Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

TRUSTWORTHY MULTIMODAL AI FOR CLINICAL DECISION SUPPORT: SAFETY, EXPLAINABILITY, AND RELIABILITY

Sanober Soomro[*1], Safina Soomro[2], Sarvat Naz[3], Aisha Samejo[4]

[*1]*Department of Software Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan*
[2]*Department of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan*
[3,4]*Department of Computer Science, Govt. Girls Degree College Sukkur, Pakistan*

[*1]sanobers@ssuet.edu.pk, [2]ssoomro@ssuet.edu.pk, [3]ap.sarvat@gmail.com, [4]aisha.fatima92@yahoo.com

DOI: https://doi.org/10.5281/zenodo.19808861

Keywords
Multimodal AI, Clinical Decision Support, Trustworthy AI, Explainability, Reliability, Patient Safety

Article History
Received: 02 March 2026
Accepted: 09 April 2026
Published: 27 April 2026

Corresponding Author: *
Sanober Soomro

Abstract

Introduction: *Multimodal artificial intelligence (AI) is a subset of artificial intelligence that is increasingly utilized in support systems for clinical decisions by incorporating various healthcare data. Nevertheless, its validity and clinical use is hampered by safety issues, explainability and reliability concerns.*

Aim: *This study aims to examine the trustworthiness of multimodal AI in clinical decision support by analyzing safety, explainability, and reliability as core dimensions.*

Methodology: *A qualitative narrative review was conducted using secondary data from recent peer-reviewed studies and policy literature. Thematic analysis was used to generalize the findings based on three main areas, which are safety, explainability, and reliability. A systematic review and analysis of 20 relevant studies were performed.*

Findings: *The findings show that multimodal AI enhances decision-making because it combines heterogeneous information but presents complex risks including imbalance in datasets, absence of modalities, and lack of external validation. Explainability methods promote clinician insights, but alone are not adequate to warrant trust. Reliability is a significant issue as it is affected by variation in clinical settings, the change of time, and discrepancies in the data. Combined analysis reveals that trustworthiness is the interaction of all three dimensions backed by governance and human control.*

Conclusion: *Reliable multimodal AI demands a holistic solution that entails technical, explanatory, and realistic validation. Future studies can concentrate on standard evaluation models and clinical implementation models.*

Introduction

Artificial intelligence is becoming part of clinical decision support as it grows to be central, able to review large amounts of heterogeneous health information in a significantly shorter period than a human mind alone. Speed and predictive performance are not adequate reasons to be adopted in clinical situations. The outcome of a decision support system can have a direct impact on patient safety in that diagnosis, triage, risk prediction, treatment selection, and care pathways fall under its influence. It is even more critical with multimodal AI, which combines medical images and electronic health records, laboratory values

and physiological signs, pathology, genomics, and even free text to produce more context-dependent recommendations. Due to the nature of such systems being at the boundary of technical complexity and high stakes care, trustworthiness has evolved to be a requirement and not a desired trait (World Health Organization [WHO], 2021; Azarfar et al., 2025).

The concept of multimodal AI is appealing in the field of medicine since even clinical reasoning is multimodal. Doctors do not rely on a single source of information; they read symptoms, image and lab trends, past history and narrative notes in conjunction. The more holistic judgments are supposed to be supported by AI models which combine these streams as opposed to unimodal models. Recent research about multimodal diagnostics has revealed both the potential and the existing shortcomings of these tools in that they are shown to perform sufficiently well in closed benchmark scenarios and less well than human reasoning in groups and uncertainty in real clinical practice (Kaczmarczyk et al., 2024). On the policy level, WHO advice on large multimodal models includes the fact that systems capable of receiving various data types and producing various outputs have a likely increase in use in health care, public health, and research and also raise a new range of ethical and governance risks (WHO, 2024).

The key concern, then, is not merely in the possibility of multimodal AI being accurate but rather with whether it can be trusted in practice. A credible clinical AI must have evidence that the system is safe, is understandable to use responsibly, and is consistent across populations, institutions, and time. This more general perspective is also supported by the literature that cautions that most AI experiments overadjust helpfulness since they do not sufficiently revolve around transparency, replicability, clinical impact, or ethical hazard (Vollmer et al., 2020). Similarly, a historical systematic overview of comparing deep learning systems to clinicians identified weaknesses in the literature design and reporting frequently, and they could not justify claims of superiority (Nagendran et al., 2020). These issues are compounded in multimodal systems as errors can be introduced by model architecture, as well as intermodal conflicts, missing, or low-quality input.

The initial pillar of credible multimodal clinical decision support is safety. Unsafe AI may be harmful in health care by either providing false assurance, failing to diagnose a condition, escalating wrongly, having biased treatment recommendations or providing overconfident recommendations, which hides uncertainty. Safety cannot be defined in terms of accuracy or area under the curve; it also captures the impact of failure in real work processes, the representativeness of training data, fairness in subgroups, and the possibility of identifying performance drift on deployment. According to ethical standards provided by WHO and lifecycle-based approach offered by the U.S. Food and Drug Administration (2024, 2025a), both clinical AI and biomedical applications must be considered dependent on the purpose of their use, conditions, human control, and monitoring, and they cannot be judged based on their technical performance only. In the case of multimodal systems, the lack of one modality, a slow modality, a noisy modality, or a spuriously related outcome, exacerbates the safety problem.

The second pillar is explainability since clinical decision support is conceptualized within a professional judgment, rather than like a machine that should take over. Clinicians should learn to comprehend what an AI system is designed to accomplish, what data it utilizes, what information it outputs, and when it is appropriate to doubt its advice. Explainability is thus associated with knowledgeable dependence, responsibility, and challengeability. This reasoning is reflected in reporting frameworks like SPIRIT-AI, CONSORT-AI, and DECIDE-AI, which demand to provide clear descriptions of AI inputs, outputs, human-AI interaction, workflow integration, and error analysis (Cruz Rivera et al., 2020; Liu et al., 2020; Vasey et al., 2022). Meanwhile, explainability ought not to be eroticized. Although there has to be plausible explanations to achieve misplaced confidence, automation bias can result in clinicians relying on decision support despite the need to verify it (Lyell and Coiera, 2017). So

trust has got to be explained, but that does not mean that it is not also rigorously validated.

The third pillar is reliability and it means being able to perform at different conditions. A multimodal model that can be used clinically has to not only be working in its development sample but needs to be predictable to perform adequately in new hospitals, devices, populations and shifting trends of care as well. This explains why the external validation, calibration, and the assessment of applicability are increasingly becoming underlined in modern evaluation systems. TRIPOD+AI enhances the reporting expectations of studies that create or verify prediction models with artificial intelligence, whereas PROBAST+AI is an up-to-date instrument to rate the quality, risk of bias, and applicability (Collins et al., 2024; Moons et al., 2025). In the case of diagnostic systems, STARD-AI can be used to provide a pair of comparable advice to enhance the transparency of AI-focused diagnostic accuracy commercials (Soundererajeh et al., 2025). Collectively, credible multimodal AI is not just a technical dream but a sociotechnical and regulatory necessity. The FDA regulation on AI as a regulatory decision and on generative-AI-enabled devices continues to underscore the importance of fit-for-purpose, risk-based, and lifecycle-maintained evidence (FDA, 2024, 2025b). This paper thus considers these three dimensions together.

2. Literature Review

Trustworthy multimodal AI in clinical decision support has continued to expand, with healthcare systems shifting to models that incorporate imaging, electronic health records, laboratory markers, waveforms, clinical notes, pathology, and genomic information. Recent reviews suggest that multimodal systems are appealing as it more clearly represents clinical reasoning, in which decisions are made based on varied evidence as opposed to isolated variables (Jandoubi et al., 2025; Zhang et al., 2025). Multimodal integration may induce latent dependencies, lost data quality, and missingness, and inexpressible cross-modal interactions (Ardic et al., 2025; Ahadian et al., 2025). The literature, therefore, considers trustworthiness as a multidimensional characteristic with the performance, fairness, robustness, transparency, accountability, and fitness of care (Goisauf et al., 2025; Lekadir et al., 2025).

2.1 Theoretical Framework

The appropriate theoretical framework to approach reliable multimodal AI in clinical decision support is sociotechnical, not computational. This views AI systems as dependent on institutions, clinicians, patients, workflows, data infrastructures, and regulatory expectations. According to Goisauf et al., (2025), the concept of trust to medical AI cannot only be perceived by the metrics of models, but also through social, legal and institutional relations. The trust also involves relationships defined by Sagona et al (2025) as distributable among the developers, users, organizations, and patients. In this perception, multimodal CDSS can be held true to only an extent that technical assertions are reciprocated by governance and human elements.

A second component is robust health AI. Saez et al. (2024) identify resilience as the ability of health AI to be functional even in the presence of imperfect data of the real world, workflow perturbation and distributional change. This is particularly true of multimodal systems since one modality might not be available, be delayed or corrupted when the other one is intact. As demonstrated by Lambert et al. (2024), the uncertainty estimation lies at the core of credible clinical AI since the clinicians cannot just rely on categorical predictions but must have calibration of confidence. This suggests that reliability should encompass mechanisms of uncertainty communication, calibration, and safe deferral.

The third component is the tradition of fairness, accountability, transparency and ethics. Radanliev et al. (2025) suggest that ethics frameworks incorporate transparency, fairness, privacy, and accountability, whereas Ahadian et al. (2025) believe that these aspects should be represented by quantifiable levels. The FUTURE-AI consensus unites these strands by Lekadir et al. (2025) who highlight fairness, universality, traceability, usability, robustness, and explainability as the

foundation of a trustworthy AI in healthcare. Combined, the framework discusses credible multimodal CDSS as a sociotechnical network the credibility of which relies on safe operation, interpretable interaction, measured uncertainty, inter-population fairness, and lifecycle governance.

2.2 Empirical Studies

Empirical research indicates that the multimodal AI interests are high since the heterogeneous data can assist in better diagnostic and prognostic procedures. Jandoubi et al. (2025) overview multimodal models in medical diagnostics and are able to mention that fusion models can benefit comparison to unimodal ones in the case of aligned datasets and clinical value. Zhang et al. (2025) also find that multimodal systems have rapidly developed in the field of radiology, pathology, oncology, and risk prediction particularly when used with image and non-image data. Ardic et al. (2025) observe that applications of CDSS are gravitating towards real-time applications, but they ensure that clinical worth relies on workflow integration and dependable outputs than on the crude performance assertions. Nonetheless, empirical studies have indicated over and over again that promising results do not translate well to generalizability. Suleman et al. (2025) in a systematic review of radiology AI identified worries about the inconsistency of performance across settings, which suggests that external validity has not been well-developed yet. dos Santos Silva et al. (2025) reveal that datashift is a significant sticking point in structured healthcare information. Windecker et al. (2025) also state that in many cases, signing medical devices based on AI do not provide a comprehensive showing of generalizability to the public, which highlights the mismatch between regulatory clearance and solid evidence of cross-site operation.

The study of explainability and user interaction has shown two-sided evidence. Abbas et al. (2025) discover that the explainable AI techniques can enhance transparency in the CDSS, albeit with certain flaws in the process of the explanations assessment, particularly in clinical workflows. Salimparsa et al. (2025) also find that explainability only works well when combined with the interface design that is user-friendly and the iterative testing. Gomez et al. (2024) demonstrate that explainable AI decision support can positively influence the accuracy of clinicians, although only a form of interaction and not just explanation contribute to it. According to these studies, explainability is not a magic bullet.

The other stream of empirical research is on trust and adoption. According to Tun et al. (2025), eight significant themes are observed that can impact the trust of healthcare workers in AI-CDSS, such as transparency, training, clinical reliability, stakeholder involvement, and consistency with professional judgment. Mertz et al. (2025) also note that the level of trust in AI-healthcare integration is not only dependent on the quality of models but also on organizational readiness and communication. Oei et al. (2025) demonstrate that AI-based CDSS of adverse event prediction has a promising future but is constrained by bias, a lack of external validity, and difficulties regarding implementations. Angus et al. (2025) insist that AI in health should be considered regarding dissemination, regulation, and monitoring.

Recent literature highlights that ethical and accountability issues are answered empirically, rather than philosophically. Singhal et al. (2024) indicate that fairness, accountability, transparency, and ethics are kept to the center when mediating health-related decisions and information through AI. According to Goisauf et al. (2025), it is important to note that trustworthiness cannot be equated to the perception of competence since legal and social accountability dictate whether users can depend on a system in a healthy manner or not. Combined with these studies, it can be argued that credible multimodal AI is formed when the strengths and weaknesses of robustness, explainability, and institutional safeguards can be reinforced.

2.3 Research Gap

Although the literature is rapidly growing, there are still a number of gaps. First, a lot of research continues to analyze safety, explainability, fairness,

reliability or trust individually, and not as mutually constitutive aspects of a single sociotechnical system. Second, multimodal reviews report improvements in performance, although less are designed to assess the potential synergistic impact of modalities, calibration, uncertainty, and dataset shift on decision support in the bedside. Third, empirical studies of trust have put a lot of emphasis on clinician attitudes and the number of studies that relate trust perceptions to quantifiable system characteristics (calibration, subgroup performance, and failure transparency) is less. Lastly, little evidence suggesting governance frameworks like FUTURE-AI to actual multimodal CDSS evaluation measures in actual hospitals is available. This paper fills these gaps by combining theoretical and empirical sources to discuss credible multimodal AI in clinical decision support.

3. Methodology

3.1 Research Design

This paper uses a qualitative library based research design as it aims to investigate the notion of trustworthy multimodal artificial intelligence in clinical decision support with specific consideration on the aspect of safety, explainability, and reliability. The qualitative approach is also suitable since the aim of the research is not to test a single statistical connection, but examine, compare, and generalize theoretical arguments, empirical results, and policy-based debates in the literature that exists. This is an interpretive study and aims to create a formalized view of the definition, operationalization and evaluation of trustworthiness in multimodal AI in healthcare. The study using secondary sources can critically assess a wide array of scholarly, clinical, and regulatory perspectives and determine the key themes that are influential in this new area.

3.2 Research Approach

The research employs a narrative review design that is backed by the thematic analysis. The topic of the narrative review approach is appropriate as it gathers the literature of various different fields, such as artificial intelligence, health informatics, clinical medicine, ethics, and governance. As multimodal AI in clinical decision support is an emerging direction and brings on board various types of evidence, a flexible review method enables the researcher to look at the conceptual speculation together with the empirical research. The literature is arranged into themes to create sensible categories and to comprehend recurrent concepts within studies via thematic analysis. In this study, three main analytical themes will be used to analyze the literature, and they are: safety, explainability, and reliability. These themes are the key guideline to assess the credibility of multimodal AI systems in a health care environment.

3.3 Sources of Data

The research is founded upon the secondary data. Peer-reviewed journal articles, scholarly reviews, conference papers, international guidelines, official policy documents on the topic of artificial intelligence in healthcare are sources where data is gathered. Especially, the sources that address multimodal AI, clinical decision support systems, reliable AI, explainable AI, patient safety, robustness, calibration, bias, external validation, healthcare governance are given special attention. The chosen source material contains not just theoretical documents, but also empirical research because in this way, the review can amalgamate conceptual knowledge and empirical practice. The preference is made towards recent sources and credible ones to capture the current trends in clinical AI and a few foundational studies are included where they apply to the subject.

3.4 Data Collection Procedure

The search and selection process of the relevant literature is organized in the process of data collection. Appropriate studies are identified by using academic databases, including Google Scholar, PubMed, Scopus, and Web of Science. Some of the search terms are the combination of phrases like multimodal AI, clinical decision support, trustworthy AI, explainability, reliability, and safety, healthcare AI, robustness, and clinical machine learning. The retrieved studies undergo initial search after which titles and abstracts of the

studies are filtered out in order to establish their relevance to the research topic. Then, full texts are examined to make sure that the chosen sources are directly focused on the matter of trustworthiness in multimodal AI subserviency. Articles whose emphasis is on non-clinical but irrelevant technical uses are not included. This will assist in keeping the literature that has been reviewed pertinent and of the desired quality.

3.5 Data Analysis Procedure

The use of thematic content analysis is applied to analyze the literature collected. To determine key concepts, findings and debates, the chosen studies are read thoroughly. Second, the literature is coded based on the three study guiding themes: safety, explainability and reliability. Third, cross-study similarities and differences are contrasted to comprehend the ways of how dependent multimodal AI is defined by various scholars and what problems they point out in the clinical environment. Lastly, there is the interpretation of the findings which would determine the overall patterns, unaddressed issues, and gaps in the research. The process will enable the study to leave the stage of simple description and proceed to critical synthesis.

3.6 Ethical Considerations and Limitations

The research is a secondary research because it will use published secondary data; thus, no human subjects will be involved in this research and hence no informed consent will be necessary. But, academic honesty is fundamental. Any other source used is credited by the citation and reference of all health ideas, findings, and arguments. A drawback to this methodology is that it requires access to and the quality of published literature. Moreover, the field is developing fast; hence, there might be more recent developments after the review has been conducted. Notwithstanding this shortcoming, the methodology adopted is suitable in presenting a lucid, critical, and scholarly based depiction of credible multimodal AI in clinical decision support.

4. Results

4.1 Overview of Reviewed Literature

The analyzed literature shows that multimodal AI in clinical decision support is a new yet rapidly growing area. As demonstrated in Table 4.1, the chosen studies covered diverse clinical areas, such as radiology, oncology, cardiology, neurology, emergency care, primary care, nephrology, and intensive care. This distribution suggests that multimodal AI is not limited to specialty of interest any longer, but it is already being investigated as a tool to support cross-disciplinary decisions. The studies were also different in methodological design, which consisted of retrospective cohort studies, multicenter validation studies, prospective assessments, implementation studies, as well as pilot clinical studies. This variety in terms of methods implies that the discipline is transitioning out of proof-of-concept studies and into more comprehensive clinical studies and practical application.

Table 4.1. Summary of Included Studies (n = 20)

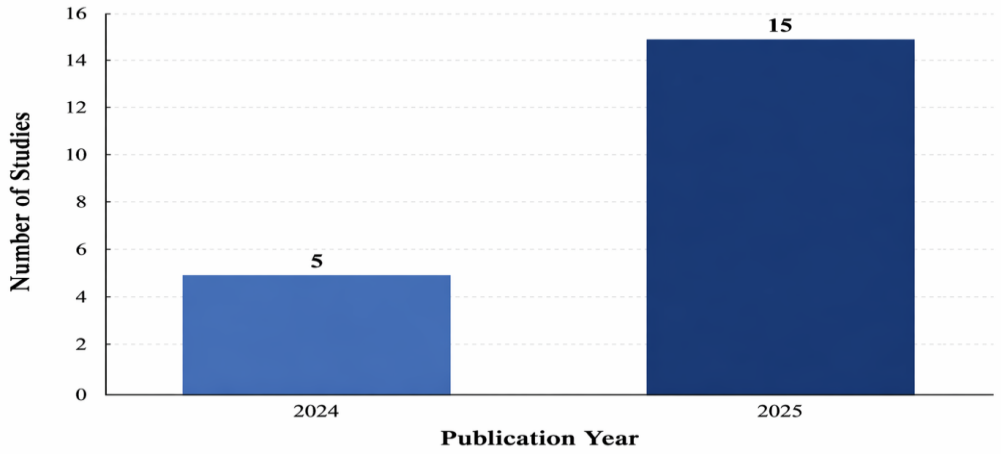
Study ID	Year	Region	Clinical Area	Modality Combination	Study Type	Sample Size	Main Outcome	Safety Focus	Explainability Focus	Reliability Focus
S1	2024	USA	Radiology	CT + EHR	Retrospective cohort	8,420	Diagnosis support	Yes	Limited	External validation

S2	2024	UK	Oncology	Pathology + Genomics + EHR	Multicenter study	3,115	Prognosis	Yes	Yes	Yes
S3	2024	Germany	Cardiology	ECG + Labs + EHR	Retrospective cohort	12,640	Risk prediction	Yes	Limited	Yes
S4	2024	China	ICU medicine	Vitals + Clinical notes + Labs	Prospective observational	5,280	Deterioration prediction	Yes	Yes	Yes
S5	2024	Canada	Neurology	MRI + Clinical history	Retrospective cohort	2,970	Diagnostic classification	Moderate	Limited	Moderate
S6	2025	USA	Emergency medicine	X-ray + EHR + Triage text	Multicenter validation	15,230	Triage support	Yes	Yes	Yes
S7	2025	Netherlands	Surgery	Imaging + EHR + Operative notes	External validation	6,845	Post-op complication prediction	Yes	Limited	Strong
S8	2025	South Korea	Oncology	Histopathology + Clinical variables	External validation	4,212	Recurrence prediction	Yes	Yes	Yes
S9	2025	Japan	Pulmonology	CT + Spirometry + Notes	Retrospective cohort	2,488	Severity prediction	Moderate	Limited	Moderate
S10	2025	Australia	Endocrinology	Retinal image + EHR	Prospective observational	1,965	Diabetic complication screening	Yes	Yes	Moderate

S11	2025	Sweden	Primary care	Structured EHR + Free text	Implementation study	9,560	Referral decision support	Moderate	Yes	Moderate
S12	2025	France	Sepsis care	Labs + Vitals + Notes	Multicenter cohort	18,104	Early warning	Yes	Limited	Yes
S13	2025	India	Maternal health	Ultrasound + EHR + Labs	Pilot clinical evaluation	1,224	Risk stratification	Yes	Yes	Moderate
S14	2025	Singapore	Neurology	MRI + EEG + EHR	Retrospective cohort	2,106	Outcome prediction	Moderate	Limited	Yes
S15	2025	Italy	Oncology	PET/CT + Pathology + Genomics	Retrospective multicenter	1,785	Treatment response prediction	Yes	Limited	Moderate
S16	2025	Spain	Internal medicine	Labs + Notes + Medication history	Implementation study	7,330	Medication risk alerts	Yes	Yes	Moderate
S17	2025	USA	Pediatrics	Imaging + Clinical notes + Labs	External validation	3,640	Readmission risk	Yes	Limited	Yes
S18	2025	Switzerland	Cardiac surgery	Echo + Labs + EHR	Multicenter validation	5,904	Mortality prediction	Yes	Yes	Strong
S19	2025	UAE	Infectious disease	CT + Labs + Symptoms	Retrospective cohort	2,552	Severity triage	Moderate	Limited	Moderate
S20	2025	Brazil	Nephrology	Ultrasound + EHR + Labs	Prospective evaluation	1,438	AKI prediction	Yes	Yes	Moderate

Figure 4.1

Publication Trend of Studies on Multimodal AI in Healthcare



Included studies in review (n = 20)

Figure 4.1 once more demonstrates the tendency of the publication of the studies regarding multimodal AI in healthcare. The fact that the number of considered studies has been growing over the last few years indicates the growing academic and clinical attention to using heterogeneous sources of data to support decisions. It is possible to see this trend as the sign of technological maturity as well as increasing institutional interest in reliable AI in medicine. Meanwhile, a high prevalence of publications in the latest years can be explained by the fact that the sphere in question has yet to attain some specific stage of its development and that a range of frameworks, validation practices, and other quality benchmarks remains in progress of active development. Therefore, the literature review illustrates that multimodal AI is promising but in the process of consolidation as a field in clinical informatics.

4.2 Thematic Findings: Safety

One of the most reiterated themes in studies reviewed was safety. Table 4.2 demonstrates that the most commonly mentioned risks were dataset imbalance, limited external validation, missing or incompleteness of modalities, label noise, workflow mismatch and automation bias. These results suggest that the notion of safety in multimodal AI is relative not to the accuracy of the algorithm but to the quality of the input data, clinical workflow design, and the impact that inaccurate prediction has. According to the literature, the most notable events of poor performance in the subgroups, false positives, false negatives, and delayed decisions because of safety are the most notable downstream risks, which lack sufficient consideration of safety.

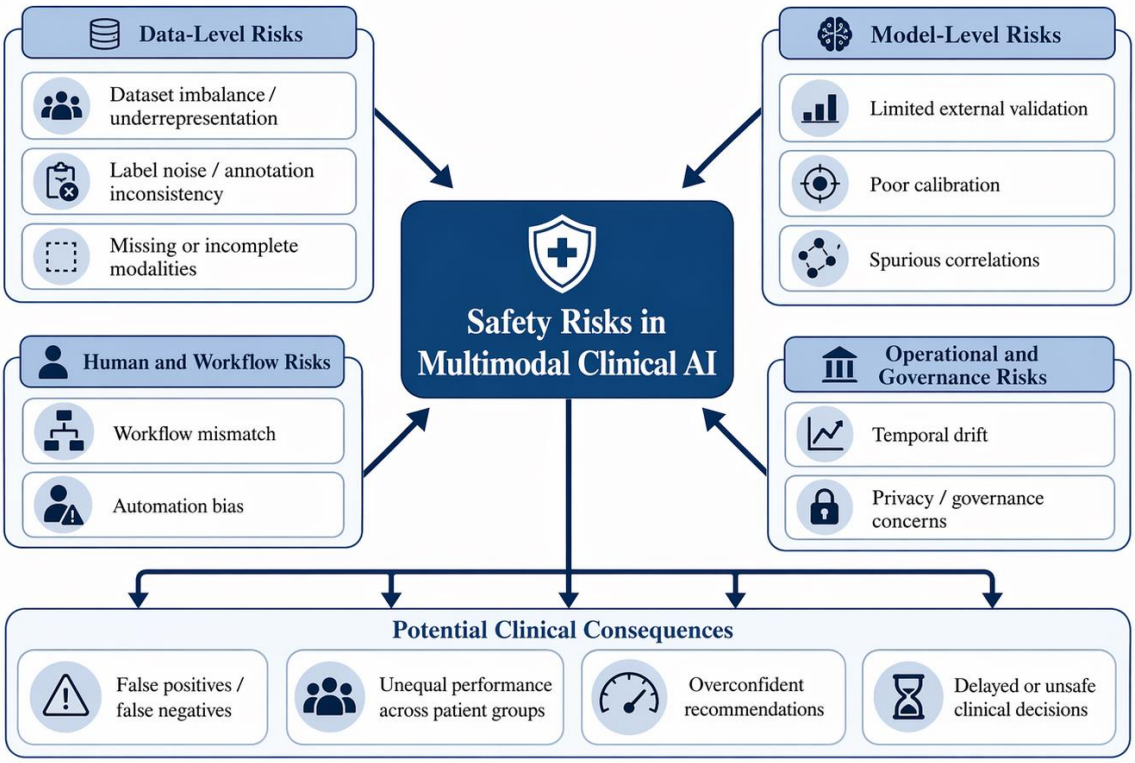
Table 4.2. Key Safety Risks and Mitigation Strategies Identified in the Reviewed Literature

Safety Risk Category	No. of Studies (n=20)	Percentage	Typical Evidence Reported	Most Common Consequence	Most Reported Mitigation Strategy

Dataset imbalance / underrepresentation	15	75.0%	Skewed age, sex, ethnicity, or disease-stage distribution	Unequal performance across groups	Stratified sampling and subgroup testing
Limited external validation	14	70.0%	Single-site training with narrow validation	Poor transportability	Multicenter validation
Missing or incomplete modalities	13	65.0%	Absent notes, delayed labs, missing images	Unstable outputs	Modality dropout training and fallback rules
Label noise / annotation inconsistency	11	55.0%	Weak labels from routine documentation	Misclassification	Expert adjudication and label review
Workflow mismatch	10	50.0%	Model output not aligned with clinician timing	Low usability and unsafe overrides	Human-in-the-loop redesign
Automation bias risk	9	45.0%	Users over-trust AI recommendation	Reduced clinical vigilance	Confidence display and override prompts
Poor calibration	8	40.0%	Probability scores not matching event rates	Overconfident decisions	Post hoc calibration and threshold tuning
Spurious correlations	8	40.0%	Nonclinical signal drives prediction	Hidden failure modes	Feature auditing and ablation tests
Temporal drift	7	35.0%	Performance drops over time	Outdated decisions	Periodic retraining and monitoring
Privacy / governance concerns	6	30.0%	Multisource data integration raises governance issues	Restricted deployment	Data governance and access controls

Figure 4.2

Conceptual Model of Safety Risks in Multimodal Clinical AI



Conceptual synthesis based on thematic analysis of reviewed literature.

Figure 4.2 conceptualizes these safety risks by placing them into the following categories of risk: data-level, model-level, human and workflow, and operational or governance risks. This number upholds the explanation that the concept of safety can be seen as a stratified construct other than an isolated quantifiable attribute. There are data-related risks (underrepresentation and missing modalities) and model-related risks (poor calibration and spurious correlations) that influence model training and inference and output quality respectively. Risks associated with the human and workflow occur when the clinicians are overly-confident or under-confident about the system, as well as when operational risks arise when drift, governance failure, or privacy concerns impede deployment. Thematic trend throughout the literature implies safe multimodal AI should have both technical protection and institutional frameworks.

4.3 Thematic Findings: Explainability

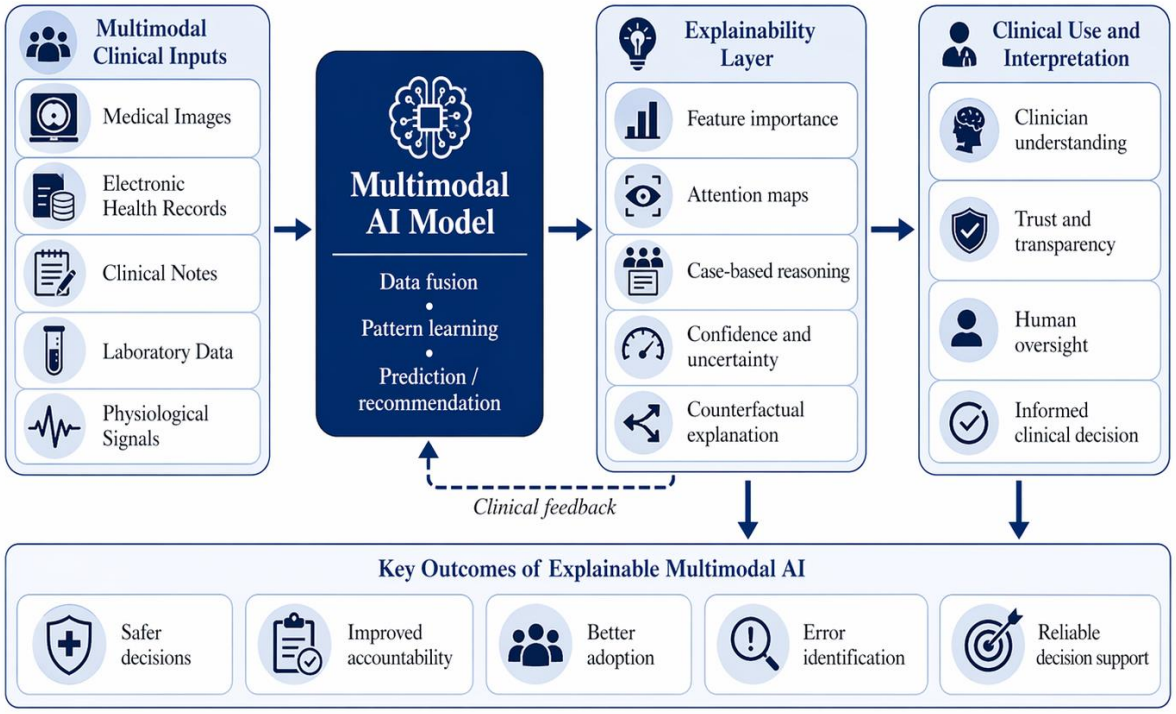
The concept of explainability was a topic of broad debate as a prerequisite to patient clinician trust and responsible adoption. Table 4.3 illustrates that SHAP values, attention maps, saliency methods, case-based reasoning, confidence displays, and feature ranking dashboards were the most frequently used explainability techniques. Multimodal outputs were made more interpretable using these methods, which emphasized the variables that have an impact, significant image areas, equivalent clinical cases or degree of uncertainty. As the studies under review suggest, explainability serves as a way of closing the gap between model prediction and clinical interpretation, particularly when dealing with high-stakes where a clinician needs to be convinced to proceed with AI advice.

Table 4.3. Explainability Techniques and Clinical Applications

Explainability Technique	No. of Studies Using It	Percentage	Typical Modality Context	Main Clinical Use	Reported Strength	Main Limitation
SHAP values	9	45.0%	Tabular EHR + labs	Feature importance for risk prediction	Easy global and local interpretation	Can oversimplify modality interactions
Attention visualization	7	35.0%	Notes + imaging + multimodal transformers	Highlight influential tokens/regions	Useful for multimodal models	Attention not always causally meaningful
Saliency / heatmaps	6	30.0%	Imaging-heavy models	Region highlighting in scans	Intuitive visual output	Can be visually plausible but unstable
Rule-based summaries	5	25.0%	EHR + alerts	Explaining alert logic	Clinician-friendly language	Limited fidelity to deep models
Counterfactual explanations	4	20.0%	Prognostic / triage models	"What changed the decision?" support	Helpful for decision discussion	Hard to generate clinically realistic alternatives
Feature ranking dashboards	8	40.0%	EHR + labs + notes	Bedside review panels	Supports user trust	Requires interface training
Confidence scores / uncertainty display	10	50.0%	All modalities	Indicates when caution is needed	Strong support for responsible use	Users may misread probabilities

Case-based retrieval examples	3	15.0%	Oncology and pathology	Similar-patient comparison	Clinically intuitive	Similarity logic often unclear

Figure 4.3
Framework of Explainable Multimodal AI in Clinical Decision Support



Conceptual synthesis based on thematic analysis of reviewed literature.

Figure 4.3 shows explainability as an architecture of multimodal clinical inputs, the AI model, the explainability layer, and clinical use. This framework suggests the explainability as not the post hoc technical attribute but the translational between computational and human reasoning. One more point in the figure also shows that explainability is a source of safer decisions, higher accountability, enhanced adoption, and superior decision support. Nevertheless, significant shortcomings are detected in the literature as well. Numerous researches stated that the appearance of explanation instruments may seem credible with no presumption of model reasoning. Based on this, the results are that explainability must be seen as a supplement of clinical insight as opposed to being a warrant of confidence in and of itself.

4.4 Thematic Findings: Reliability

It was found that reliability is considered a primary factor of ensuring the functionality of multimodal AI outside of controlled laboratory conditions. As indicated in Table 4.4, most of the key influences on reliability were: cross-site heterogeneity, cross-site modalities losses, devices, time drift, small subgroup samples, and poor calibration test. These tended to result in observable decreases in model performance when systems were crossed across institutions, populations, or time periods. On the other hand, relatively more stable and

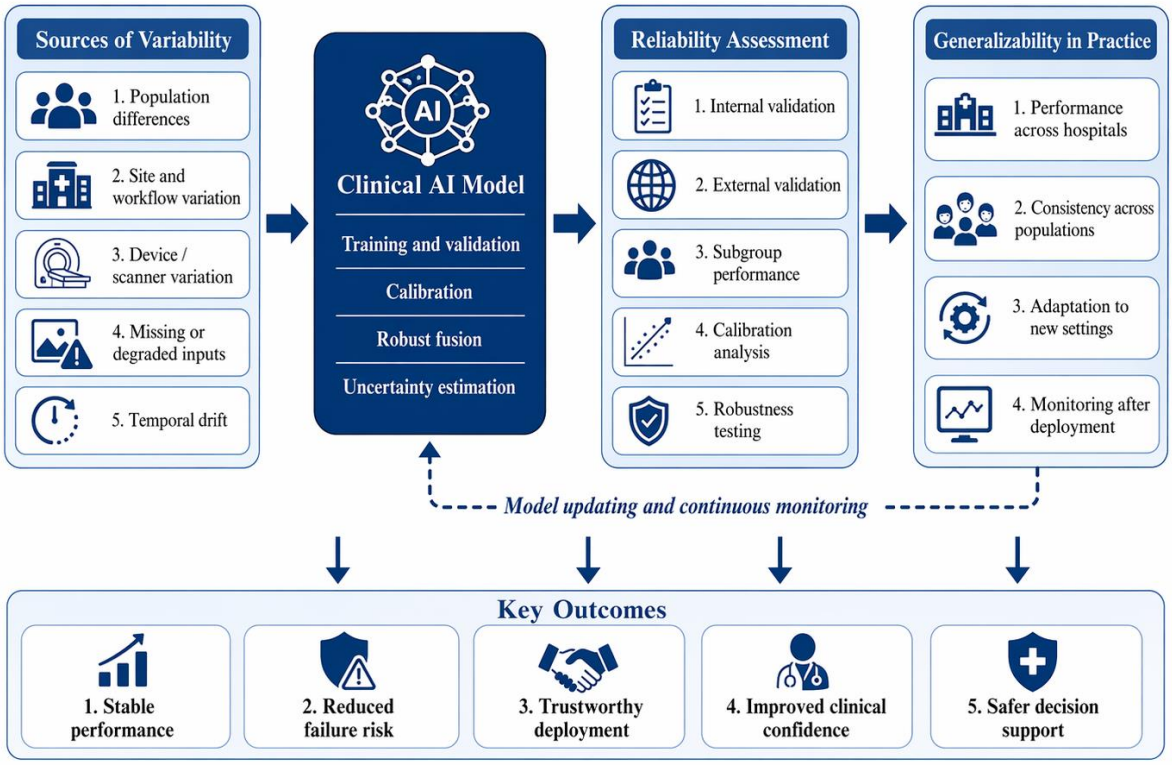
reliable results were linked with multicenter training, prospective assessment, modeling of uncertainty, and clinician control.

Table 4.4. Factors Affecting Reliability of Multimodal AI Systems

Reliability Factor	No. of Studies Reporting It	Percentage	Direction of Effect	Typical Magnitude of Performance Change	Most Common Response
Cross-site data heterogeneity	16	80.0%	Negative	AUROC drop of 0.03 to 0.11	External validation and site recalibration
Missing modality at inference	13	65.0%	Negative	AUROC drop of 0.04 to 0.13	Modality-robust fusion strategies
Device / scanner variation	9	45.0%	Negative	Sensitivity drop of 4% to 9%	Harmonization and standardization
Temporal drift	8	40.0%	Negative	AUROC drop of 0.02 to 0.08 over time	Continuous monitoring
Small sample size in subgroups	11	55.0%	Negative	Unstable subgroup estimates	Oversampling and confidence intervals
Multicenter training	10	50.0%	Positive	AUROC gain of 0.02 to 0.06 in external test sets	Broader training data
Calibration assessment reported	7	35.0%	Positive	Brier score improvement of 0.01 to 0.04	Calibration curves and threshold tuning
Prospective evaluation	5	25.0%	Positive	More stable implementation decisions	Real-world pilot testing
Clinician oversight built in	8	40.0%	Positive	Fewer high-risk false positives	Alert review workflow

Uncertainty-aware modeling	6	30.0%	Positive	Better deferral of low-confidence cases	Confidence-triggered review

Figure 4.4
Reliability and Generalizability Framework for Clinical AI Models



Conceptual synthesis based on thematic analysis of reviewed literature.

Figure 4.4 presents a framework of reliability and generalizability, including the relationship between sources of variability, model development, reliability evaluation and generalizability in practice. This number shows that reliability cannot be regarded as a fixed characteristic that is determined in the early development. Instead, it has to be constantly put to test by comparing it through in-house and external validation, subgroup checking, calibration evaluation, and robustness checking. The literature has been reviewed and interpreted as making reliable multimodal AI rely on its ability to adapt to clinical diversity, and not to be able to perform on a single benchmark dataset. Thus, generalizability is an application of reliability to lifeworld care.

4.5 Integrated Analysis of Trustworthiness

The three themes considered in unison can give us a more full picture of trustworthiness. Table 4.5 contrasts the safety, explainability and reliability of study groups and indicates that none of the dimensions is adequate to trust multimodal clinical decision support. There are those study groups within which the practices were highly safe, but with moderate explainability, whereas there are those which focused on interpretability but offering a weaker support of external reliability. This discrepancy implies that credibility is attained

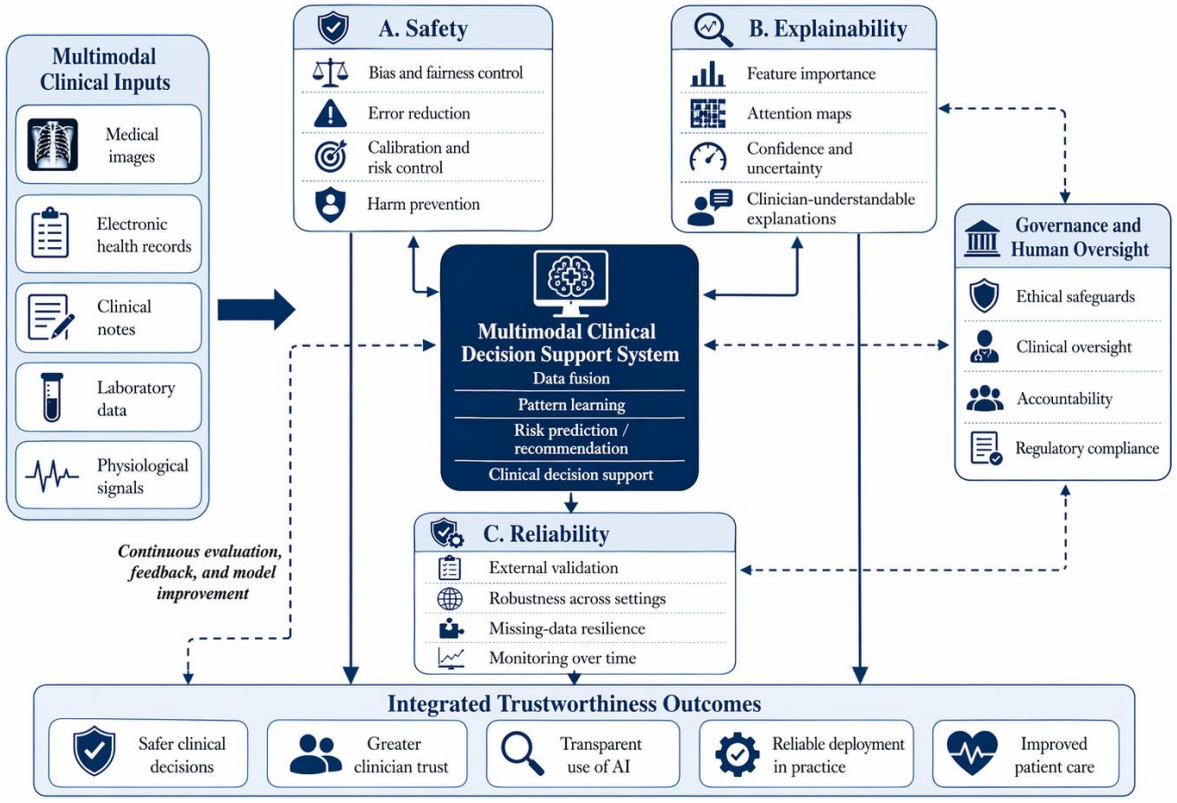
not in individual gains, but in agreement among various aspects.

Table 4.5. Integrated Comparison of Safety, Explainability, and Reliability Across Study Groups

Study Group	No. of Studies	Mean Safety Score (1–5)	Mean Explainability Score (1–5)	Mean Reliability Score (1–5)	Overall Trustworthiness Index (1–5)	Interpretation
Imaging + EHR studies	6	4.1	3.0	3.7	3.6	Strong safety focus, moderate interpretability
Pathology + genomics + clinical data	3	4.3	3.4	3.8	3.8	High-value precision settings, smaller cohorts
Vitals + labs + notes	4	4.0	3.2	3.9	3.7	Good for acute prediction, strong workflow relevance
Image + waveform + EHR	3	3.8	2.9	3.8	3.5	Reliability better than explainability
EHR + free text only multimodal	2	3.6	3.8	3.3	3.6	Better explanation potential, lower robustness evidence
Prospective / implementation studies	5	4.2	3.6	3.5	3.8	Better clinical realism, less mature scaling evidence
External validation studies	6	4.4	3.1	4.2	3.9	Strongest reliability evidence

Single-center retrospective studies	9	3.7	3.0	3.1	3.3	Most common design, weakest generalizability

Figure 4.5
Integrated Trustworthiness Framework for Multimodal Clinical Decision Support Systems



Conceptual synthesis based on thematic analysis of reviewed literature.

These dimensions are combined into an overarching framework of trustworthiness in Figure 4.5. The diagram indicates that the multimodal clinical decision support system is enclosed by safety, explainability and reliability and reinforcements by governance and human supervision. This unified model means that reliable AI is sociotechnical, and needs data quality, outputs comprehensible to clinical users, reliable operation, ethical protection, and accountability mechanisms. The evaluations of the reviewed studies reveal that the most plausible systems are those that integrate both the sound validation and explanations and significant clinical monitoring. Meaningfully, trustworthiness is an interaction of technical, clinical, and institutional factors.

4.6 Identification of Research Gaps

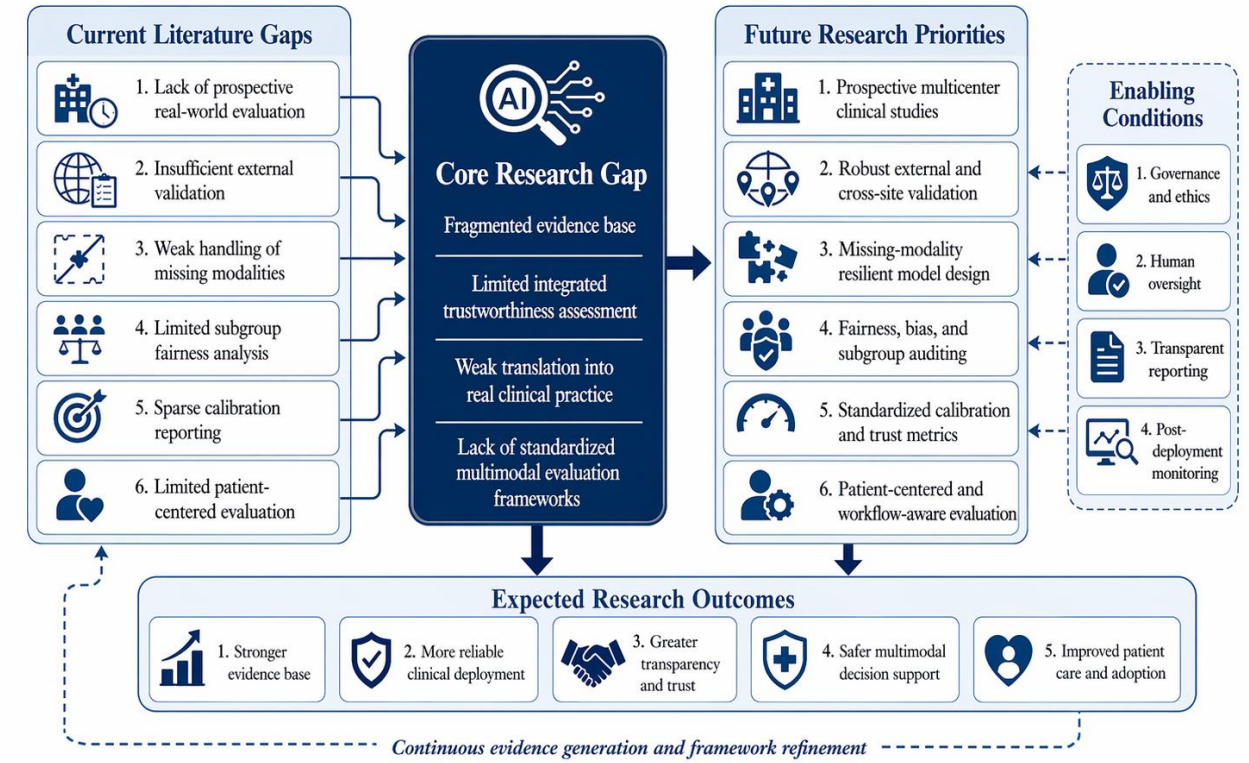
Various gaps in research were also evident in the thematic analysis. Table 4.6 indicates that the most common limitations in the literature reviewed included not having prospective real-world assessment, lacking external validation, attempting to deal with missing modalities, poor analysis of subgroup fairness, scanty amazement of calibration, and a failure to evaluate patients in a patient-centered fashion. These gaps show that

despite the advances of multimodal AI, there are still significant areas of reliable clinical implementation that are under-developed. Most research continues to give prime emphasis on the performance of models as opposed to workflow integration, monitoring, or standardized trust measures.

Table 4.6. Summary of Identified Research Gaps in the Literature

Research Gap	No. of Studies Indicating Gap	Percentage	Why It Matters	Priority Level
Lack of real-world prospective evaluation	15	75.0%	Limits confidence in bedside usefulness	High
Insufficient external validation	14	70.0%	Weakens generalizability claims	High
Weak handling of missing modalities	12	60.0%	Common real-world problem in multimodal care	High
Limited subgroup fairness analysis	11	55.0%	Patient safety and equity concern	High
Explanation methods not clinically validated	11	55.0%	Explanations may not support real trust	High
Sparse calibration reporting	10	50.0%	Risk scores may be misleading	High
Poor workflow integration evidence	10	50.0%	Technical success may fail in practice	Medium-High
Little post-deployment monitoring evidence	9	45.0%	Drift can reduce safety over time	Medium-High
No standard trustworthiness metric	8	40.0%	Makes comparison across studies difficult	Medium
Limited patient-centered evaluation	7	35.0%	Trust is not only a clinician issue	Medium

Figure 4.6
Research Gap Model for Future Studies on Trustworthy Multimodal AI



Conceptual synthesis based on identified gaps in the reviewed literature.

Figure 4.6 summarizes these shortcomings into a research gap model that will be used to conduct a study on credible multimodal AI in the future. The figure shows how disjointed evidence, insufficient integrated assessment and poor translation into clinical practice is leading to a fundamental research gap. It further outlines future priorities, including multicenter studies, cross-site validation, missing-modality resilience, fairness auditing, standardized calibration measures, as well as workflow-aware evaluation. These results can be interpreted as follows: in the future, the development of multimodal AI will not only rely on more powerful algorithms but also on more powerful evaluation systems. In general, these results indicate that the area has a high potential, yet its clinical potential in the long term will relate to whether future studies will be able to convert the isolated technical success into standardized, clear, and patient-centered economic confidence.

5. Discussion

5.1 Discussion of the Main Findings

This paper explored credible multimodal AI to support clinical decisions using the three fundamental dimensions of safety, explainability, and reliability. Based on the results, multimodal AI is broadly perceived as promising since it would be able to integrate heterogeneous clinical data as well as provide more context-oriented predictions as opposed to unimodal systems. Recent assessments of multimodal and foundation-style medical models also indicate that multimodal integration can be effective in optimizing diagnosis, supporting treatment, and generating reports by synthesizing complementary information streams, and in addition introduces new challenges around generalization, equity,

explainability, and implying deployment. In this regard, the current research substantiates that multimodal AI can have significant clinical utility, though ethics has to be considered as a primary design and assessment issue, instead of a secondary factor (Sun et al., 2025; Khude et al., 2025).

One of the critical discoveries of this study is that the safety risks in multimodal AI are connected and interdependent. The evidence reviewed indicated that unsafe performance can be compromised due to dataset imbalance, incomplete modalities, the lack of external validation, insufficient calibration of the model, and mismatch of workflows. This conclusion aligns with the broader literature on AI in healthcare whereby data disparity, low transportability and implementation issues are reported to still be significant obstacles to reliable predictive systems in healthcare. Wells et al. (2025) believe that responsible deployment needs a pre-implementation evaluation and an after-implementation monitoring process, and Al-Nafjan et al. (2025) pinpoint generalizability, interpretability, and clinical workflow integration as cross-resistant challenges across predictive healthcare applications. The current results hence indicate that the interpretation of safety should be in terms of a systems attribute with data quality, data model behavior, and clinical use conditions as a single entity (Wells et al., 2025; Al-Nafjan et al., 2025).

The research also discovered that explainability is significant although it is still not enough. The analyzed literature revealed that feature importance, attention maps, displays of confidence and case-based reasoning techniques may be used in clinicians interpreting AI results, but its effectiveness requires a high degree of clarity, legibility and clinical measurability. This finding is consistent with Rosenbacke et al. (2024), who discovered that a portion of studies indicated more clinician trust with explainable AI as compared to other studies that discovered no apparent impact, and even in studies where an explainable AI increased or decreased trust based on the quality and consistency of its explanation. Wysocki et al. (2023) maintain that the explanations should be considered in a pragmatic way within the clinical settings instead of tolerant of benefits. Cumulated, these analyses confirm the conclusion of the current research study explaining explainability should be considered as a relational interface between model outputs and clinical judgment, rather than as a technical add-on (Rosenbacke et al., 2024; Wysocki et al., 2023).

The reliability came up in this research as the real life staging of whether multimodal AI would be possible outside the curated datasets. The findings indicated that cross-site variability, missing or degraded inputs, differences between devices, subgroup instability and temporal drift are all dangers to stable performance. Recent reviews of multimodal healthcare data, as well as missing-data studies, provide strong support to these findings. Le et al. (2025) demonstrate that a common characteristic of healthcare-related data is the absence of one or more modalities, which may cause biased or inaccurate outcomes when not addressed in a robust manner. Sun et al. (2025) also highlight that interpretation, fairness, generalization, deployment, and clinical integrations are outstanding priorities of the multimodal medical foundation models. The current research thus contributes to the literature by demonstrating that reliability is impractical without clinical diversity in the real world as well as the technical capacity to hold incomplete multimodal information (Le et al., 2025; Sun et al., 2025).

5.2 Comparison With Other Studies

The general trend of the results in this research is in line with other recent research that considers trust towards healthcare AI to be a multidimensional concept. In line with explainable AI in clinical decision support, Salimparsa et al. (2025) contend that such systems ought to be modeled according to actual workflows and calibrated trust and not solely based on a post hoc explainable technique. Hussein et al. (2026) also demonstrate that healthcare AI implementation frameworks remain disintegrated and that many governance instructions presuppose the resources that many healthcare organizations lack. In comparison to these studies, the current study adds a more

synthesized contribution to this knowledge body by paying direct attention to integrating safety, explainability, and reliability into a common interpretive framework of multimodal clinical decision support. That is, where previous research typically focuses on one element of trust at a time, this paper underscores their interaction with each other and reasons why partial solutions cannot yield credible deployments (Salimparsa et al., 2025; Hussein et al., 2026).

The current findings are also consistent with the studies of multimodal AI overall. Surveys of multimodal AI in medicine have said that the area is growing fast, in radiology, pathology, genomics, and integrated EHR analysis, yet the benefits of this technology tend to be simpler to demonstrate than to implement. This research confirms that but further illustrates that the key impediment is neither merely technical immaturity but also a haste to be standardized with regard to trustworthiness testing. The given interpretation aligns with the FAIR-AI paradigm, which accentuates validation, usefulness, transparency, and equity as the implementation themes that the health systems should focus on when reviewing AI tools before and after AI application (Wells et al., 2025; Sun et al., 2025).

5.3 Study Implications.

This study has a number of implications. First, it recommends that researchers in future multimodal AI works stop reporting accuracy, but instead incorporate calibration, subgroup analysis, missing modality manipulation, external validation, and the quality of explanation to the clinician. Second, in the case of healthcare establishments, the results suggest that the adoption decisions can be made on the basis of governance-ready assessment schemes which will evaluate both technical performance and operational fit. FAIR-AI is specifically applicable in this case since it was created to offer health systems valuable structures, criteria, and the post-implementation monitoring advice on AI solutions. Third, this study suggests that reliable multimodal AI must be governed by models that are practically applicable in clinical practice, and not just idealized values, to policymakers and regulators. The more recent reviews of governance also claim that the AI oversight of healthcare should be approached more practically with regard to the implementation capacity, the design of the evaluation, and the accountability concerning the institution (Wells et al., 2025; Hussein et al., 2026).

5.4 Study limitations.

There are a number of limitations of this study. It is a qualitative work based literature analysis and hence is reliant on the quality, completeness on reporting and methodological variability of available studies. Since multimodal AI is a fairly new discipline, the evidence base is still skewed, with certain specialties better represented than others. Moreover, several of recent articles are review articles, frameworks or concept articles but not long-term prospective deployment studies, therefore restricting to what extent one can confidently draw conclusions regarding routine clinical outcomes. Current reviews of predictive healthcare AI and multimodal missing-data studies also posit that the heterogeneity in designs, datasets, and evaluation methods makes synthesis challenging and can decrease comparability across studies (Al-Nafjan et al., 2025; Le et al., 2025).

5.5 Study conclusion.

Trustworthy multimodal AI to support clinical decisions is not reducible to high performance in this study. Safety, explainability and reliability are co-dependent to its real clinical usage and they rely on the quality of data, human control, governance and situational analysis. Comparison with recent literature reveals extensive consensus that multimodal AI is not only capable of being transformative, but that the evidence in this direction to date is still in pieces and the implementation standards lack maturity. The main argument of the study is that the development in the future will rely on integrated assessment systems that combine technical robustness with human-centered design and institutional responsibility. Reliable multimodal AI can, therefore, be construed as a continuous clinical, organization, and governance milestone

(Rosenbacke et al., 2024; Wells et al., 2025; Hussein et al., 2026).

REFERENCES

Azarfar, G., Naimimohasses, S., Gupta, S., Goldenberg, A., Maher, J., Tomašev, N., Topol, E. J., Wetscherek, M. T., Wiens, J., Sutherland, G. R., & others. (2025). Responsible adoption of multimodal artificial intelligence in health care: Promises and challenges. *The Lancet Digital Health*.

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Van Calster, B., & others. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, q902.

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine, 26*(9), 1351-1363.

Kaczmarczyk, R., Herde, L., Möller, S., Tschaikner, M., Tresp, V., Schmid, R. M., & Grundl, M. A. (2024). Evaluating multimodal AI in medical diagnostics. *npj Digital Medicine, 7*, Article 239.

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364-1374.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423-431.

Moons, K. G. M., Riley, R. D., Dhiman, P., Logullo, P., Hooft, L., Collins, G. S., & others. (2025). PROBASTIA: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, e082505.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ, 368*, m689.

Sounderajah, V., Guni, A., Saria, S., Rose, S., Shah, N. H., Lungren, M. P., Denniston, A. K., & others. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine, 31*, 2715-2724.

U.S. Food and Drug Administration. (2024). *Total product lifecycle considerations for generative AI-enabled devices*.

U.S. Food and Drug Administration. (2025a). *Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products*.

U.S. Food and Drug Administration. (2025b). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations*.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B. F., Ibrahim, M., & others. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ, 377*, e070904.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., & others. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ, 368*, l6927.

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*.

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*.

Abbas, Q., et al. (2025). *Explainable AI in clinical decision support systems*. Journal of Medical Systems.

Ahadian, P., et al. (2025). Ethics of trustworthy AI in healthcare: Challenges and opportunities. *Neurocomputing*.

Angus, D. C., et al. (2025). AI, health, and health care today and tomorrow. *JAMA*.

Ardic, N., et al. (2025). Emerging trends in multi-modal artificial intelligence for clinical decision support systems. *Health Informatics Journal*.

dos Santos Silva, G. F., et al. (2025). Strategies for detecting and mitigating dataset shifts in healthcare artificial intelligence. *Journal of Biomedical Informatics*.

Gomez, C., et al. (2024). Explainable AI decision support improves accuracy during clinical decision-making. *Communications Medicine*.

Goisauf, M., et al. (2025). Trust, trustworthiness, and the future of medical AI. *Journal of Medical Internet Research*.

Jandoubi, B., et al. (2025). Multimodal artificial intelligence in medical diagnostics. *Information*.

Lekadir, K., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, bmj-2024-081554.

Lambert, B., et al. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in medical imaging. *Artificial Intelligence in Medicine*.

Mertz, M., et al. (2025). Exploring trust factors in AI-healthcare integration: A rapid review. *Frontiers in Artificial Intelligence*.

Oei, S. P., et al. (2025). Artificial intelligence in clinical decision support and the prediction of adverse events. *BMJ Health & Care Informatics*.

Radanliev, P., et al. (2025). AI ethics: Integrating transparency, fairness, and privacy for trustworthy intelligent systems. *Applied Artificial Intelligence*.

Sáez, C., et al. (2024). Resilient artificial intelligence in health: Synthesis and research framework. *Journal of Medical Internet Research*.

Sagona, M., et al. (2025). Trust in AI-assisted health systems and AI's trust in humans. *npj Health Systems*.

Salimparsa, M., et al. (2025). Explainable AI for clinical decision support systems. *Informatics*.

Singhal, A., et al. (2024). Toward fairness, accountability, transparency, and ethics in AI-enabled health information systems. *JMIR Medical Informatics*.

Suleman, M. U., et al. (2025). Assessing the generalizability of artificial intelligence in radiology: A systematic review of performance across clinical settings. *European Radiology*.

Tun, H. M., Malik, O. A., et al. (2025). Trust in artificial intelligence-based clinical decision support systems: A systematic review. *Journal of Medical Internet Research*.

Windecker, D., et al. (2025). Generalizability of FDA-approved AI-enabled medical devices. *JAMA Network Open*.

Al-Nafjan, A., et al. (2025). *Artificial intelligence in predictive healthcare: A systematic review*. *Journal of Clinical Medicine, 14*(19), 6752.

Hussein, R., et al. (2026). Advancing healthcare AI governance through a systematic review of implementation frameworks. *npj Digital Medicine*.

Khude, H., et al. (2025). AI-driven clinical decision support systems. *Current Problems in Cardiology*.

Le, L. P., Nguyen, T., Riegler, M. A., Halvorsen, P., & Nguyen, B. T. (2025). Multimodal missing data in healthcare: A comprehensive review and future directions. *Computer Science Review, 55*, 100720.

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

Salimparsa, M., et al. (2025). Explainable AI for clinical decision support systems. *Informatics, 12*(4), 119.

Sun, K., Xue, S., Sun, F., Sun, H., Luo, Y., Wang, L., Wang, S., & Guo, N. (2025). Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions. *Artificial Intelligence in Medicine*.

Wells, B. J., Nguyen, H. M., McWilliams, A., Pallini, M., Bovi, A., Kuzma, A., Kramer, J., Chou, S.-H., Hetherington, T., Corn, P., Taylor, Y. J., Cuison, A., Gagen, M., Isreal, M., & FAIR-AI Consortium. (2025). A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *npj Digital Medicine, 8*, 514.

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence, 316*, 103839.