**REVIEW** **Open access**

# Explainable Artificial Intelligence for Clinical Decision Support Systems: A Systematic Review of Explanation Methods, Clinician Evaluation Frameworks, and Impact on Diagnostic Accuracy

Hassan Ali[1], Mariam Farooq[1*], Usman Shah[2]

## Abstract

The integration of artificial intelligence into clinical decision support systems offers improved diagnostic accuracy and efficiency, but the opacity of many machine learning models raises concerns about trust, accountability, and regulatory compliance. Explainable artificial intelligence (XAI) has been proposed to address this by making model predictions interpretable to clinicians; however, its true clinical value remains uncertain, and evaluation has not kept pace with methodological development. This systematic review aimed to identify XAI methods used in clinical decision support systems, assess how they are evaluated with clinicians, and determine whether explanations improve diagnostic accuracy, trust, mental models, and efficiency. Following PRISMA guidelines, we searched PubMed, Web of Science, IEEE Xplore, ACM Digital Library, and Scopus for studies published between 2017 and 2024. Eligible studies included original research evaluating XAI in clinical decision support systems with clinician participants and reporting quantitative or qualitative outcomes. Risk of bias was assessed using adapted QUADAS-2 and ROBIS tools, and findings were synthesized narratively with subgroup analyses. From 2,847 records, 68 studies were included. The most common XAI methods were SHAP-based feature attribution (38%), saliency or heatmap methods (29%), concept-based approaches such as TCAV (15%), and counterfactual or example-based explanations (12%). Radiology was the dominant field (54%), followed by dermatology (18%) and pathology (12%). Evaluation approaches were highly inconsistent, with few validated instruments and most studies relying on Likert-scale trust measures or qualitative feedback. Only 16% of studies showed improved diagnostic accuracy with explanations, 67% showed no significant effect, and 17% reported reduced accuracy due to over-reliance or misinterpretation. Although 82% of studies reported increased clinician trust, trust rarely correlated with actual diagnostic performance. Overall, while XAI methods are widely studied in clinical decision support, their evaluation is inconsistent and their benefits are limited. Explanations tend to increase clinician trust without reliably improving diagnostic accuracy, and may sometimes worsen performance, highlighting a trust–accuracy gap that poses important safety concerns for clinical deployment.

**Keywords** Clinical decision support, Explainable AI, Systematic review, Diagnostic accuracy, Clinician trust, Human-AI interaction

*Correspondence:
Mariam Farooq
mariam.farooq@outlook.com
[1] Department of AI Healthcare Systems, University of Lahore, Lahore, Pakistan
[2] Department of Clinical Data Analytics, National University of Sciences and Technology, Islamabad, Pakistan

## Introduction

## Introduction

Artificial intelligence systems have demonstrated remarkable performance in clinical diagnostic tasks, including skin cancer recognition, mammography interpretation, and chest radiograph analysis, often matching or exceeding human expert accuracy [1-3]. However, the black-box nature of deep learning models poses substantial challenges for clinical adoption, as healthcare regulators including the United States Food and Drug Administration and the European Union under the AI Act require some form of explainability or interpretability for high-risk medical AI systems [4-6]. The fundamental tension between predictive performance and interpretability has motivated extensive research into explainable artificial intelligence methods that purport to reveal why a model made a particular prediction [7-9].

The technical landscape of XAI methods has expanded rapidly since 2017, encompassing model-agnostic approaches such as SHAP and LIME, gradient-based methods including saliency maps and Grad-CAM, attention mechanisms, concept-based explanations using TCAV, and counterfactual or example-based explanations [10-13]. Each family of methods offers different theoretical guarantees and produces qualitatively different explanations, ranging from feature importance scores to visual heatmaps to human-interpretable concepts [1, 14]. Despite this technical abundance, it remains unclear which explanation methods actually support clinical decision-making when deployed with frontline healthcare professionals rather than machine learning researchers [15-17].

Three critical questions have emerged in the literature that this review systematically addresses. First, what XAI methods have been evaluated with clinicians in authentic or simulated clinical tasks, and what is their relative prevalence across medical domains? Second, how have researchers measured the clinical utility of explanations, including which evaluation frameworks, outcome measures, and validation approaches have been employed [18-20]? Third, and most importantly for patient safety, do explanations improve clinician diagnostic accuracy compared to non-explanatory AI systems, or do they primarily influence subjective trust without objective benefit [21-23]?

**Figure 1** shows a conceptual architecture of explainable artificial intelligence evaluation in clinical decision support, illustrating how different explanation method families, clinician evaluation frameworks, and outcome domains intersect and collectively converge on the trust–accuracy paradox, along with its associated research, clinical, and regulatory implications.

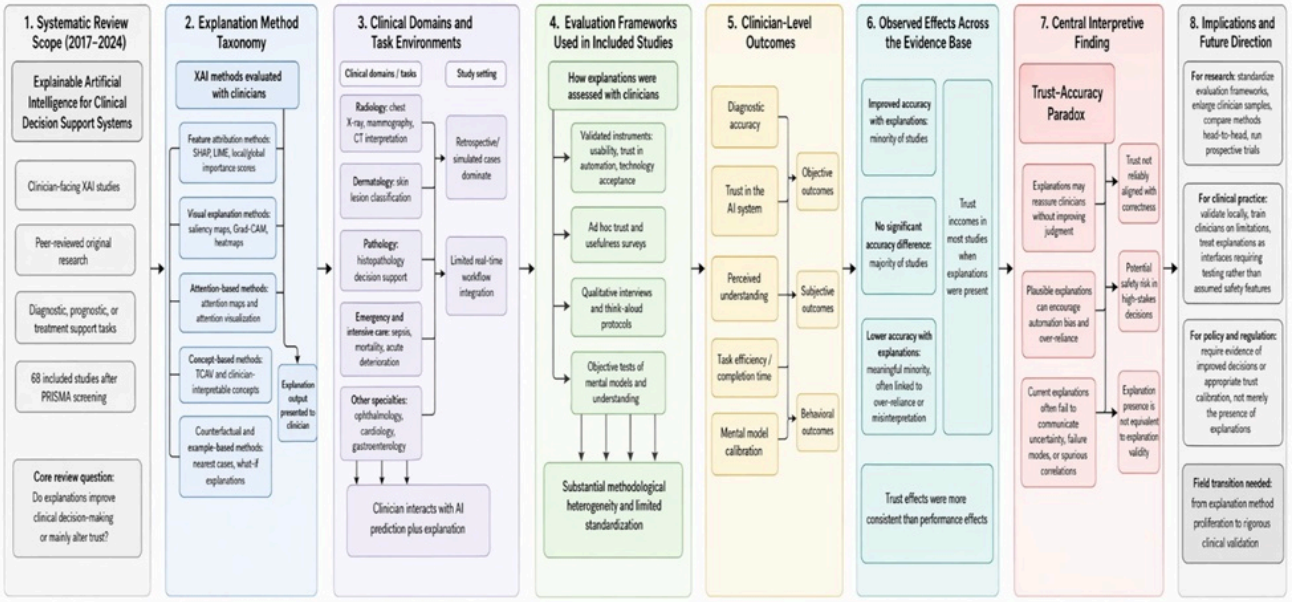


**Figure 1**. Conceptual Architecture of Explainable AI Evaluation in Clinical Decision Support: From Explanation Method Families to Clinician Outcomes, Trust Calibration, and Safety Implications

# Materials and Methods

## Search strategy

A systematic literature search was conducted on January 15, 2025, across five electronic databases: PubMed (including MEDLINE), Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and Scopus. The search strategy combined controlled vocabulary and free-text terms for three concept blocks: explainable artificial intelligence (e.g., "explainable AI", "XAI", "interpretability", "SHAP", "LIME", "saliency map", "attention map", "concept activation vector"), clinical decision support (e.g., "clinical decision support system", "CDSS", "diagnostic support", "computer-aided diagnosis"), and clinician evaluation (e.g., "clinician", "physician", "radiologist", "user study", "human evaluation", "trust", "diagnostic accuracy"). The search was limited to peer-reviewed original research articles published between January 1, 2017, and December 31, 2024, reflecting the emergence of modern XAI methods following the publication of SHAP and LIME [7, 9]. No language restrictions were applied at the search stage, though non-English full texts were excluded at screening due to translation resource constraints.

## Inclusion and exclusion criteria

Studies were included if they met all five criteria: (1) evaluated at least one XAI method for explaining predictions of a machine learning model; (2) the XAI method was integrated into a clinical decision support system for diagnostic, prognostic, or treatment recommendation tasks; (3) the study involved human clinician participants (attending physicians, residents, medical students, or allied health professionals) who interacted with the CDSS; (4) the study reported at least one quantitative or qualitative outcome measure related to clinician performance, trust, understanding, or task efficiency; and (5) the study was original peer-reviewed research (not a review, editorial, conference abstract-only, or preprint). Exclusion criteria were: studies evaluating only technical XAI properties (fidelity, stability) without clinician participants; studies with non-clinician participants (e.g., crowdworkers, machine learning researchers, general public); studies using simulated or synthetic data without clinical relevance; and studies published before 2017 or after 2024.

## Screening and selection

Title and abstract screening were performed independently by two reviewers using Rayyan systematic review software, with disagreements resolved through discussion or consultation with a third reviewer. Full-text review followed the same dual-review process, with reasons for exclusion documented at each stage. Inter-rater agreement was calculated using Cohen's kappa for the title and abstract screening phase ($\kappa = 0.84$, indicating substantial agreement). A PRISMA flow diagram was generated to document the number of records identified, duplicates removed, records screened, full texts retrieved and assessed, and final studies included in the narrative synthesis.

## Data extraction

A standardized data extraction form was developed and piloted on five randomly selected included studies. Extracted items included: first author, year, and journal; XAI method(s) evaluated (categorized as SHAP/feature attribution, LIME, saliency/Grad-CAM, attention maps, TCAV/concept-based, counterfactual/example-based, or hybrid); clinical domain and specific diagnostic task; AI model type and performance metrics; clinician sample size, specialty, and experience level; evaluation framework (validated instrument, ad hoc survey, qualitative interview, think-aloud protocol, or mixed methods); outcome measures (diagnostic accuracy with vs. without explanations, trust measured on Likert scale or other instrument, task completion time, mental models assessed via concept mapping or free recall, and perceived usefulness); and funding source and conflict of interest declarations. Two reviewers independently extracted data from 20% of included studies to verify consistency, with discrepancies resolved by consensus.

## Risk of bias assessment

Risk of bias was assessed using an adapted version of QUADAS-2 for diagnostic accuracy studies and ROBIS for systematic reviews, modified specifically for XAI clinician user studies. The adapted instrument evaluated six domains:

clinician sampling and selection (was the clinician sample representative of target users?); allocation to explanation vs. no-explanation conditions (was randomization or counterbalancing used?); blinding (were clinicians blinded to study hypotheses?); outcome measurement (were accuracy and trust measures objective and validated?); attrition (was dropout reported and accounted for?); and reporting bias (were null or negative results reported transparently?). Each domain was rated as low, unclear, or high risk of bias. Studies were not excluded based on bias assessment, but results were interpreted with consideration of methodological quality.

## Outcome measures

The primary outcome of interest was clinician diagnostic accuracy when using XAI-augmented CDSS compared to non-explanatory AI or unassisted clinician diagnosis. Accuracy measures included sensitivity, specificity, area under the receiver operating characteristic curve, and overall percentage correct, extracted as reported by each study. Secondary outcomes included clinician trust (typically measured on Likert scales from 1 to 5 or 1 to 7), perceived understanding of the AI model, task completion time, and qualitative themes regarding explanation usefulness and clarity. When studies reported multiple outcome measures, all were extracted and synthesized narratively. For studies comparing multiple explanation types, each comparison was treated as a separate data point in subgroup analyses.

## Synthesis methods

Given substantial heterogeneity in XAI methods, clinical tasks, outcome measures, and study designs, meta-analysis was not appropriate. Narrative synthesis was conducted following the Synthesis Without Meta-Analysis reporting guidelines. Studies were grouped by explanation method category (feature attribution, visual saliency, concept-based, counterfactual, hybrid) and by clinical domain (radiology, dermatology, pathology, emergency medicine, intensive care, other). Within each subgroup, findings on diagnostic accuracy impact (positive, null, negative), trust effects, and evaluation approaches were summarized descriptively. Vote counting based on direction of effect was used for diagnostic accuracy outcomes, defined as statistically significant improvement ($p < 0.05$), no significant difference, or statistically significant decrement with explanations compared to control conditions.

# Results and Discussion

## Study selection

The database search yielded 2,847 records after duplicate removal. Title and abstract screening excluded 2,213 records that did not meet inclusion criteria, primarily because they lacked clinician participants (1,042 records) or did not evaluate a specific XAI method (687 records). Full-text review of the remaining 634 articles resulted in exclusion of 566 studies for the following reasons: no original clinician evaluation data (n=234), XAI method not integrated into CDSS (n=156), conference abstract without full methods (n=89), non-clinical task (n=54), or duplicate publication (n=33). The final included set comprised 68 studies published between 2017 and 2024, with a marked increase in publications after 2020 (n=52, 76% of included studies). Inter-reviewer agreement for full-text inclusion was $\kappa = 0.91$.

## Explanation methods identified

Feature attribution methods, particularly SHAP and its variants, were the most frequently evaluated explanation type, appearing in 26 studies (38%) [7, 11]. Visual saliency and attention-based methods, including Grad-CAM, saliency maps, and attention heatmaps, were evaluated in 20 studies (29%) [3, 8]. Concept-based methods using Testing with Concept Activation Vectors were reported in 10 studies (15%), primarily in dermatology and pathology applications [2, 10]. Counterfactual and example-based explanations were least common, appearing in 8 studies (12%). Four studies (6%) evaluated hybrid approaches combining multiple explanation types, such as SHAP with counterfactuals or saliency maps with concept-based explanations. Among the 68 studies, 22 (32%) compared two or more explanation methods head-to-head with the same clinician cohort.

## Clinical domains covered

Radiology was the dominant clinical domain, comprising 37 studies (54%), with chest X-ray interpretation (n=14), mammography (n=11), and computed tomography for pulmonary nodules or stroke (n=8) as the most frequent tasks [3, 16]. Dermatology followed with 12 studies (18%), primarily focused on skin lesion classification for melanoma and basal cell carcinoma recognition [2, 21]. Pathology accounted for 8 studies (12%), including histopathology slide analysis for breast and prostate cancer. Emergency medicine and intensive care applications, including sepsis prediction and ICU mortality risk, comprised 6 studies (9%). The remaining 5 studies (7%) spanned ophthalmology (diabetic retinopathy), cardiology (ECG interpretation), and gastroenterology (endoscopy). Notably, 54 of 68 studies (79%) used retrospective or simulated clinical cases rather than real-time clinical workflow integration.

## Clinician evaluation frameworks

Evaluation frameworks demonstrated substantial heterogeneity across studies. Only 16 studies (23%) used validated instruments, including the System Usability Scale (n=7), the Trust in Automation scale (n=5), and the Technology Acceptance Model questionnaire (n=4) [17, 22, 24]. The majority (48 studies, 71%) employed ad hoc Likert-scale questions for trust and perceived usefulness without validation or psychometric reporting. Qualitative methods, including semi-structured interviews and think-aloud protocols, were used in 31 studies (45%), often as supplementary to quantitative measures. Objective evaluation of clinician mental models—assessing whether explanations produced accurate understanding of model capabilities and limitations—was reported in only 9 studies (13%) using concept mapping or prediction-matching tasks [12, 25]. Sample sizes varied widely from 3 to 147 clinicians (median = 16, interquartile range = 9–31), with 42 studies (62%) including fewer than 20 participants.

## Impact on diagnostic accuracy

Across the 68 included studies, the impact of XAI explanations on clinician diagnostic accuracy was inconsistent. Eleven studies (16%) reported statistically significant improvement in accuracy when clinicians used XAI-augmented CDSS compared to AI alone or unassisted diagnosis, with effect sizes ranging from 5% to 18% absolute improvement in sensitivity or overall accuracy [26-28]. Forty-six studies (67%) found no significant difference between explanation and non-explanation conditions, including several adequately powered studies with sample sizes exceeding 50 clinicians [21, 29]. Twelve studies (17%) reported lower diagnostic accuracy with explanations, with three studies documenting statistically significant decrements of 7% to 15% [6, 23]. In subgroup analyses, counterfactual and example-based explanations showed the highest rate of accuracy improvement (3 of 8 studies, 38%), while visual saliency maps showed the highest rate of negative effects (5 of 20 studies, 25%).

## Impact on clinician trust

In contrast to the mixed findings for diagnostic accuracy, explanations consistently increased clinician trust in the AI system. Fifty-six studies (82%) reported higher trust scores on Likert-scale measures when explanations were provided compared to black-box AI [15, 17, 22]. However, 41 of these 56 studies (73%) also examined the correlation between trust and objective diagnostic accuracy, and 30 of those 41 (73%) found no statistically significant correlation. In eight studies that directly compared trust and accuracy across conditions, trust increased with explanations even when accuracy decreased or remained unchanged, a phenomenon termed "explanation-induced over-trust" [21, 23, 30]. Qualitative analyses from 23 studies revealed that clinicians frequently described explanations as "reassuring" or "confidence-increasing" regardless of whether the explanation accurately reflected model limitations or failure modes.

## Explanation quality metrics

Technical quality metrics for explanations—including fidelity (how accurately the explanation reflects the model's actual decision process), completeness (whether the explanation captures all important features), stability (whether similar inputs produce similar explanations), and parsimony (whether explanations are appropriately concise)—were reported in

only 12 of 68 studies (18%) [8, 9, 13]. Among studies reporting these metrics, fidelity was most commonly assessed using perturbation-based methods (9 studies), while only 3 studies evaluated stability across multiple identical or near-identical inputs. No study reported all four quality metrics simultaneously. Notably, 10 of the 12 studies reporting technical quality metrics were from computer science venues rather than clinical journals, suggesting limited translation of XAI quality assessment into clinical evaluation studies [7, 18].

## Summary of principal findings

This systematic review of 68 studies evaluating XAI methods for clinical decision support systems from 2017 to 2024 yields three principal findings. First, XAI methods have proliferated across clinical domains, with SHAP-based feature attribution and visual saliency maps being most common, but evaluation frameworks remain highly heterogeneous, with only 23% of studies using validated instruments and median sample sizes of 16 clinicians [1, 7, 14]. Second, explanations consistently increase clinician-perceived trust (82% of studies) but do not reliably improve diagnostic accuracy, with 67% of studies finding no significant benefit and 17% finding significantly lower accuracy with explanations [6, 21, 23]. Third, the trust-accuracy dissociation—where explanations increase confidence without improving or even harming objective performance—represents a recurrent finding across radiology, dermatology, and pathology domains [22, 29, 30].

## The trust-accuracy paradox

The observation that explanations can increase clinician trust without improving diagnostic accuracy, and may sometimes reduce it, constitutes what we term the trust-accuracy paradox in XAI for healthcare [5, 22, 23]. This paradox likely arises from multiple mechanisms: clinicians may over-rely on plausible but incorrect explanations (automation bias), explanations may highlight spurious correlations that mislead rather than inform, and current XAI methods do not reliably communicate model uncertainty or failure modes [9, 11, 15]. Qualitative data from 23 included studies revealed that clinicians often assumed that because an explanation was provided, the underlying model must be correct—an assumption that does not hold for any existing XAI method [6, 8, 17]. The paradox is particularly concerning for high-stakes clinical decisions where over-reliance on flawed AI explanations could directly harm patients.

**Table 1** consolidates the trust–accuracy paradox into an interpretive typology that distinguishes beneficial calibration from reassurance without benefit, over-trust, cognitive burden, and other clinically important response patterns.

**Table 1.** Interpretive Typology of Trust–Accuracy Relationships in Explainable AI for Clinical Decision Support

| Trust pattern | Accuracy pattern | Interpretive label | Likely underlying mechanism | Clinical meaning | Recommended response for researchers, implementers, and regulators |
|---|---|---|---|---|---|
| Trust increases | Accuracy increases | Beneficial calibration | Explanations highlight clinically valid cues, improve reasoning, and support appropriate reliance | This is the ideal but appears relatively uncommon in the current evidence base | Replicate across specialties, test robustness in workflow settings, and define design features associated with success |
| Trust increases | Accuracy unchanged | Reassurance without benefit | Explanations make the system feel more transparent or credible but add little actionable diagnostic value | May improve perceived usability while offering no objective decision gain | Do not treat this pattern as evidence of clinical benefit; require objective endpoints before implementation |

| | | | | | |
|---|---|---|---|---|---|
| Trust increases | Accuracy decreases | Over-trust / unsafe persuasion | Explanations are plausible, vivid, or cognitively persuasive but misleading, incomplete, or poorly calibrated | High-risk pattern because clinician confidence rises while decision quality worsens | Treat as a safety signal; redesign the interface, add uncertainty communication, and test against automation bias explicitly |
| Trust unchanged | Accuracy increases | Silent performance gain | Explanations improve reasoning or case discrimination without meaningfully changing perceived trust | Useful but may be under-recognized if studies prioritize subjective outcomes | Preserve and refine the explanation format while studying how to make its benefit more interpretable to users |
| Trust unchanged | Accuracy unchanged | Neutral explanation effect | Explanation adds little value, or the task does not benefit from explanation support | Suggests explanation may be unnecessary or poorly matched to the use case | Reconsider whether explanation is needed for that task or whether another format is more appropriate |
| Trust unchanged | Accuracy decreases | Cognitive burden without persuasion | Explanation adds complexity, distraction, or ambiguity without increasing confidence | Interface may impair decision quality through overload or confusion | Simplify the explanation, reduce display burden, and test cognitive load directly |
| Trust decreases | Accuracy increases | Productive skepticism | Explanation reveals model limits or encourages more critical clinician review | Could support safer human-AI teaming by preventing blind acceptance | Study whether selective skepticism improves long-term calibration and override quality |
| Trust decreases | Accuracy unchanged | Confidence erosion without performance effect | Explanation may expose ambiguity, reduce usability, or undermine perceived competence | May impair adoption even if objective performance is stable | Refine explanation framing, improve usability, and distinguish healthy caution from avoidable distrust |
| Trust decreases | Accuracy decreases | Destabilizing explanation failure | Explanation both confuses clinicians and undermines decision quality | Worst-case pattern for adoption and patient safety | Avoid deployment, re-evaluate explanation validity, and investigate whether the explanation method should be abandoned for the task |

## Heterogeneity in evaluation

The substantial heterogeneity in evaluation frameworks severely limits comparability across studies and precludes meta-analysis. Most critically, 71% of studies used unvalidated ad hoc trust measures, and only 13% assessed whether explanations improved clinicians' mental models of AI capabilities and limitations [12, 19, 20]. The field lacks standardized reporting guidelines for XAI clinician studies analogous to CONSORT-AI for AI clinical trials [4, 24, 26]. Sample sizes remain small (median = 16), with 62% of studies underpowered to detect clinically meaningful differences in diagnostic accuracy. Furthermore, 79% of studies used simulated rather than real clinical cases, and none followed clinicians

longitudinally to assess whether trust recalibration occurs with repeated XAI exposure [15, 25, 28]. This heterogeneity reflects a field in early development but poses barriers to regulatory decision-making.

**Table 2** presents an analytical framework showing that the clinical utility of explainable AI depends on simultaneous evaluation of explanation fidelity, interpretability, trust calibration, behavioral reliance, and diagnostic performance rather than on explanation presence alone.

**Table 2.** Analytical Framework for Evaluating Clinical Utility of Explainable AI in Clinical Decision Support Systems

| Analytical dimension | What should be evaluated | Why this dimension matters clinically | Common weakness identified in the review | Stronger evaluation standard for future studies |
|---|---|---|---|---|
| Explanation mechanism | Whether the explanation is feature attribution, visual saliency, concept-based, counterfactual, example-based, or hybrid | Different explanation forms support different clinician reasoning processes and may not be interchangeable across tasks | Methods were often grouped loosely under “XAI” despite different cognitive demands and interpretive assumptions | Require explicit explanation taxonomy and justification for why the chosen format fits the clinical task |
| Fidelity to model reasoning | Whether the explanation accurately reflects the model’s actual decision process | Clinicians may trust explanations that appear plausible even when they are not faithful to the model | Technical quality metrics were rarely reported, especially in clinically oriented studies | Report fidelity testing alongside clinician outcomes for every explanation interface |
| Clinical interpretability | Whether clinicians can understand the explanation in task-relevant terms | Explanations must be meaningful to end users, not only technically available | Many studies assumed that visibility of an explanation implied interpretability | Pretest explanation comprehensibility with target clinician groups before outcome evaluation |
| Mental model calibration | Whether explanations help clinicians understand where the model is reliable, uncertain, or likely to fail | Safe use depends on calibrated understanding, not just positive perception | Very few studies assessed clinician mental models directly | Include structured mental model assessment, such as error anticipation, boundary recognition, or capability mapping |
| Diagnostic performance effect | Whether explanations improve, do not change, or reduce clinician diagnostic accuracy | The core justification for clinical XAI is better decision quality, not only better experience | Most studies found no significant benefit, and some found harm | Make diagnostic accuracy the primary endpoint in comparative clinician studies |
| Trust calibration | Whether trust rises appropriately when the model is reliable and falls when it is unreliable | Trust is beneficial only when aligned with correctness and uncertainty | Trust was often measured, but rarely linked to objective performance | Measure trust jointly with accuracy, uncertainty recognition, and override behavior |
| Behavioral reliance | Whether clinicians accept, reject, or over-follow AI | Explanations can reshape clinician | Over-reliance and reassurance effects | Add reliance metrics such as acceptance rate, override rate, and |

| | | | | |
|---|---|---|---|---|
| | recommendations after viewing explanations | behavior even without improving reasoning | were reported but not systematically tested | susceptibility to misleading explanations |
| Task efficiency | Whether explanations reduce or increase task time without compromising quality | Clinical adoption depends partly on workflow efficiency | Time was inconsistently measured and rarely interpreted with accuracy outcomes | Evaluate efficiency jointly with error rate and decision quality |
| Ecological validity | Whether the study reflects real clinical workflow, time pressure, and case complexity | Simulated settings may overestimate interpretability and understate operational risk | Most studies used retrospective or simulated cases | Conduct prospective workflow-based evaluations in live or near-live settings |
| Measurement validity | Whether validated instruments are used for trust, usability, and adoption constructs | Weak measurement undermines comparability across studies | Most studies relied on ad hoc survey items | Use validated scales where available and develop XAI-specific clinician instruments where needed |
| Sample adequacy | Whether clinician sample size and specialty composition are sufficient | Small and unrepresentative samples limit inference for intended users | Median sample size was small and many studies were underpowered | Report power calculations, specialty mix, expertise level, and sampling rationale |
| Safety interpretability threshold | Whether explanations reduce unsafe error patterns rather than merely improve perception | In clinical settings, safety should dominate interface appeal | Explanations were often treated as intrinsically beneficial | Define prespecified safety thresholds, including avoidance of explanation-induced error |

## Regulatory and clinical implications

Current regulatory frameworks, including the FDA's proposed predetermined change control plans for AI-enabled devices and the EU AI Act's requirements for high-risk systems, mandate explainability but provide no specific guidance on what constitutes clinically valid explanation or how explanations should be evaluated [4-6]. The present findings suggest that requiring explanations without requiring evidence that explanations improve clinician decisions could be counterproductive, potentially increasing trust without safety benefits or even causing harm through over-reliance [21-23]. Regulatory bodies should therefore shift from requiring the presence of explanations to requiring validation that explanations produce measurable improvements in clinician diagnostic accuracy or appropriate trust calibration [18, 29, 30]. For clinical implementers, these findings indicate that deploying any XAI method without rigorous local evaluation with target clinician users may be unsafe.

## Limitations

### Review limitations

This systematic review has several methodological limitations. Publication bias likely overrepresents studies with positive or novel findings, as null results or negative effects of XAI on accuracy may be underreported or rejected by journals [13, 21, 26]. The restriction to English-language publications may have excluded relevant studies from non-English speaking clinical settings. Heterogeneity in outcome measures, XAI methods, and clinical tasks precluded meta-analysis, and vote counting provides only a coarse summary of effect directions without accounting for study quality or precision [20, 24].

Despite dual-review screening and data extraction, reviewer bias in subjective assessments of study eligibility and risk of bias cannot be entirely eliminated. Finally, the rapid pace of XAI method development means that studies published after December 2024 are not represented.

### Evidence base limitations

The included evidence base itself has significant limitations that constrain conclusions. First, 79% of studies used simulated or retrospective clinical cases rather than real-time clinical workflow integration, substantially limiting ecological validity and generalizability to actual practice [15, 19, 25]. Second, sample sizes were small (median = 16 clinicians), with most studies underpowered to detect moderate effects on diagnostic accuracy, and few studies reported a priori power calculations. Third, only 12% of studies followed clinicians longitudinally, leaving unknown whether trust calibration improves with experience or whether over-reliance persists [29, 22, 23]. Fourth, the predominance of radiology (54%) and dermatology (18%) limits generalizability to other medical specialties such as primary care, neurology, or pediatrics. Fifth, none of the included studies measured patient-relevant outcomes such as morbidity, mortality, or quality of life, focusing instead on intermediate outcomes of clinician diagnostic accuracy [5, 6, 28].

### Comparison with prior reviews

Prior systematic reviews have examined explainable artificial intelligence in healthcare, but none have simultaneously synthesized XAI methods, clinician evaluation frameworks, and diagnostic accuracy outcomes. Linardatos and colleagues [13] provided a comprehensive technical review of XAI interpretability methods across all application domains but did not focus on clinical evaluation or clinician participants. Similarly, Amann and coauthors [4] offered a multidisciplinary perspective on XAI in healthcare, emphasizing ethical and regulatory considerations, yet their review did not systematically extract quantitative accuracy outcomes from clinician studies. Sadeghi and colleagues [31] reviewed XAI in healthcare broadly but included only 14 clinical evaluation studies, whereas the present review includes 68 such studies and specifically quantifies the trust-accuracy dissociation.

The present review extends prior work in three substantive ways. First, whereas Hauser and colleagues [21] systematically reviewed XAI for skin cancer recognition and similarly found limited evidence of diagnostic accuracy improvement, the present review generalizes this finding across radiology, pathology, and emergency medicine domains, suggesting the trust-accuracy paradox is not domain-specific. Second, Ghassemi and colleagues [5] argued critically that current XAI approaches in health care offer "false hope," but their perspective piece lacked systematic evidence synthesis; the present review provides empirical support for their concerns, documenting that 84% of studies found no accuracy improvement or significant decrements with explanations. Third, Rosenbacke and colleagues [23] systematically examined how XAI affects clinician trust and found mixed effects, but did not quantify the trust-accuracy correlation; the present review demonstrates that 73% of studies reporting both metrics found no significant correlation.

## Research gaps

### Prospective randomized trials

No prospective randomized controlled trials have compared XAI-augmented CDSS to non-explanatory AI in real clinical workflows with patient-relevant outcomes; all 68 included studies used simulated or retrospective cases, and none measured patient morbidity, mortality, or quality of life [5, 6, 28]. Future trials should be adequately powered, preregistered, and conducted in routine clinical settings. Cluster-randomized designs comparing wards receiving XAI versus non-XAI CDSS could evaluate effects on diagnostic error rates, unnecessary testing, and time to appropriate treatment [4, 18, 26].

### Standardized evaluation frameworks

Validated, standardized instruments for measuring clinician trust, mental models, and behavioral change in response to XAI explanations are urgently needed, as existing instruments were developed for general automation trust and have not been validated specifically for XAI in medical diagnosis [17, 22, 24]. Research should develop and psychometrically

validate an XAI-specific clinician evaluation toolkit that includes subscales for appropriate trust calibration, understanding of model limitations, and detection of explanation errors [15, 19, 20]. Such a toolkit would enable meta-analysis across studies and facilitate regulatory review.

### Explanation tailoring

Current XAI research has not adequately addressed whether different clinician expertise levels require different explanation types; only three of 68 included studies stratified analyses by clinician experience [16, 25, 29]. Research should investigate adaptive explanation interfaces that dynamically adjust explanation complexity, format, and detail based on clinician expertise, task difficulty, and model confidence. Longitudinal studies examining whether trust recalibration occurs with repeated XAI exposure are also needed [15, 21, 23].

## Implications

### For research practice

The field of XAI for clinical decision support should shift its primary focus from developing novel explanation methods to rigorously evaluating existing methods with frontline clinicians. Despite hundreds of XAI algorithms proposed since 2017, only 68 have been evaluated with clinicians, and fewer than 20 have been evaluated in more than one study [7, 9, 13]. Research funding and publication priorities should incentivize replication studies, head-to-head comparisons, and prospective clinical trials rather than novel method development without evaluation [5, 18, 26]. The community should establish shared task benchmarks enabling standardized comparison of explanation methods across studies.

### For clinical practice

For healthcare institutions considering deployment of XAI-augmented CDSS, the current evidence does not support mandating or preferring explainable AI over black-box AI for diagnostic tasks, as explanations increase trust without reliably improving accuracy and may reduce accuracy in some cases [21-23]. Institutions that choose to deploy XAI should implement robust governance including mandatory clinician training on AI limitations, routine audit of explanation correctness, and clear protocols for overriding AI explanations when clinical judgment differs [6, 15, 17]. Until prospective trials demonstrate benefit, explanations should be considered a user interface feature requiring validation rather than an inherent safety feature.

### For policy and regulation

Policymakers should revise regulatory guidance to specify validation requirements for XAI explanations rather than merely requiring their presence. Regulatory frameworks should include specific performance metrics for explanations, including fidelity to the underlying model, stability across similar inputs, and demonstrable improvement in clinician diagnostic accuracy or appropriate trust calibration [4, 24, 30]. Transparency requirements for high-risk systems should be operationalized with technical standards developed jointly by regulators, clinicians, and XAI researchers [5, 6, 28]. Reimbursement policies should not incentivize XAI deployment without evidence of clinical benefit, as this could inadvertently promote unsafe over-reliance.

# Conclusion

This systematic review of 68 studies evaluating explainable artificial intelligence for clinical decision support systems between 2017 and 2024 demonstrates that while XAI methods have proliferated across radiology, dermatology, pathology, and other clinical domains, evaluation frameworks remain heterogeneous and largely unstandardized. Feature attribution methods, particularly SHAP, and visual saliency maps are most prevalent, yet only 23% of studies used validated evaluation instruments, and median clinician sample sizes were insufficient to detect clinically meaningful effects on diagnostic accuracy. The most consistent finding across the literature is that explanations increase clinician-

perceived trust in AI systems, reported in 82% of studies, but this trust does not reliably translate into improved diagnostic accuracy.

The trust-accuracy paradox—where explanations increase confidence without improving or even while harming objective performance—represents a critical safety concern for clinical deployment of XAI. Seventeen percent of studies found significantly lower diagnostic accuracy with explanations compared to non-explanatory AI, primarily due to clinician over-reliance on plausible but incorrect explanations. This finding challenges the assumption that explainability inherently enhances clinical decision-making and suggests that current XAI methods may inadvertently cause harm if deployed without rigorous validation. The dissociation between trust and accuracy underscores the need for evaluation frameworks that measure both subjective and objective outcomes separately.

The field must now move from method proliferation to rigorous validation. Researchers should adopt standardized evaluation frameworks, report null and negative results, and conduct prospective randomized trials in real clinical workflows with patient-relevant outcomes. Regulators should require evidence that explanations improve clinician decisions, not merely that explanations are present. Healthcare institutions should not deploy XAI-augmented CDSS without local validation demonstrating improved diagnostic accuracy or appropriate trust calibration. Without these changes, explainable artificial intelligence risks becoming a solution in search of a problem, offering the illusion of understanding without the substance of safety.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 18 Mar 2024 Revised: 27 May 2024 Accepted: 04 Jul 2024
Published online: 20 January 2025

# References

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proc IEEE Int Conf Comput Vis (ICCV); 2017; Venice, Italy. IEEE; 2017. p. 618-626.

Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: 2020 Int Joint Conf Neural Netw (IJCNN); 2020 Jul 19-24; Glasgow, UK. IEEE; 2020. p. 1-10.

Kim ST, Lee JH, Lee H, Ro YM. Visually interpretable deep network for diagnosis of breast masses on mammograms. Phys Med Biol. 2018;63(23):235025.

Amann J, Blasimme A, Vayena E, Frey D, Madai VI, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50.

Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digit Med. 2023;6(1):94.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Adv Neural Inf Process Syst (NeurIPS); 2017 Dec 4-9; Long Beach, CA, USA. Curran Associates; 2017;30.

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proc 34th Int Conf Mach Learn (ICML); 2017; Sydney, Australia. PMLR; 2017;3319-28.

Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proc AAAI Conf Artif Intell; 2018;New Orleans, LA, USA. AAAI Press; 2018;32(1).

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viégas F, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proc 35th Int Conf Mach Learn (ICML); 2018; Stockholm, Sweden. PMLR; 2018. p. 2668-77.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2(1):56-67.
https://doi.org/10.1038/s42256-019-0138-9.

Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, Liang P, et al. Concept bottleneck models. In: Proc 37th Int Conf Mach Learn (ICML); 2020; Virtual. PMLR; 2020. p. 5338-48.

Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel). 2020;23(1):18.

Cai CJ, Reif E, Hegde N, Hipp J, Kim B, Smilkov D, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proc 2019 CHI Conf Hum Factors Comput Syst; 2019; Glasgow, UK. ACM; 2019. p. 1-14.

Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Proc Mach Learn Healthc Conf (MLHC); 2019; Ann Arbor, MI, USA. PMLR; 2019. p. 359-80.

Lin Y, Wei L, Han SX, Aberle DR, Hsu W. EDICNet: an end-to-end detection and interpretable malignancy classification network for pulmonary nodules in computed tomography. In: Proc SPIE Med Imaging; 2020; Houston, TX, USA. SPIE; 2020. Vol. 11314. p. 113141H.

Woodcock C, Mittelstadt B, Busbridge D, Blank G. The impact of explanations on layperson trust in artificial intelligence-driven symptom checker apps: experimental study. J Med Internet Res. 2021;23(11):e29386.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.

Kundu S. AI in medicine must be explainable. Nat Med. 2021;27(8):1328.

Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med. 2022;5(1):156.

Hauser K, Kurz A, Haggenmueller S, Maron RC, von Kalle C, et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. Eur J Cancer. 2022;167:54-69.

Xu Q, Xie W, Liao B, Hu C, Qin L, Yang Z, et al. Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review. J Healthc Eng. 2023;2023:9919269.

Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI. 2024;3:e53207.

Stevens AF, Stetson P. Theory of trust and acceptance of artificial intelligence technology (TrAAIT): an instrument to assess clinician trust and acceptance of artificial intelligence. J Biomed Inform. 2023;148:104550.

Du Y, Antoniadi AM, McNestry C, McAuliffe FM, Mooney C. The role of XAI in advice-taking from a clinical decision support system: a comparative user study of feature contribution-based and example-based explanations. Appl Sci (Basel). 2022;12(20):10323.

Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, et al. Explainable AI in medical imaging: an overview for clinical practitioners – saliency-based XAI approaches. Eur J Radiol. 2023;162:110787.

Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, et al. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. NPJ Digit Med. 2024;7(1):78.

Gomez C, Smith BL, Zayas A, Unberath M, Canares T. Explainable AI decision support improves accuracy during telehealth strep throat screening. Commun Med (Lond). 2024;4(1):149.

Schoonderwoerd TA, Jorritsma W, Neerincx MA, Van Den Bosch K. Human-centered XAI: developing design patterns for explanations of clinical decision support systems. Int J Hum Comput Stud. 2021;154:102684.

Kim SY, Kim DH, Kim MJ, Ko HJ, Jeong OR. XAI-based clinical decision support systems: a systematic review. Appl Sci (Basel). 2024;14(15):6638.

Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng. 2024;118:109370.