISMA-ScR reporting checklist and the updated methodological guidance from the Joanna Briggs Institute [29,30], this scoping review defined the scope, characteristics, stages of translation, and gaps for implementation of the clinician-oriented AI-CDSS literature (Supplementary File 1). The objective was to perform descriptive mapping of the evidence instead of conducting pooled effect size or intervention effectiveness evaluation.

2.2. Review Questions

The review addressed: (1) the clinical applications and decisions supported; (2) specialties, settings, populations, users, AI methods, and input modalities; (3) evaluation and validation approaches; (4) technical and workflow integration; and (5) adoption, usability, clinical impact, safety, fairness, explainability, economics, privacy, governance, and postimplementation monitoring.

2.3. Eligibility Criteria and Population–Concept–Context (PCC) Framework

Eligibility criteria were defined using the Population–Concept–Context framework. Population was defined as a set of clinicians, health care professionals, or multidisciplinary teams that had access to information produced by the AI, understood it, and was capable of acting on it for a clinical decision. Concept represented data-based techniques including machine learning, deep learning, natural language processing, large language models, computer vision, reinforcement learning, fuzzy systems, and Bayesian methods, which are used to create information useful for diagnosis, prognosis, risk assessment, treatment, prescribing, triaging, monitoring, alerting, referring, reporting, and workflow decisions. Context included actual health care data or workflow data. The clinical decision-making process or potential integration into workflow were not enough for inclusion.

Studies based on primary empirical reports were only considered to be eligible if the full text exhibited an element of AI, real-life human health care data, and a concrete clinician-facing component that could be useful for patient care decision-making. Studies on model development and validation were not automatically considered to be eligible based on the fact that they used clinical data, had external or temporal validation, compared AI models with the clinicians, and their future possible integration. Tools or outputs with a concrete layer that had not been tested in real-life users or workflows were considered to be Clinical validation. The reports were designated as Prospective/user evaluation if clinicians used or evaluated the AI in a prospective, reader, simulated-workflow, and user-task setting without evaluation of the effect of its usage in live-care; Implementation/impact evaluation if live workflow usage, integration, adoption, operational outcome, patient outcome, or implementation impact were being studied. F5 was the exclusion criterion if a model development/validation lacked a concrete clinician-facing layer.

2.4. Information Sources and Search Strategy

Source corpus extraction was conducted in PubMed/MEDLINE and was seen as a design boundary as opposed to an affirmation of comprehensive multidisciplinary coverage. A query was conducted in the database on 15 June 2026 based on identical concept syntax per year, beginning from 1 January 2020 to the search date without any language filter used. The start date of 1 January 2020 was kept as a practical boundary for our contemporary times and was reflected retrospectively in the protocol amendment record. Dates, year limits, and output are available in Supplementary File 2. Other data characteristics are available as Tables S1–S5 in Supplementary File 3 and data is given in Supplementary File 4. The annual PubMed/MEDLINE exports contained 14,111 raw records in total. There were no surveillance post-cutoff records included.

2.5. Record Management and Deduplication

Records were managed reproducibly in Python (Supplementary File 5). Duplicate removal occurred in two documented stages: 14 within-year exact-title duplicates and 14 cross-year duplicates after annual concatenation, yielding 14,083 unique screened records. PMID, normalized DOI, and exact normalized title were used; fuzzy matching was not used.

2.6. Title and Abstract Screening

Title/abstract prescreening involved the use of transparent, rule-based approaches in combination with AI-generated recommendations, followed by an assessment by the authors (Supplementary File 6). The rules were designed to prioritize sensitivity by retaining records when the title or abstract did not provide sufficient grounds for exclusion. This design objective does not establish corpus-wide screening sensitivity. The AI recommendations were only for reference purposes and did not define eligibility criteria. The entire corpus was not screened by two humans independently. The prespecified TA200 and FT150 quality control samples were assessed independently by Ali Olamat.

2.7. Full-Text Retrieval

After applying the filters of title/abstract and metadata screens, there were a total of 5140 reports sought for full-text retrieval (Supplementary File 7). For those for which PMC XML was available, it was used. The others not in PMC were searched for through legal means of open access, the publishers' websites, library databases, or PDF files stored locally. A systematic effort at retrieving reports that could not be found through an interlibrary loan or direct contact with authors was not done for all the reports found to be unavailable. The retrieval summary recorded 1373 of 1449 outside-PMC reports (94.8%) as not retrieved and all 3691 PMC records as assessed. A candidate-level metadata comparison between retrieved and non-retrieved reports was not available.

2.8. PMC Identifier-Integrity Repair and XML Verification

An identifier-integrity audit found that a prior PubMed-to-PMC mapping process could assign a PMCID that was not an article-level identifier. The mapping was rebuilt using only PubMedData/ArticleIdList identifiers, removing 207 incorrect PMCIDs and yielding 3691 verified PMC candidates plus 1449 outside-PMC candidates. Each acquired PMC XML article was verified against its PMID, DOI, and PMCID.

2.9. Full-Text Eligibility Assessment

The eligibility-assessment set comprised 3767 reports, recorded as 3691 PMC XML articles and 76 outside-PMC reports. The same eligibility rules were applied to both sources. A total of 1181 reports were included and 2586 were not included. Included reports were classified as Clinical validation, Prospective/user evaluation, or Implementation/impact evaluation. The 1373 records assigned to the not-retrieved category were recorded separately. Recorded decisions and supporting documentation are provided in Supplementary File 8; Table 1 clarifies the distinction between non-inclusion and a definitive Exclude decision.

Table 1. Study selection and final report-level corpus.

Characteristic	Count	Percent (%)
Raw annual exports	14,111	-
Within-year duplicates removed	14	0.1
Canonical annual records	14,097	-
Cross-year duplicates removed	14	0.1
Unique records screened	14,083	100.0
Title/abstract exclusions	8504	60.4
Metadata-stage exclusions	439	7.9 of 5579
Reports sought	5140	-
Reports not retrieved	1373	26.7
Reports considered for eligibility	3767	-
Reports not included	2586	68.6
Included reports	1181	31.4
PMC	1137	96.3 of included
Outside PMC	44	3.7 of included
Clinical validation	741	62.7

Table 1. *Cont.*

Characteristic	Count	Percent (%)
Prospective/user evaluation	270	22.9
Implementation/impact evaluation	170	14.4

The two duplicate-removal stages are distinct. The 2586 non-included reports include four Unclear/F10 records rather than definitive Exclude decisions. For these four records and one F5 record, the available documentation does not fully establish full-text retrieval and/or the decision basis. Code definitions and record-level details are provided in Supplementary Table S5 (Supplementary File 3) and Supplementary File 8.

2.10. AI-Assisted Structured Author Adjudication

The process of full-text processing involved the use of structured extraction, artificial intelligence (AI)-assisted coding assistance, targeted auditing, and human adjudication. The machine-produced results focused on the identification of priority for records or suggestions for code assignment but not in the final decision on eligibility or charting. Targeted human review covered 666 risk-flagged records. Documented remediation activities included consensus resolution of the 152-record disagreement workbook, reconciliation of 24 FT150 disagreements, and targeted integrity re-verification of 90 provisional new inclusions.

2.11. Data Charting

A master charting dataset from internal sources that contains 117 variables was used to aid the extraction and validation process. The released public dataset contains 66 variables at the level of reports out of a total of 1181 reports used in this study. Supplementary Data S1 (Supplementary File 4) presents a provenance matrix that captures, for each public variable, the derivation pathway, source variable, basis for verification, scope of dual review, reliability tier, denominator source, and maximum allowable interpretation. This matrix clearly shows that a corpus-wide dual review process was not carried out; DX120 and targeted correction provide bounded validation but not source verification. Study_Design analysis was carried out independent of the final Evaluation_Stage taxonomy and there are 170 Study_Design entries missing.

2.12. Independent Validation and Adjudication

With a predefined seed of 20260730, Ali Olamat individually scored 200 title/abstracts and 150 full texts using blinded pre-consensus workbooks. These are comparator-specific quality control modules rather than a complete dual-human review of the entire corpus. For DX120, Pinar Özel and Ali Olamat independently completed the core fields in 120 publications; Ali Olamat served as the independent second reviewer. The public package includes deidentified pairs and reproducible metrics; named raw reviewer workbooks and signed author approvals are retained in the restricted/source-lock archive.

Percent exact agreement and Cohen's κ , Gwet's AC1, class-specific sensitivity, specificity, precision, negative predictive value, and F1 score were calculated. Confidence intervals of 95% were computed using 2000 bootstrap samples based on the seed 20260809. TA200, FT150, and DX120 interval sampling was achieved by bootstrapping publication records; for DX120, the pooled intervals are obtained through cluster bootstrap by Final_DX_Slot to ensure all the fields from the chosen record are included. TA200, FT150, and DX120 were used as fixed quality control modules rather than sample sizes computed via statistical power analysis. Metrics relative to the final FT150 reference are post-adjudication concordance estimates because the reference contains consensus resolution.

2.13. Descriptive Synthesis

Descriptive analyses were performed based on the report level ($N = 1181$). The Evaluation_Stage taxonomy used in the final analysis includes Clinical validation, Prospective/user evaluation, and Implementation/impact evaluation. Study_Design was evaluated separately as a multi-coded, source-derived category. Minimal identification/title-based linking revealed 1181 reports in relation to 1180 potential report families. It should be noted that linking was not meant as an attempt to reconstruct the report families in a systematic way and is presented only as a lower-bound duplicate report sensitivity estimate.

For missing-data and absence-detection handling, NR denoted information not reported or not captured and was not interpreted as No. Analysis-ready three-state fields distinguished Detected, Not detected in available extraction, and NR or uncaptured. A not-detected signal was treated as an extraction result rather than proof of absence in the original report. Where a positive source field conflicted with a rule-derived absence flag, the positive field took precedence.

2.14. Risk-of-Bias Assessment

Consistent with the evidence-mapping purpose, no formal report-level risk-of-bias or critical-appraisal tool was applied. Evidence maturity was described through evaluation stage, validation type, prospective or user evaluation, workflow integration, patient or clinical outcomes, calibration, decision-curve analysis, safety, fairness, explainability, economics, and governance or postimplementation evaluation.

2.15. Protocol or Registration

No prospective registration occurred for the review. A retrospective protocol, dated 16 August 2026, describes the clinical eligibility criteria, quality control modules, the replacement protocol, the phases of consensus and remediation, the duplicate concordance, the corrections made to the publicly available dataset, the search cutoff date of 15 June 2026, and the resulting corpus of $N = 1181$ records. A dated explanation, provided on 18 August 2026, provides the justification for the start date of 2020 and the final bootstrap-uncertainty methodology; no changes were made to either eligibility or PRISMA counts.

3. Results

3.1. Study Selection

There were 14,111 raw records across the annual PubMed/MEDLINE exports. In total, 14 within-year duplicates were eliminated before annual canonicalization, which resulted in 14,097 records; moreover, after concatenation, 14 cross-year duplicates were discarded, leaving 14,083 records for screening. A total of 8504 records were excluded during title/abstract screening, and 439 records were excluded during metadata screening. Of 5140 reports sought, 1373 were recorded as not retrieved and 3767 were considered for eligibility. The retrieval summary recorded 1373/1449 outside-PMC candidates (94.8%) as not retrieved and all 3691 verified PMC candidates as assessed. A total of 2586 reports were not included and 1181 reports were included. A summary of validation point estimates is provided in Supplementary File 9 and the complete included-report bibliography is provided in Supplementary File 10. Recorded F1–F10 counts were F1 = 56, F2 = 729, F3 = 194, F4 = 151, F5 = 1120, F6 = 42, F7 = 12, F8 = 11, F9 = 267, and F10 = 4. The final corpus comprised 1137 PMC and 44 outside-PMC reports (Table 1; Figure 1).

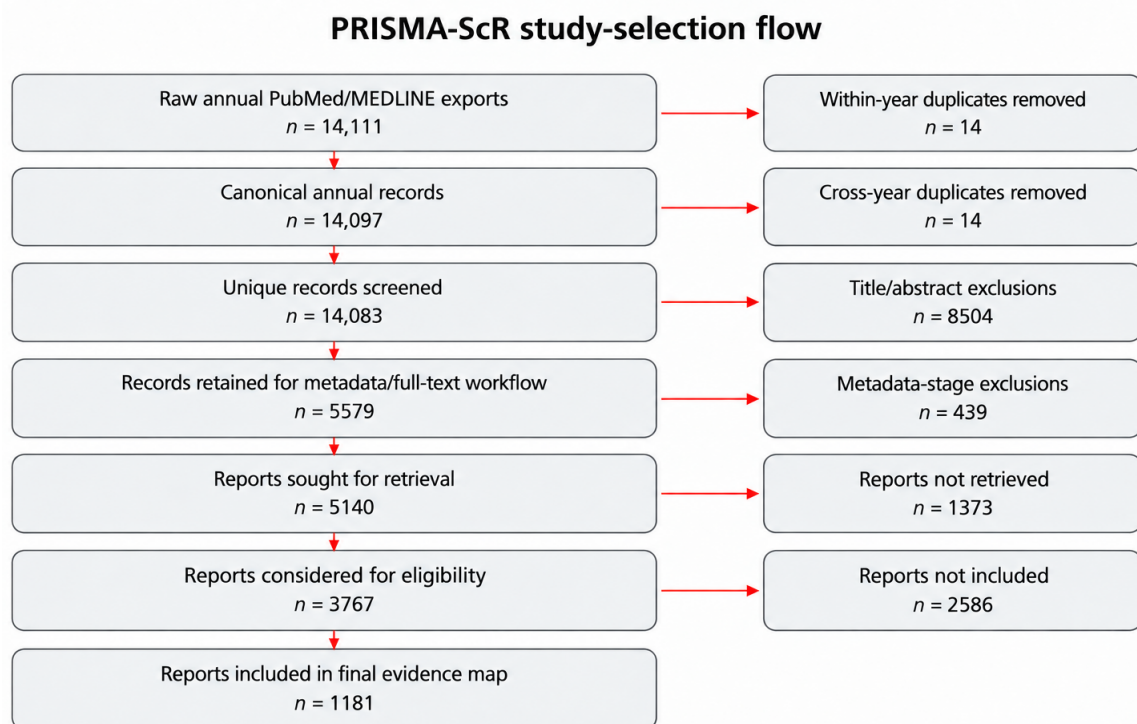


Figure 1. PRISMA-ScR study-selection flow, including within-year and cross-year duplicate removal. Non-inclusion is defined in the Table 1 note.

No formal risk-of-bias assessment was performed. Descriptive maturity indicators are reported in the main tables and Supplementary Data S1. The public source-level charting dataset contains all 1181 final included publication reports and 66 variables covering bibliographic data, final class, clinical scope, AI approach, input modality, application, evaluation

and validation features, integration variables, outcome domains, implementation-gap indicators, and analysis-ready three-state fields.

3.2. Publication Trends and Geographical Distribution

Publication volume rose markedly after 2024, with 654 of 1181 included reports published in 2025–2026. Country or region terms were captured for 1057 reports. China, the United States, and the United Kingdom were most frequently detected; 249 reports were multicountry and 124 were NR or uncaptured (Figure 2; Supplementary Figure S2 (Supplementary File 7); Supplementary Tables S1 and S2).

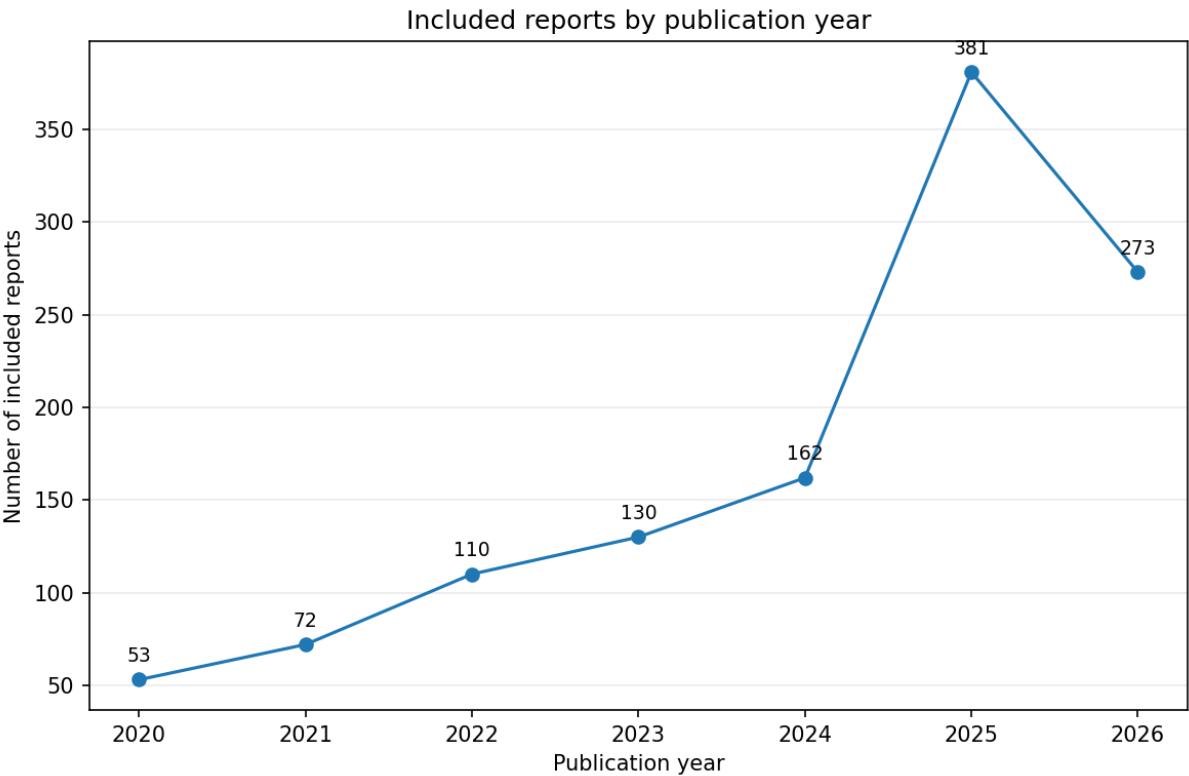


Figure 2. Included reports by publication year; the 2026 search ended on 15 June 2026.

3.3. Clinical Specialties, Settings, Populations, and User-Related Lexical Signals

The specialties that were coded as the most frequent were oncology and surgery/perioperative care, followed by endocrinology, pulmonology, nephrology, critical care and cardiology. Nephrology was coded in 207 reports (17.5%). Surgery/perioperative specialty was coded in 350 reports (29.6%), whereas an operating-room/perioperative setting was captured in 66 reports (5.6%). The healthcare setting was not mentioned or captured in 702 publications; out of those mentioned, the most common settings were intensive care units and emergency rooms. Lexemes that were associated the most with users were clinicians, physicians, specialists, care team, nurses, and radiologists. The term clinical decision support was kept separate from the professional role as a use-layer signal. Multi-coding was common in signals and many were based on rule-derived coding rather than manually assigned intended user information (Supplementary Table S3).

At least one user-related lexical signal was captured in 1162/1181 (98.4%) reports. The most frequent signals were clinician (962/1181 [81.5%]), physician (695/1181 [58.8%]), specialist (305/1181 [25.8%]), care team (244/1181 [20.7%]), nurse (163/1181 [13.8%]), and radiologist (149/1181 [12.6%]). A separate clinical decision-support use-layer signal was detected in 660/1181 (55.9%) reports. These nonexclusive text-detection counts should not be interpreted as a validated taxonomy of intended users without source-level confirmation (Supplementary Table S3).

3.4. AI Methods and Input Data Modalities

Machine learning algorithms that can be used in generic cases including gradient boosting, random forest, deep learning, support vector machines, and large language models or generative AI were the most common types of methods. More than one type of method was found in 733 studies. Structured EHR/EMR and medical imaging were the top two types

of inputs, while there were 509 papers where no input type was identified. Multiple types of input were coded in 256 studies (Figure 3, Table 2, Supplementary Figure S5).

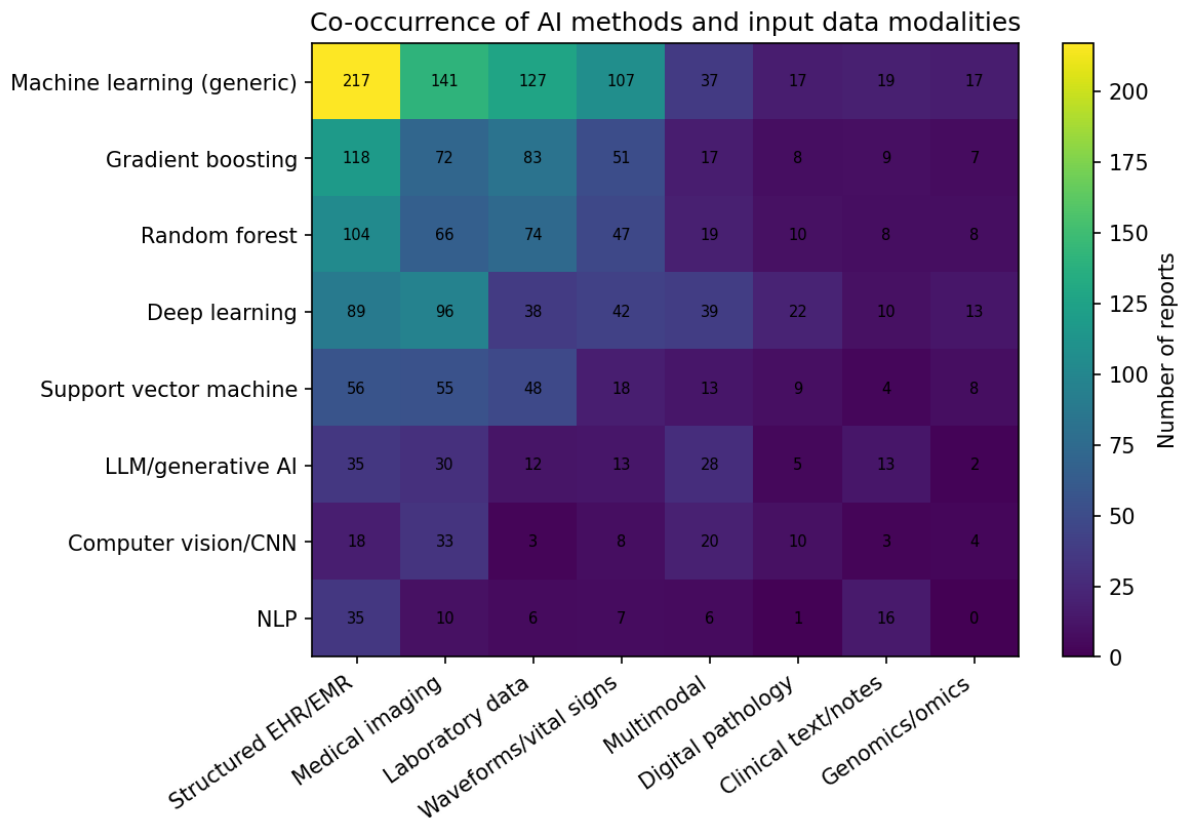


Figure 3. Co-occurrence of AI methods and input data modalities.

Table 2. Selected clinical, technical, and integration characteristics.

Domain	Category	Count	Percent (%)
Clinical application	Diagnosis/detection	774	65.5
Clinical application	Prognosis/risk prediction	588	49.8
Clinical application	Screening/referral	235	19.9
Clinical specialty	Oncology	355	30.1
Clinical specialty	Surgery/perioperative	350	29.6
AI method	Machine learning (generic)	796	67.4
AI method	Gradient boosting	415	35.1
Input modality	Structured EHR/EMR	301	25.5
Input modality	Medical imaging	236	20.0
Workflow signal	Broad clinician-facing workflow/use layer	933	79.0
Technical integration	Specific technical-integration signal (exploratory; numeric output withheld)	—	—

Categories may be multi-coded. Specific technical-integration status is exploratory.

3.5. Clinical Applications and Decisions Supported

Diagnosis/Detection and Prognosis/Risk Prediction took primacy in this area, followed by Screening/Referral, Monitoring, Triage/Prioritization, and Treatment Recommendation/Planning. “Diagnosis/Detection” was encoded in 774 studies and “Prognosis/Risk Prediction” in 588; lack of application information was found in 86. The output categories “Risk/Probability” and “Diagnosis/Classification” were the most common ones. Since applications were multi-coded and mostly rule-derived, they should be treated as evidence map markers but not mutually exclusive functions (Table 2; Figure 4; Supplementary Figure S4).

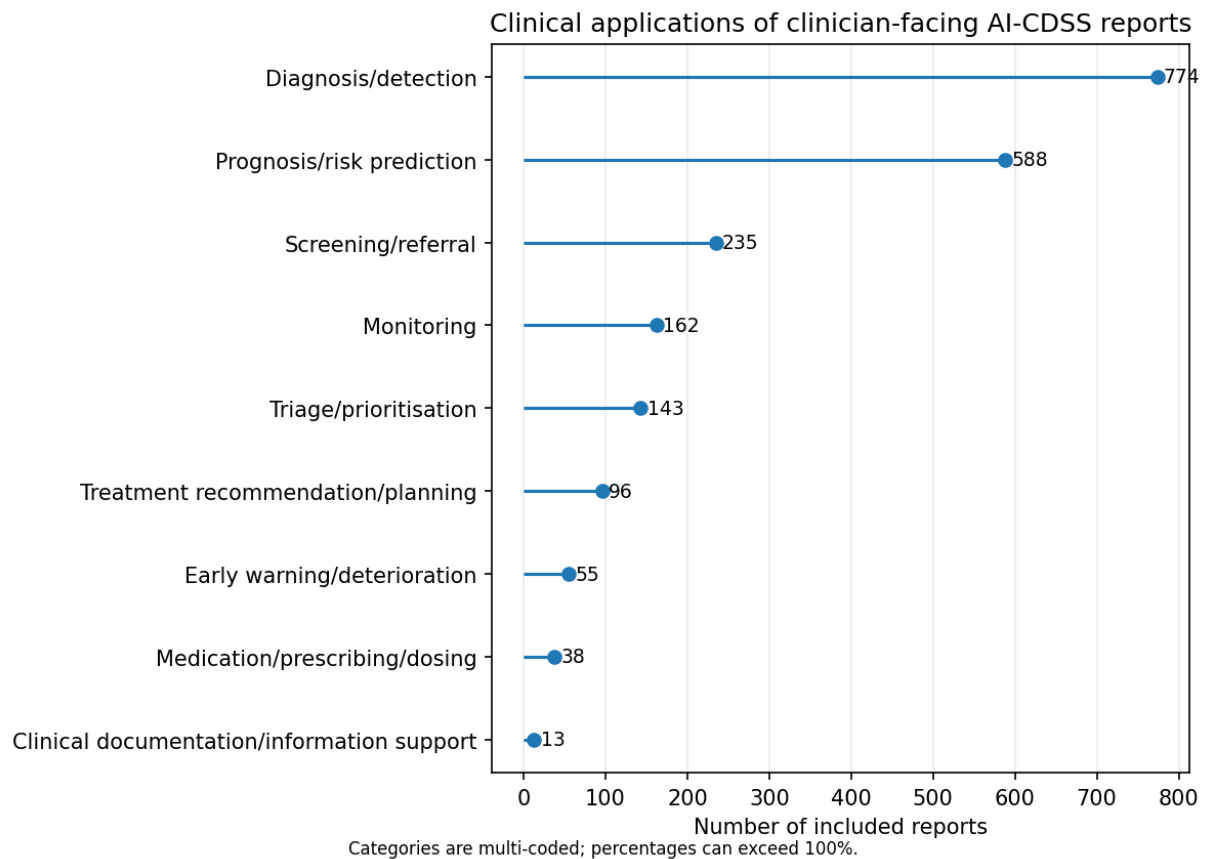


Figure 4. Clinical applications of included clinician-facing AI-CDSS reports.

3.6. Evaluation Approaches and Validation Strategies

Final translational stages were Clinical validation (741/1181 [62.7%]), Prospective/user evaluation (270/1181 [22.9%]), and Implementation/impact evaluation (170/1181 [14.4%]). Study-design information was populated in 1011/1181 (85.6%) reports. The most frequent multi-coded design categories were retrospective internal validation (704), external validation (433), prospective clinical evaluation (363), multicenter validation (259), qualitative/mixed methods (116), and usability/human-factors evaluation (109); 170 reports were NR or uncaptured (Table 3; Figure 5).

Table 3. Final translational stages and corrected study-design categories.

Domain	Category	Count	Percent (%)
Final stage	Clinical validation	741	62.7
Final stage	Prospective/user evaluation	270	22.9
Final stage	Implementation/impact evaluation	170	14.4
Study design	Retrospective internal validation	704	59.6
Study design	External validation	433	36.7
Study design	Prospective clinical evaluation	363	30.7
Study design	Multicenter validation	259	21.9
Study design	Qualitative/mixed methods	116	9.8
Study design	Usability/human factors	109	9.2
Study design	NR or uncaptured	170	14.4

Study_Design is source-derived and multi-coded; Evaluation_Stage is the final human-locked taxonomy.

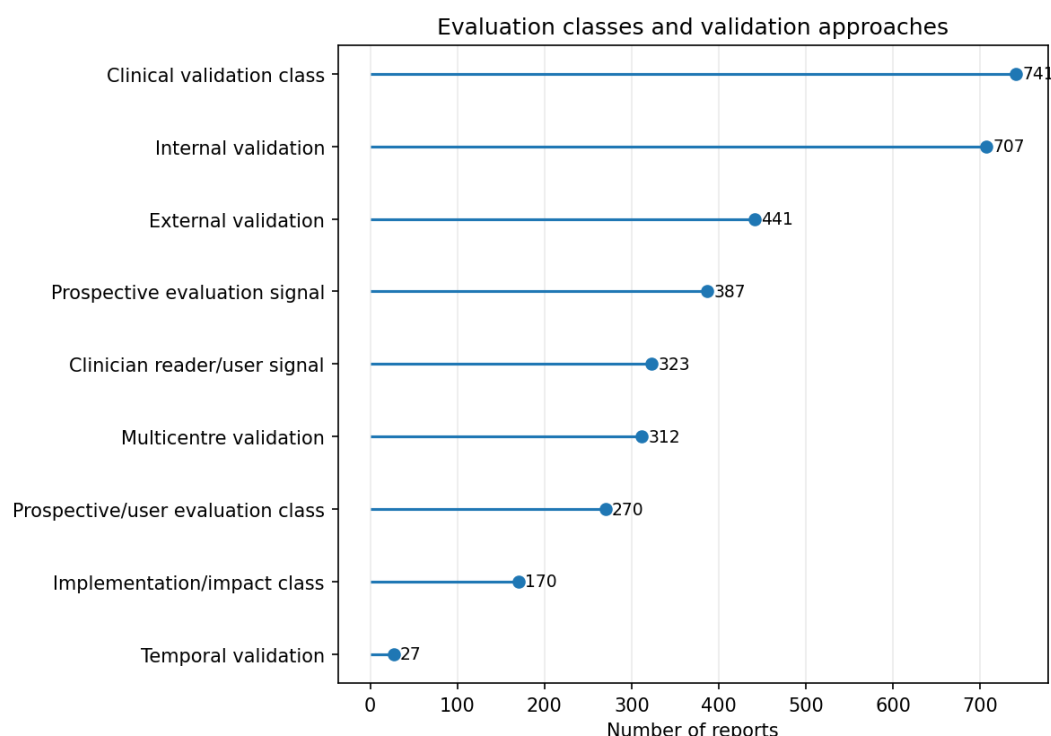


Figure 5. Evaluation classes and validation approaches.

The internal validation was described in 707 papers, while external validation was mentioned in 441 papers, multicenter validation in 312 papers, and temporal validation in only 27 papers. The prospective assessment and reader/user evaluation were stated in 387 and 323 papers, respectively. The most often mentioned performance domains were discrimination and sensitivity/specificity. Calibration and decision-curve analysis were found in 367 and 277 papers, correspondingly.

3.7. Workflow and Technical Integration

A workflow signal was created and characterized for clinicians. As the specific domain for technical integration only showed moderate reliability, it is not included in the main body of the report and is available exclusively in the Supplementary Data S1 and Supplementary File 9.

3.8. Adoption, Usability, and Implementation Evaluation

Adoption/use outcomes were detected in 422 reports and usability/human-factors outcomes in 184. The combined workflow/adoption/usability status was positive in 506, not detected in 584, and NR or uncaptured in 91 reports (Table 4).

Table 4. Outcome and responsible-AI extraction signals with reliability labels; low-reliability numeric outputs are withheld from main-text interpretation.

Domain	Detected	Not Detected	NR/Uncaptured	Reliability
Clinical/patient outcomes	—	—	—	Exploratory; numeric output withheld (Supplementary File 9)
Workflow/adoption/usability	506	584	91	Descriptive
Safety	98	1083	0	Descriptive
Fairness/equity	61	996	124	Descriptive
Explainability	155	405	621	Descriptive
Economic/resource	182	886	113	Descriptive
Privacy/governance/post-deployment	—	—	—	Exploratory; numeric output withheld (Supplementary File 9)
Technical integration	—	—	—	Exploratory; numeric output withheld (Supplementary File 9)

Not detected is an extraction state, not proof of article-level absence. Em dashes indicate numeric outputs withheld from main-text interpretation; complete exploratory values are available in Supplementary Data S1 and Supplementary File 9.

3.9. Clinical and Patient Outcomes

Extraction of clinical/patient-outcome information showed very low inter-reviewer reliability in this review. Therefore, quantitative results for this domain are not interpreted in the main text and are provided only in Supplementary Data S1 and Supplementary File 9 as exploratory extraction states.

3.10. Safety, Fairness, Explainability, Economics, and Governance

Overall, safety occurred in 98 reports, fairness/equity in 61, explainability/interpretability in 155, and economic/resource evaluation in 182. Extraction of privacy/governance/post-deployment information showed very low inter-reviewer reliability, and its numerical findings are thus omitted from main-text interpretation. Not-detected extraction status does not necessarily mean absence from an article (Table 4).

3.11. Implementation Gaps

Numerical results for low/moderate reliability concerning clinical outcomes, privacy/monitoring, and technical integration are not included in the main manuscript. Exploratory workflow/technical-integration and implementation-gap extraction results are provided in Supplementary File 9, Exploratory Figures E1 and E2, for audit transparency only; they are not prevalence estimates.

3.12. Independent Validation and Adjudication Results

Table 5 summarizes the TA200, FT150 and DX120 validation results with bootstrap 95% confidence intervals. In TA200, original automation achieved 42.5% agreement (95% CI 35.5–49.5), $\kappa = 0.133$ (95% CI 0.066–0.204), and Include F1 = 0.428 (95% CI 0.337–0.513); calibrated automation achieved 46.0% agreement (95% CI 39.0–53.0), $\kappa = 0.183$ (95% CI 0.100–0.266), and Include F1 = 0.476 (95% CI 0.353–0.583). In FT150, the primary reference comparison yielded 59.3% agreement (95% CI 52.0–67.3), $\kappa = 0.212$ (95% CI 0.068–0.356), and Include F1 = 0.558 (95% CI 0.450–0.657); the final adjudicated reference comparison yielded 76.0% agreement (95% CI 68.7–82.7), $\kappa = 0.477$ (95% CI 0.346–0.604), and Include F1 = 0.651 (95% CI 0.524–0.758). The latter is post-adjudication concordance, not wholly independent validation. In the same TA200 sample, include sensitivity was 88.9% (40/45) for original automation and 55.6% (25/45) for calibrated automation against the independent reviewer reference. Thus, the increase in F1 was accompanied by lower sensitivity in this comparator-specific sample; these estimates do not measure the sensitivity of the complete human-assisted review workflow (Supplementary File 5).

In the replacement-adjusted DX120 sample, 1411 field comparisons were evaluable. Agreement was 79.4% (cluster-bootstrap 95% CI 77.2–81.7), $\kappa = 0.579$ (95% CI 0.534–0.624), AC1 = 0.602 (95% CI 0.557–0.645), precision = 0.843, recall = 0.675, specificity = 0.894, NPV = 0.766, and F1 = 0.750 (95% CI 0.721–0.782). Strict three-level stage agreement was 96.7% (95% CI 93.3–99.2), $\kappa = 0.916$ (95% CI 0.818–0.981), and macro-F1 = 0.949 (95% CI 0.889–0.990). Full class-specific metrics and confidence intervals are reported in Supplementary File 9 and the authoritative validation workbook.

Table 5. Reproducible validation results with bootstrap 95% confidence intervals.

Module	Comparison	<i>n</i>	Agreement % (95% CI)	κ (95% CI)	F1/Macro-F1 (95% CI)
TA200	Original automation vs Ali Olamat	200	42.5 (35.5–49.5)	0.133 (0.066–0.204)	0.428 (0.337–0.513)
TA200	Calibrated automation vs Ali Olamat	200	46.0 (39.0–53.0)	0.183 (0.100–0.266)	0.476 (0.353–0.583)
FT150	Primary reference vs Ali Olamat	150	59.3 (52.0–67.3)	0.212 (0.068–0.356)	0.558 (0.450–0.657)
FT150	Final adjudicated reference vs Ali Olamat	150	76.0 (68.7–82.7)	0.477 (0.346–0.604)	0.651 (0.524–0.758)
DX120 pooled	13-field human–human comparison	1411	79.4 (77.2–81.7)	0.579 (0.534–0.624)	0.750 (0.721–0.782)
DX120 stage	Strict three-level crosswalk	120	96.7 (93.3–99.2)	0.916 (0.818–0.981)	0.949 (0.889–0.990)

TA200 and FT150 are examples of quality control analysis for specific comparators; the last FT150 comparison is a post-adjudication concordance comparison. DX120 is the independent human–human extraction comparison. All of the complete diagnostic measures, including AC1, McNemar test, class-specific stage F1, and confidence intervals, are presented in Supplementary File 9 and the validation workbook.

In total, the source locked-remediation involved 666 risk flagged files. Prior to the application of New90 integrity overlay, the distribution included 105 Inclusion, 557 Exclusion and four Unclear/F10 files. For re-verification, 90 targeted files were included, 37 of which were still Inclusion and 53 became Exclusion and, therefore, the final distribution was 52 Inclusion, 610 Exclusion and four Unclear/F10. Overall, 3767 reports were considered for eligibility: 1181 were included and 2586 were not included.

4. Discussion

4.1. Principal Findings

The resulting evidence map, which is based on PubMed/MEDLINE data and strongly skewed toward PMC, is made up of 1181 accessible reports on AI-CDSS for clinicians. The taxonomy that was constructed manually and signed off by humans was primarily focused on Clinical validation, while Prospective/user and Implementation/impact evaluation were performed less often. The system could not replicate the clinician-facing threshold with the necessary accuracy and thus replace human moderation, and structured data extraction reliability was dependent on the topic. The main value of this work is establishing translational stages in an auditable evidence map with clear boundaries.

4.2. Comparison with the Previous Literature

These results agree with previous review studies that highlight the existence of abundant technical evidence but little evidence regarding clinical use [3–6]. The inclusion of the SWOT analysis from Sallam et al. makes it clear that the identified opportunities in terms of diagnosis and efficiency are threatened by other issues, namely, bias, privacy, transparency, and oversight [9].

This body of work illustrates why real-world data are not synonymous with real clinical use. Retrospective EHR, registry, imaging, laboratory, or waveform data may support clinically relevant validation but do not show that clinicians used the tool or that care changed. The strict eligibility rule therefore required a concrete clinician-facing layer and preserved distinctions among Clinical validation, Prospective/user evaluation, and Implementation/impact evaluation. The final 666-record remediation and 90-record integrity re-verification materially changed the corpus, highlighting the importance of explicit translational criteria and signed human approval.

4.3. From Validation Performance to Clinical Impact

Extraction of clinical/patient-outcome information showed low inter-reviewer reliability in this review; therefore, these extraction states were not used for corpus-wide quantitative generalization. This is a limitation of the review's extraction process, not a judgment on the reliability of clinical-outcome research as a field. Future AI-CDSS studies should relate model outputs to clinician actions, process changes, and patient outcomes.

4.4. Validation, Calibration, and Clinical Utility

Internal validation was common, while external validation, multicenter validation, and especially temporal validation were rare. Transportability might be affected by factors such as prevalence, measurement, documentation, practice, and structural data differences. External validation does not guarantee long-term stability, and local validation, temporal validation, recalibration, and updating are still necessary [15–19].

Fewer studies included calibration and decision-curve analysis than discrimination. TRIPOD+AI and PROBAST+AI also focus on calibration, generalizability, quality of reporting, and bias in addition to discrimination [23,24]. Since clinical tools usually prompt action above some threshold value, future studies should consider prespecifying threshold values, calibration in the large and slope where relevant, net benefit analysis, and local recalibration or updating.

4.5. Workflow Integration and Clinician–AI Interaction

More general characterizations of entire workflows or layers of use were common. Indicators of technical integration were more problematic, and thus the frequency distribution of these is not discussed in the body of the paper. The impact of workflow depends on task design, interface, timing, training, organizational setting, and accountability systems. Thus, integration must be thought of as a sociotechnical intervention, in terms of who gets the product, when they get it, how they get it, what training comes along with it, how disputes are settled, and what follows.

Explainability or interpretability was detected in 155/1181 (13.1%) reports, while 621/1181 (52.6%) were NR or uncaptured. A missing positive extraction signal does not prove that no explanation was provided; nevertheless, external analyses indicate that clinician-centered evaluation of explanations remains uncommon [8]. Future studies should test whether explanations improve appropriate reliance, error detection, confidence calibration, or decision quality rather than assume that feature importance or saliency is inherently useful.

4.6. *Clinical and Patient Outcomes*

The extraction results did not provide a reliable corpus-based estimate of prevalence for any of the clinical or patient outcomes. Even if death or disease occurrence is not necessary as the outcome measure in all AI-CDSSs, there needs to be a description of how the output is converted to decisions and processes and ultimately to patient outcomes in each study.

4.7. *Responsible AI, Safety, Fairness, and Explainability*

Signals regarding safety and equity/fairness were reported in relatively few papers; nonetheless, the lack of positive extraction of safety signals in the articles considered does not mean their overall lack. The assessment of safety should go beyond average error of models and include false positive and false negative rates, automation bias, alert fatigue, subgroup performance, failure modes of workflows, and possible misuse.

The assessment of fairness should move beyond measures of parity. Subgroup differences may occur due to representation, data quality, measurement, prevalence, access, or care pathways; conversely, equality on one measure does not necessarily mean equal benefit. Fairness should be assessed through representation, subgroup calibration, threshold impact, missingness, access, and change in disparities due to deployment [13,28].

4.8. *Economic Evidence, Governance, and Post-Deployment Monitoring*

Economic or resource evaluation was detected in 182/1181 reports and showed better field reliability than the exploratory outcome and governance variables. Economic analyses should address acquisition, integration, local validation, training, workflow redesign, maintenance, monitoring, cybersecurity, and the cost of incorrect decisions.

Extraction of privacy/governance/post-deployment information showed very low inter-reviewer reliability. Positive, not-detected, and NR/uncaptured states in this domain should therefore be treated only as exploratory extraction results, not as evidence establishing the presence or absence of governance in the original reports.

4.9. *Regulatory and Policy Implications*

It seems clear that there is a link between the gaps identified and current regulatory expectations. In addition to the focus on ethical issues, the WHO highlights the importance of human rights, governance, evaluation and monitoring, describing their guidelines as non-binding [13,14]. FDA issued final guidance on the use of AI for device software functions in PCCP in August 2025 and CDS software in January 2026 with continued support for risk-based monitoring through lifecycle and real-world performance [15–17]. The international guidelines set out by the IMDRF N41 and N88 establish principles for the clinical evaluation of SaMD and good machine-learning practices [18,19]. The regulation (EU) 2026/1744 “Digital Omnibus on AI” is currently in effect and has updated the framework and timelines for the implementation of the AI Act [20].

Reporting standards can help align research with these requirements. CONSORT-AI and SPIRIT-AI can be used for the transparent reporting of trials and protocols for AI interventions [21,22]; TRIPOD+AI and PROBAST+AI can be applied to the reporting and critical evaluation of predictive models [23,24]; STARD-AI can be used for diagnostic accuracy reporting [25]; CLAIM provides imaging-specific guidance, including its 2024 revision [26,27]; and FUTURE-AI provides a comprehensive lifecycle framework [28].

4.10. *Strengths and Limitations*

Methodological strengths include a comprehensive synthesis of 1181 reports, specification of Clinical validation, Prospective/user evaluation, and Implementation/impact evaluation; a 66-item charting form and a new provenance matrix at the field level; validation of all 3691 PMC identifiers; recording of independent, consensus, and final audit values; prespecified TA200, FT150, and DX120 quality-control modules; remediation of human errors on 666 risk-flagged records; joint signatures on the final eligibility lock and 52 new charted reports; bootstrapped uncertainty that is reproducible; and minimal duplicate-report sensitivity check.

A number of limitations have far-reaching implications. First, the search covered only PubMed/MEDLINE and publications starting from 2020; 2020 was retrospectively noted as a pragmatic design decision. Second, the retrieval summary recorded 1373 of 1449 outside-PMC candidates (94.8%) as not retrieved and all 3691 PMC candidates as assessed. Systematic interlibrary-loan retrieval or author contact was not implemented across all unavailable reports, and no candidate-level retrieved-versus-not-retrieved comparison was available; hence the direction and magnitude of availability bias cannot be quantified. Third, the corpus was not fully independently double-screened and double-extracted. TA200 and FT150 are quality-control modules specific to each comparator; the eligibility criteria and the protocol were revised retrospectively, and there was no risk-of-bias or critical appraisal tool formally used. Fourth, Study Design was retrospectively mapped back, and 170 reports were not reported/captured. Fifth, the reliability of clinical/patient outcomes,

privacy/governance/post-deployment monitoring, and technical integration was low to moderate; numeric results for them are not reported here. Sixth, the unit of analysis is the publication report, and $N = 1180$ is just a minimal duplicate-report check on the identifier/title level. Documentation limitations affecting five non-included reports are described in Table 1 and Supplementary File 8. TA200 results are comparator-specific and do not establish corpus-wide screening sensitivity.

4.11. Implications

The assessment of the performance of AI-based CDSS should include calibration, integration into workflow, training of users, safety, and monitoring along with other considerations. Health care organizations need to identify the decision to be supported, the user group, the action threshold, the escalation procedure, and situations under which an output could be overridden.

Validation and implementation by developers are interlinked activities throughout the product lifecycle. Documentation needs to cover data sources, users, error modes, choice of action thresholds, human–AI interaction, implementation context, updating practices, and monitoring plan. Responsibility needs to be defined within health care organizations regarding governance, auditing, adverse event reporting, equity, security, maintenance, and, eventually, retiring of the AI-based system.

Regulatory and policy makers must ensure representativeness of data, human oversight, transparency, post-market surveillance, and controlled model modification. Research must move past accuracy-focused work and examine the impact of AI on clinicians' practice in terms of when, how, who and at what cost the changes occur.

4.12. Research Priorities

Future research must concentrate on temporal and external validation, local calibration, prospective silent validation, multicenter implementations, and direct measures of adoption, usability, trust, workload, and clinical outcomes. Prospective safety and equity testing should include subgroup and failure modes testing wherever possible. Economic evaluation needs to start early to gather information about implementation costs and adoption trajectories. Postimplementation research must set out in advance thresholds for action, drift monitoring, need for retraining, governance requirements, and reactions to unanticipated harms. In low-resource settings, local capacity building of data, infrastructure, personnel, and governance needs to occur rather than assuming that a model will transfer without modification [12,28].

5. Conclusions

Among the 1181 PubMed/MEDLINE-indexed papers available for review, the human-reviewed taxonomy showed that evidence of AI-CDSS for clinicians was mostly limited to Clinical validation, with relatively lesser number of papers on prospective application and implementation/impact. Temporal validation was rarely found. Findings related to technical integration, clinical outcomes, and privacy/governance/monitoring after deployment of the system were not considered as part of the main paper because of poor field extraction reliability at a low to moderate level.

Future research should treat clinical utility, implementation, human factors, equity, safety, local calibration, economic value, and post-deployment monitoring as core evidence domains rather than optional additions after technical development. Auditable human oversight, clear action pathways, lifecycle reporting standards, and pragmatic evaluation are needed to demonstrate accountable clinical value rather than technical functionality alone.

Supplementary Materials: The following supporting information accompanies this article in <https://tijfs.jams.pub/conversion/supplementary/1ce41410041b4b91e8f21e7a04f3ba59>: Supplementary File 1. PRISMA-ScR Checklist. Reporting checklist for the final corpus of 1181 reports. Supplementary File 2. Complete PubMed/MEDLINE Search Strategy and Record Reconciliation. Full search syntax, annual limits, and record reconciliation to the final corpus. Supplementary File 3. Supplementary Tables S1–S5. Publication-year/source, country/region, clinical-scope, retrieval, and non-inclusion summaries for 1181 reports. Supplementary File 4. Public Source-Level Charting Dataset (Supplementary Data S1). The 1181-report, 66-variable public dataset, with operational codebook, field provenance, missingness, field-reliability labels, correction audits, attestation-evidence metadata, and the 52-row merge audit. Codebook and missingness are worksheets within this file. Supplementary File 5. Deidentified Reproducibility Runner. Deidentified validation pairs, public data, reproduction scripts, expected metrics, uncertainty outputs, and the authoritative validation workbook. Retrospective Review Protocol and Amendment Log. Retrospective protocol, search and selection methods, human-review roles, and amendment history. Supplementary File 7. Supplementary Figures S1–S6. Validation and independent-review workflow; geographical distribution; temporal validation-strategy extraction signals; clinical applications; AI method/model families; and DX120 field-level F1. Supplementary File 8. PRISMA and Full-Text Exclusion Audit. PRISMA counts, recorded decisions for 2586 non-included reports, F1–F10 code definitions, and a record-level retrieval-provenance audit. Supplementary File 9. Exploratory Extraction Figures and Validation Uncertainty. Exploratory Figures E1 and E2 and complete validation point estimates with bootstrap 95% confidence inter-

vals. Exploratory numeric outputs are not prevalence estimates. Supplementary File 10. Complete Bibliographic Citations for the 1181 Included Reports. Complete included-report bibliography; original manuscript reference numbers 31–1211 are retained for traceability.

Author Contributions: P.Ö.: Conceptualization; Methodology; Software; Formal Analysis; Investigation; Data Curation; Visualization; Project Administration; Supervision; Writing—Original Draft; Writing—Review and Editing. A.O.: Validation; Investigation; Data Curation; Methodology Review; Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable. This scoping review analyzed the published literature and publicly available bibliographic data and did not involve human participant recruitment or identifiable participant-level data.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supplementary File 4 (Supplementary Data S1) contains the final public dataset of 1181 reports and 66 variables, with the operational codebook, missingness summary, field-reliability labels, field-level provenance matrix, correction audits, attestation-evidence metadata, and 52-row merge audit. Supplementary Files 1–10 are provided in the consolidated package `tijfs-219_Supplementary_Files.zip`. Deidentified validation pairs and reproducibility scripts are provided in Supplementary File 5; the PRISMA counts, recorded non-inclusion decisions and retrieval-provenance audit are provided in Supplementary File 8. Named raw reviewer workbooks, actual signed author approvals and copyrighted full texts are retained in the restricted/source-lock archive.

Conflicts of Interest: The authors declare no conflicts of interest.

Code Availability: Deidentified reproducibility runner files are provided in Supplementary File 5. These include the publicly available $N = 1181$ dataset, the recorded F1–F10 audit for 2586 non-included reports, the minimal duplicate report mapping, de-identified validation pairs, scripts, expected point metrics, bootstrap output files, and manifest files. This runner recreates PRISMA statistics, stage and source distribution, exclusion code counts, report summary statistics, TA200/FT150/DX120 statistics, record-level bootstrap intervals, and clustered DX120 interval pooled estimates. No operational scripts with local path specifications are redistributed. The runner reproduces the supplied data; source-level documentation is described in Supplementary File 8.

Use of Generative AI Tools: OpenAI ChatGPT, including GPT-5-family models accessed through the ChatGPT Work project environment during July and August 2026, was used for structured prescreening support, coding assistance, code drafting and debugging, consistency checks, figure and document preparation, and language editing. The interface did not consistently expose a stable deployment or build identifier for all sessions; no identifier was inferred. Generative AI did not make final eligibility, exclusion-code, evaluation-stage, or source-level charting decisions. PÖ and AO reviewed, corrected, and approved all AI-assisted outputs and accept full responsibility for the integrity and accuracy of the work. Supplementary File 5 contains the deidentified reproducibility materials described above. The original prompt–response transcript is not included in Supplementary Files 1–10. Additional AI-assisted language editing and document preparation took place in September 2026.

Abbreviations

AC1: Gwet agreement coefficient 1; AI: artificial intelligence; AI-CDSS: artificial intelligence-based clinical decision support system; API: application programming interface; CDSS: clinical decision support system; CI: confidence interval; DOI: Digital Object Identifier; DX: data extraction; EHR: electronic health record; EMR: electronic medical record; EU: European Union; FDA: US Food and Drug Administration; FHIR: Fast Healthcare Interoperability Resources; FT: full text; ICMJE: International Committee of Medical Journal Editors; IMDRF: International Medical Device Regulators Forum; ML: machine learning; NR: not reported or not captured; NPV: negative predictive value; ORCID: Open Researcher and Contributor ID; PACS: picture archiving and communication system; PCC: Population–Concept–Context; PMC: PubMed Central; PMID: PubMed identifier; PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; SaMD: software as a medical device; TA: title/abstract.

References

1. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [[CrossRef](#)] [[PubMed](#)]
2. Marwaha, J.S.; Kvedar, J.C. Crossing the chasm from model performance to clinical impact: The need to improve implementation and evaluation of AI. *npj Digit. Med.* **2022**, *5*, 25. [[CrossRef](#)] [[PubMed](#)]
3. Zhou, Q.; Chen, Z.H.; Cao, Y.H.; Peng, S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: A systematic review. *npj Digit. Med.* **2021**, *4*, 154. [[CrossRef](#)] [[PubMed](#)]

4. Han, R.; Acosta, J.N.; Shakeri, Z.; Ioannidis, J.P.A.; Topol, E.J.; Rajpurkar, P. Randomized controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *Lancet Digit. Health* **2024**, *6*, e367–e373. [CrossRef] [PubMed]
5. van de Sande, D.; Chung, E.F.F.; Oosterhoff, J.; van Bommel, J.; Gommers, D.; van Genderen, M.E. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *npj Digit. Med.* **2024**, *7*, 58. [CrossRef] [PubMed]
6. Wenderott, K.; Krups, J.; Zaruchas, F.; Weigl, M. Effects of artificial intelligence implementation on efficiency in medical imaging: A systematic literature review and meta-analysis. *npj Digit. Med.* **2024**, *7*, 265. [CrossRef] [PubMed]
7. Tun, H.M.; Rahman, H.A.; Naing, L.; Malik, O.A. Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review. *J. Med. Internet Res.* **2025**, *27*, e69678. [CrossRef] [PubMed]
8. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*(11), e745–e750. [CrossRef] [PubMed]
9. Sallam, M.; Snygg, J.; Allam, D.; Kassem, R.; Damani, M. Artificial intelligence in clinical medicine: A SWOT analysis of AI progress in diagnostics, therapeutics, and safety. *J. Innov. Med. Res.* **2025**, *4*(3), 1–20. [CrossRef]
10. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*(6464), 447–453. [CrossRef] [PubMed]
11. Panch, T.; Mattie, H.; Atun, R. Artificial intelligence and algorithmic bias: Implications for health systems. *J. Glob. Health* **2019**, *9*(2), 020318. [CrossRef] [PubMed]
12. Ciecierski-Holmes, T.; Singh, R.; Axt, M.; Brenner, S.; Barteit, S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review. *npj Digit. Med.* **2022**, *5*, 162. [CrossRef] [PubMed]
13. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*; World Health Organization: Geneva, Switzerland, 2021; Available online: <https://www.who.int/publications/i/item/9789240029200> (accessed on 17 August 2026).
14. World Health Organization. *Regulatory Considerations on Artificial Intelligence for Health*; World Health Organization: Geneva, Switzerland, 2023; Available online: <https://www.who.int/publications/i/item/9789240078871> (accessed on 17 August 2026).
15. US Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. Final Guidance for Industry and Food and Drug Administration Staff. August 2025. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence> (accessed on 17 August 2026).
16. US Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Final Guidance. January 2026. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software> (accessed on 17 August 2026).
17. US Food and Drug Administration. Measuring and Evaluating Artificial Intelligence-Enabled Medical Device Performance in the Real World: Request for Public Comment. 2025. Available online: <https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device> (accessed on 17 August 2026).
18. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): Clinical Evaluation. IMDRF/SaMD WG/N41FINAL:2017. 21 September 2017. Available online: <https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation> (accessed on 17 August 2026).
19. International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. IMDRF/AIML WG/N88 FINAL:2025. 29 January 2025. Available online: <https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles> (accessed on 17 August 2026).
20. European Parliament; Council of the European Union. Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139, and (EU) 2023/1230 as regards simplification of the implementation of harmonized rules on artificial intelligence. *Off. J. Eur. Union* **2026**, *L* 2026, 1744. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744 (accessed on 17 August 2026).
21. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef] [PubMed]
22. Cruz Rivera, S.; Liu, X.; Chan, A.W.; Denniston, A.K.; Calvert, M.J.; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat. Med.* **2020**, *26*, 1351–1363. [CrossRef] [PubMed]
23. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
24. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef] [PubMed]
25. Sounderajah, V.; Guni, A.; Liu, X.; Collins, G.S.; Karthikesalingam, A.; Markar, S.R.; Golub, R.M.; Denniston, A.K.; Shetty, S.; Moher, D.; et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **2025**, *31*, 3283–3289. [CrossRef] [PubMed]
26. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*(2), e200029. [CrossRef] [PubMed]
27. Tejani, A.S.; Klontzas, M.E.; Gatti, A.A.; Mongan, J.T.; Moy, L.; Park, S.H.; Kahn, C.E., Jr.; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **2024**, *6*(4), e240300. [CrossRef] [PubMed]
28. Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **2025**, *388*, e081554. [CrossRef] [PubMed]

29. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*(7), 467–473. [[CrossRef](#)] [[PubMed](#)]
30. Peters, M.D.J.; Marnie, C.; Tricco, A.C.; Pollock, D.; Munn, Z.; Alexander, L.; McInerney, P.; Godfrey, C.M.; Khalil, H. Updated methodological guidance for the conduct of scoping reviews. *JBIM Evid. Synth.* **2020**, *18*(10), 2119–2126. [[CrossRef](#)] [[PubMed](#)]