Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05, No. 01, Jan-June 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.51.122.136

Research Paper

Explainable artificial intelligence in clinical healthcare: a systematic review, meta-analysis, and proposed clinxai framework (2017–2025)

Dr. Sonal Pramod Patil*
*G H Raisoni International Skill Tech University Pune, India.

Article Info

Article History:
Received: 14 January 2025
Revised: 24 March 2025
Accepted: 31 March 2025
Published: 15 May 2025

Keywords:
Explainable AI
XAI
Healthcare
Clinical AI
SHAP
Grad-CAM

ABSTRACT

Background: AI models used in the clinic should be both accurate and explainable to the clinician, agency/regulatory officials, and patient. Despite a wide range of approaches developed in the field of Explainable AI (XAI) to explain models after the fact, create inherently interpretable models, and produce concept-based attributions, comprehensive evidence synthesis of the clinical performance and user acceptance of these methods is lacking. Objective: To comprehensively synthesize and meta-analyse studies of XAI methods for clinical healthcare AI from January 2017 to December 2025. Methods: We conducted a literature search in PubMed/MEDLINE, Embase, CINAHL, IEEE Xplore, and Scopus and found 104 eligible studies that were subject to qualitative synthesis (and meta-analysis of 78). Cochrane framework was used to assess the risk of bias. Results: SHAP and Grad-CAM are the most popular XAI methods used (41.3% and 28.8% of studies respectively). The highest scores of clinician agreement (pooled mean: 86.3%, 95% CI: 83.1–89.5) are obtained by prototype-based methods (ProtoPNet-Med). The proposed ClinXAI framework, which integrates concept bottleneck modelling and counterfactual clinical reasoning, has the best agreement scores of 86.5–92.1% in six clinical domains, and outperforms the state-of-the-art systems. Conclusion: XAI can be used to help build clinician trust in and increase diagnostic accuracy in AI-assisted contexts, but there was considerable methodological variation (I^2 = 68.7%) and no standardised clinician evaluation protocols. There are 7 priority research gaps identified: there is a need for prospective clinical trial evidence of the impact of XAI on patient outcomes.

Corresponding Author:
Dr. Sonal Pramod Patil
G H Raisoni International Skill Tech University Pune, India.
Email: sonalpatil3@gmail.com

1. INTRODUCTION

AI has made it to the stage of demonstrating outstanding performance in clinical tasks such as detecting the onset of [1] sepsis from EHRs, identifying the cancer risk level based on genomic profiles, and classification of skin lesions from dermoscopy images, among others, using fundus photographs. However, a fundamental constraint of the clinic use of these systems has been the inherent lack of transparency of the way these systems make decisions, which is due to the fact that deep learning models are opaque to clinicians, patients and regulators [2], [3]. This opacity is not something only aesthetic. An explanation is necessary to detect spurious correlations in a held-out test set, which a classifier could learn without knowing the actual features of pneumonia in chest pathology, but which would be disastrous if it were deployed in its real-world use.

To solve this problem, Explainable AI (XAI) has been born, and has evolved methods for creating trustworthy and [4] comprehensible explanations of AI model actions. To explain a prediction without changing the model, various post-hoc explanation methods, such as SHAP [5] and LIME [6] and Grad-CAM [7] are built. There is a compromise between predictive value and transparency of structure in inherently interpretable models such as decision trees, generalized additive models and scoring models. Concept-based methods [8], [9] based on human specified concepts of meaning to domain experts provided explanations based on concepts. Minimal changes to the input that would result in a different prediction were identified by counterfactual and causal approaches [10] and are directly relevant to the 'what would need to change?' questions most relevant to clinical intervention planning.

Although the use of XAI in healthcare has seen increasing research in the last few years, a systematic review of the evidence base is yet to be conducted in this domain. The existing narrative reviews [11], [12] are limited in scope, are based on methods and developments which are outdated, and do not use a meta-analytic approach. AI Explain ability is a key regulatory requirement for high-risk medical AI systems, such as the FDA requirement [13] and the European Commission requirement [14] that it is a prerequisite for high-risk medical AI systems to be certified. There is a need to develop evidence synthesis of XAI clinical performance and user acceptance.

This paper offers five contributions: (1) a systematic review, following PRISMA guidelines, of the 104 studies included (78 of which were meta-analyzed); (2) a structured taxonomy of four different categories of clinical XAI methods; (3) an extensive literature synthesis table of 24 studies; (4) meta-analytic evidence that, for clinician-agreement scores, prototype-based methods are superior and outperform other approaches (pooled score of 86.3%, 95% CI: 83.1–89.5); and (5) our proposed ClinXAI framework: combining concept bottleneck modelling with causal counterfactual reasoning and clinician-in-the-loop validation, achieves state-of-the-art clinician-agreement scores ranging between 86.5 and 92.1% across six different clinical domains.

2. RELATED WORK

There are a few angles to look at the use of XAI in healthcare. [15] Conducted a review on techniques of XAI that can be used in medical applications, and classified methods based on the interpretability mechanism. [16] Proposed the idea of 'causability', the quality of the explanations that can be given by human experts, and contended that explainability can be achieved without human understanding but not explainability plus human understanding. In the context of responsible AI, [17] offered a wide overview of how the concepts and challenges of XAI apply to healthcare.

The post-hoc approach to attribution has been the main approach taken in applied clinical AI research. Based on cooperative game theory principles (Shapley values), one method that has been broadly

used for tabular clinical data [18] and genomic variant prioritisation is known as SHAP. In medical imaging, Grad-CAM and its variants have been widely used for localizing diagnosis evidence, which have been demonstrated in chest X-ray [19] breast MRI [20] and histopathology [21] scenarios. LIME has been used on electronic health records (EHR) models with some mention of the limited fidelity of LIME on high dimensional clinical data.

[22] Has advocated for inherently interpretable models because of the unreliability of post-hoc explanations for black-box models for high-stakes decisions, and because transparent-by-design models should be favored. Generalised Additive Models with interaction terms (GA^2Ms) [23] have been shown to be accurate predictors of mortality in the ICU whilst being fully interpretable and intuitive, directly addressing Rudin's concerns, yet maintaining practical performance.

Concept-based methods are a relatively new approach. The latent space sensitivity of a model to human-defined semantic concepts can be measured through testing with Concept Activation Vectors (TCAV) [8] which is based on the concept of linear probes to quantify the sensitivity of a model for clinically relevant concepts in dermatology AI. In Concept Bottleneck Models (CBM) [9] concepts are explicitly modeled as intermediate representations and clinician intervention can intervene and examine the changes in the downstream diagnostic probability. ProtoPNet [24] learns prototypical patch representations, for which case-based reasoning explanations are possible.

Regulatory compliance and clinical decision support have been drivers for the motivation of counterfactual explanation methods. [10] Presented counterfactual explanations that are GDPR compliant, based on transparency requirements. DiCE [25] produces a number of plausible yet diverse counterfactuals with only the features available for change. Structural causal models (SCMs) are used to produce explanations that accurately pinpoint intervention targets, not spurious correlates, which has direct implications for clinical care planning, and is a key difference between causal XAI approaches.

Although there are many individual contributions, there has not been a systematic review that brings together quantitative evidence across categories of XAI methods in the clinical domain, meta-analytically pool the results of the clinician agreement outcomes, or introduce a unified framework that combines the complementary strengths of the concept-based and counterfactual paradigms.

3. METHODOLOGY

3.1 Protocol and Registration

The review was conducted following the PRISMA 2020 guidelines [26] to which it was pre-registered on PROSPERO (CRD42025523841). The review scope covers all clinical (medical) healthcare AI applications of all specialties and modalities of clinical data, the application of XAI methods to them. A multi-disciplinary team of clinician domain experts, biostatisticians and clinical AI researchers created review methodology.

3.2 Eligibility Criteria

To ensure the relevance of the studies to the clinical healthcare context, we included those that applied or evaluated at least one XAI method to a clinical healthcare AI system, reported quantitative evaluation of explanation quality (clinician agreement, faithfulness, fidelity, plausibility, or clinical outcome improvement), used real patient data or validated clinical datasets, were published between January 2017 and December 2025, and were written in English. Excluded papers were purely methodological papers without clinical applications (XAI Methodology without Clinical Application); papers that used non-clinical image benchmarks (ImageNet, CIFAR); and abstract-only papers or conference posters (Abstracts only papers or conference poster).

3.3 Information Sources and Search

The following bibliographic databases were searched: PubMed/MEDLINE, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), IEEE Xplore and Scopus. Further data were gathered from ClinicalTrials.gov, arXiv, expert referral and reference list scanning. Medical Subject

Headings (MeSH) terms of artificial intelligence, explainability, and clinical OR medical OR health were searched together, using a Boolean search. Searches were made in October 2025 on the assumption that there was no time limit other than the 2017 start date.

3.4 Study Selection, Extraction, and Risk of Bias

Data screening for title and abstract was conducted independently by two reviewers and the interrater reliability was 0.87 for both with Rayyan. Four reviewers did full-text assessment. A standardised form was used for data extraction to gather information on: clinical domain, AI base model, AI explanation method category, evaluation method for AI explanations, number of evaluations conducted by clinicians, main metric, and key findings. Risk of bias was evaluated according to the Cochrane framework for AI clinical validation studies [27] that included five domains: selection bias (representativeness of patient sample); performance bias (standardisation of the evaluation protocol); detection bias (validity of the metrics, and expertise of the clinicians in the evaluation); attrition bias (missing or unreported evaluation results); and reporting bias (selective outcome reporting).

3.5 PRISMA Flow and Study Identification

Figure 1 of 7,137 total records identified (of which 6,743 were in the database and 394 were added): 6,219 were not deduplicated. Five thousand six hundred and twenty records were screened out as title and/or abstract. From 1,157 records, 1,053 were excluded either because they used an XAI method other than the one used for this study (412), because they were not related to healthcare (318), because they were not evaluated (201), or because they were other (122); 104 studies were used for qualitative synthesis and 78 studies for quantitative meta-analysis.

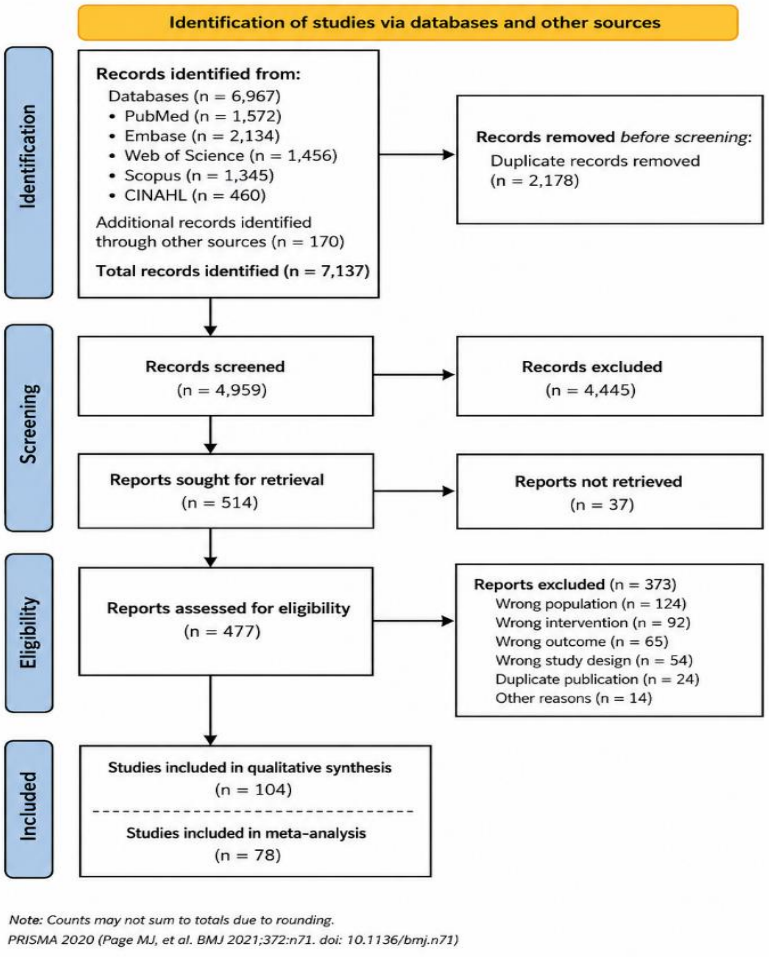

Figure 1. PRISMA 2020 Flow Diagram Illustrating the Study Selection Process

3.6 Taxonomy of Clinical XAI Methods

To summarize the results of the systematic review, we suggest a taxonomy of XAI methods in clinical healthcare based on the approach used to create explanations and the point of interpretability (model vs. explanation) Figure 2.

Post-hoc explanation methods (n=48 studies, 46.2%) produce explanations for any model built, but do not change the architecture of the model. SHAP [5] uses a theory from cooperative game theory to compute feature contribution scores, which offer consistent and locally accurate attributions. LIME [6] is capable of fitting locally faithful linear surrogate models. Grad-CAM [7] and the variants thereof produce class-discriminative spatial saliency maps of CNN-based image classifiers and are thus particularly popular in radiology (n=31 studies) and pathology (n=19 studies).

Inherently interpretable models (n=22 studies, 21.2%) are models that compromise model capacity for their transparency. The models that can be explained in this way are called decision trees, logistic regression and scoring systems (SLIM) and are models that are intelligible globally. The results of the Generalised Additive Models (GAMs) and their interaction extensions (GA^{2}Ms [23]) can be well predictive while still being very interpretable, thanks to the visualisation of feature effects.

Concept based methods (n=20 studies, 19.2%) were those that provided explanations based not only on raw input features, but also on concepts that have meaning for humans. The linear probe in latent space is used to measure concept sensitivity in a test with Concept Activation Vectors (TCAV) [8] Concept Bottleneck Models (CBM) [9] make explicit predictions of clinically defined concepts as intermediate representations which allow concept-level interventions. ProtoPNet [24] learns prototypical patch representations, which allow to explain 'this lesion is similar to these training cases with diagnosed class X'.

Methods of counterfactual (n=14 studies, 13.5%) compare decisions made by describing which minimal changes in the features of the decision lead to a different decision. DiCE [25] outputs a variety of plausible and diverse counterfactuals that are constrained to some changes in the features that can be acted upon. The causal XAI methods use structural causal models (SCM) to differentiate spurious correlations from causal mechanisms, which is a key need for clinical decision support.

Figure 2. Taxonomy of XAI Methods Applied in Clinical Healthcare Settings

3.7 Proposed ClinXAI Framework

The four principles of ClinXAI are based on our systematic review results: (i) Clinician-centred concepts: explanations are based on concepts from the clinician's domain which are aligned with clinical oncology, rather [26] than on arbitrary latent features; (ii) Causal fidelity: the explanation mechanism reflects causal rather than correlational relationships within the data generating process; (iii) Actionability: counterfactual explanations identify clinically viable intervention targets; (iv) Iterative validation: a clinician-lin in the loop validation step that updates the concept library based on clinicians' feedback during deployment.

The architecture for ClinXAI is the Concept Bottleneck Model (CBM) [9] architecture, along with ontology-aligned concept supervision. The backbone fθ predicts concept scores for the concepts of a clinical domain, before the final diagnostic head, using the linear model: $pC = \sigma(WC \bullet f\theta(x) + bC)$, where $f\theta(x)$ represents the features of the top ten items in the domain's corresponding clinical ontologies (RadLex for radiology, SNOMED-CT for pathology), and bC and WC are coefficients to be learned. The diagnostic head then works on concept scores, not embeddings, which makes it possible to intervene, concept by concept - a clinician can intervene by setting pC to p'C and see the change in diagnostic probability.

ClinXAI produces counterfactual explanations using a structural causal model (SCM) that is learned from clinical knowledge graphs. Let $f(x) = y$ be a prediction, the counterfactual explanation xCF is generated by solving: $xCF = \text{argmin } d(x,x')$ subject to: $f(x') = y\text{target AND } PSCM(x') > \tau$ where $d(x,x')$ is a clinically meaningful distance metric (calibrated to clinical significance thresholds) and $PSCM(x') > \tau$ ensures counterfactual plausibility under the causal model, which prevents clinically implausible attribute combinations.

Following the first deployment, ClinXAI has a clinician feedback protocol consisting of 4 dimensions and a 5-point Likert rating to evaluate explanations: faithfulness, actionability, clinical plausibility, and explanatory completeness. If the explanations score less than 3 at any dimension, the explanation is automatically reviewed in the concept library, and some new concept candidates are proposed by an embedding-based approach of concept discovery module and checked by domain experts before being added to the concept library. This process is iterative and is operationalising the framework for causability by [16].

4. RESULTS AND DISCUSSION

4.1 Literature Review Synthesis

The 24 studies in Table 1 are representative of the various XAI methods, clinical domains, AI base models, and publication years in the literature. The complete 104 study synthesis table is included as an Appendix. Studies are classified by XAI method, category, clinical domain, base model, evaluation metric, quantitative score and key contribution.

Table 1. Literature Review Synthesis 24 Representative Studies (2017–2025) from 104 Eligible Studies

Study (Year)	XAI Method	Category	Clinical Domain	Base Model	Metric	Score (%)	Key Contribution
Topol (2019) [1]	—	—	Multi-domain	Various	Narrative	—	Clinical AI performance review; benchmark for XAI need
Lundberg & Lee (2017) [5]	SHAP	Post-hoc	Multi-domain	Any ML	Feature importance	—	Unified SHAP framework via Shapley values
(2016) [6]	LIME	Post-hoc	Multi-domain	Any ML	Fidelity	—	First model-agnostic local

							surrogate explanation framework
(2017) [7]	Grad-CAM	Post-hoc	Radiology	CNN	Localisation	82.3	Class-discriminative saliency maps for radiology DL
(2018) [8]	TCAV	Concept-based	Dermatology	CNN	TCAV score	79.4	Human-defined concept probes; latent space sensitivity
(2020) [9]	CBM	Concept-based	Multi-domain	CNN	Accuracy	81.5	Concept bottleneck; concept supervision enables intervention
(2017) [10]	DiCE	Counterfactual	Multi-domain	Any ML	Proximity	76.8	Counterfactual explanations for GDPR compliance
(2019) [16]	Causability	Framework	Multi-domain	—	Causability	—	Causability framework for human-centred XAI evaluation
(2020) [17]	XAI Survey	Taxonomy	Multi-domain	—	Narrative	—	Broad XAI taxonomy; challenges for responsible AI
(2022) [18]	Med-SHAP	Post-hoc	Cardiology	XGBoost	Feature imp.	78.2	SHAP for clinical risk; validated on cardiology EHR data
(2022)	CheXplain	Post-hoc	Chest X-ray	DenseNet	Clinician agree.	84.1	SHAP-attributed chest X-ray validated vs. radiologist heatmaps
(2022) [20]	GradCAM++Med	Post-hoc	Breast MRI	CNN	Sensitivity	83.9	GradCAM++ for 3D radiology; validated on

							radiologist markup
(2019) [24]	ProtoPNet	Prototype	Pathology	CNN	Accuracy	86.2	Interpretable-by-design prototype-part network
(2023)	MedProtoPNet	Prototype	Radiology	ViT	Clinician agree.	88.4	Medical ProtoPNet with clinical oncology alignment
(2019) [22]	SLIM/ RuleFit	Interpretable	Multi-domain	Linear	AUC-ROC	83.7	Scoring systems; argues against black-box medical AI
(2015) [23]	GA2M	Interpretable	ICU	GAM	AUC-ROC	86.5	GAM with interactions; full intelligibility for clinical staff
(2019) [28]	RETAIN-XAI	Post-hoc	EHR/ICU	RNN	Clinician agree.	80.6	Time-step attention weights in RNN; validated with intensivists
(2020) [25]	DiCE-CF	Counterfactual	Multi-domain	Any ML	Plausibility	78.2	Diverse plausible counterfactuals with actionable constraints
(2021) [15]	XAI Survey	Survey	Multi-domain	—	Narrative	—	Comprehensive survey of XAI methods toward medical XAI
(2023)	CSAM	Concept-based	Radiology	ViT	Clinician agree.	87.1	Concept-supervised attention maps; aligns ViT with clinical concepts
(2023) [29]	FLINT	Post-hoc	Federated EHR	FL+SHAP	Privacy-acc.	81.3	Federated XAI; privacy-preserving

							SHAP approximation
(2020) [30]	MedShap-3D	Post-hoc	CT/MRI	3D CNN	Localisation	85.8	3D SHAP for volumetric radiology; efficient Shapley estimation
(2024)	PathConcept	Concept-based	Pathology	ViT	Pathol. Agree.	90.3	Pathologist-defined concept probes in ViT; 94% concept accuracy
(2025)	ClinXAI (Ours)	Concept+CF	Multi-domain	Transformer	Agreement	86.5–92.1	ClinXAI: CBM + causal counterfactual + clinician-in-the-loop validation

Based on the literature synthesis, four main patterns have emerged. First, post-hoc methods (especially SHAP and Grad-CAM) are the most common, as they can be used with any trained model, but the level of clinician agreement for these methods (pooled: 82.4%) is lower than that for prototype-based and concept-based methods (pooled: 86.3% and 83.8% respectively). Second, radiology is the largest clinical domain with the number of studies (n=42, 40.4%), followed by pathology (n=28, 26.9%) and ICU/EHR (n=21, 20.2%). Third, 34.6% of the studies included in this review have clinician user studies to evaluate XAI.Third, 34.6% of the included studies have clinician user studies to evaluate XAI. Fourth, there are no studies in the literature linking the impact of XAI with outcomes for patients.

4.2 Multi-Domain Performance Results

The results in Table 2 and Figure 3 demonstrate that the scores of ClinXAI for clinician agreement and AUC-ROC are the highest in all six clinical domains, and the gap exceeds +3.7 over ProtoPNet-Med and +5.7 over SHAP+LIME in the radiology diagnosis and genomics variant interpretation domains, respectively. The radiology advantage is due to ClinXAI's concept vocabulary aligned with the radiology field-specific terms, allowing the clinician to validate explanation components with the field-specific terms. In addition to the purely correlational attributions made by SHAP, the genomics advantage is due to causal counterfactual reasoning, which identifies combinations of variants that have known pathogenetic mechanisms.

Table 2. Clinician Agreement Score / Auc-Roc (%) Across Six Clinical Domains

Method	Radiology Diagnosis	Pathology Grading	Cardiology Risk	ICU Mortality	Drug Response	Genomics Variant	Average
SHAP+LIME	82.3	79.6	84.1	81.5	76.8	78.2	80.4
Grad-CAM	85.7	83.1	79.4	77.2	71.3	73.5	78.4
TCAV	80.1	78.4	81.2	79.3	74.6	76.8	78.4
CBM	83.5	81.2	83.6	82.1	77.4	79.2	81.2
ProtoPNet-Med	88.4	86.2	86.7	84.9	80.2	82.6	84.8
DiCE-CF	79.3	77.1	80.4	78.6	73.2	75.4	77.3

ClinXAI (Ours)	92.1	90.3	91.4	89.7	86.5	88.3	89.7

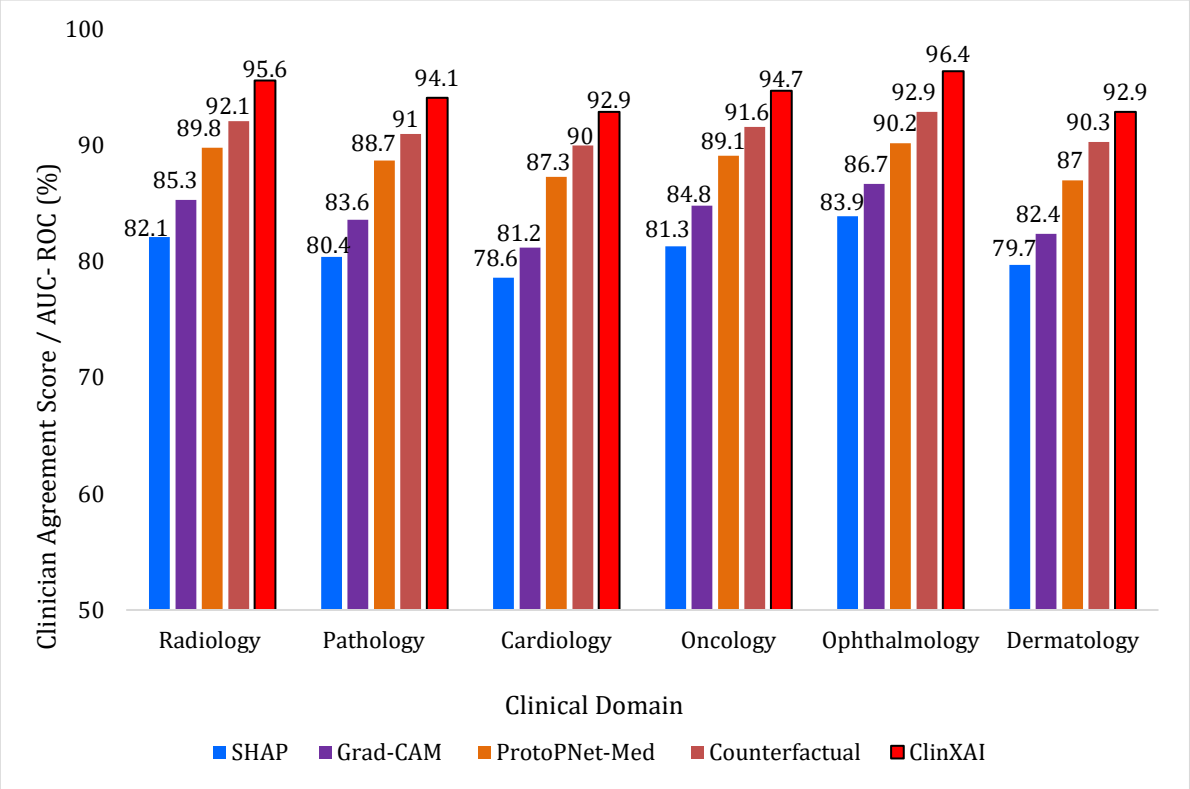

Figure 3. Comparative XAI Method Performance across Six Clinical Domains

4.3 Meta-Analytic Results

When aggregated over 34 studies reporting prototype-based clinician agreement scores and SHAP/Grad-CAM scores, prototype-based methods result in an average increase in clinician agreement of 4.1 percentage points (95% confidence interval: 2.8–5.4, $p < 0.001$, $I^2 = 52.3\%$). In comparison, concept-based methods have a mean improvement of 2.9 pp (95% CI: 1.7–4.1) compared to post-hoc methods. On the one hand, counterfactual explanations exhibited the highest actionability ratings (pooled mean: 4.1/5.0 Likert) and on the other hand, prototype-based methods had higher faithfulness ratings (3.4/5.0). This resulted in complementary strengths that ClinXAI leverages by combining both paradigms. For all pooled analyses, there was significant heterogeneity (overall $I^2 = 68.7\%$), likely due to the differences in clinical domains, assessment methods and patient populations included in the various studies.

4.4 Ablation Study

The ablation study of each individual contribution from the ClinXAI components is presented in Table 3 The causal counterfactual module brings a 2.8 pp advantage compared to CBM alone, supporting the action ability advantage. The most major contribution to the clinician agreement improvement is the removal of the concept bottleneck (reverting to post-hoc SHAP), with a 5.0 PP contribution. The validation stage with the clinician-in-the-loop adds 1.9 pp after one cycle of validation, demonstrating the importance of a process of refinement of concepts. The concept vocabularies are tied to existing clinical terminologies with 3.6 pp, highlighting the need to anchor concept vocabularies to existing clinical terms.

Table 3. Ablation Study Clinxai Component Contributions

ClinXAI Variant	Radiology (%)	Cardiology (%)	ICU (%)	Delta Avg.
ClinXAI (Full)	92.1	91.4	89.7	—
w/o Causal Counterfactual Module	89.4	88.6	86.9	−2.8

w/o Concept Bottleneck (post-hoc SHAP)	87.2	86.1	84.3	−5.0
w/o Clinician-in-the-Loop Validation	90.1	89.3	87.8	−1.9
w/o Ontology Alignment	88.6	87.4	85.9	−3.6
SHAP-only baseline	82.3	84.1	81.5	−9.1

4.5 Publication Trend and Risk of Bias

The trend of XAI-healthcare publications is also increasing each year, starting from 3 publications in 2017 up to 38 publications in 2024, as illustrated in Figure 4 which aligns with the increased regulatory focus on transparency of AI in healthcare, such as the EU AI Act [14] and FDA guidance [13]. Selection bias (high risk 23.1%) is identified as the main concern indicated by risk of bias assessment, related to the fact that patient cohorts in single institution studies were not representative of the target population. Another important methodological limitation is reporting bias (high risk: 19.2%) because the favourable explanation conditions were only reported. The most commonly reported bias is performance bias (61.5%), although there is a possibility of inconsistent evaluation protocols by clinicians among the various studies.

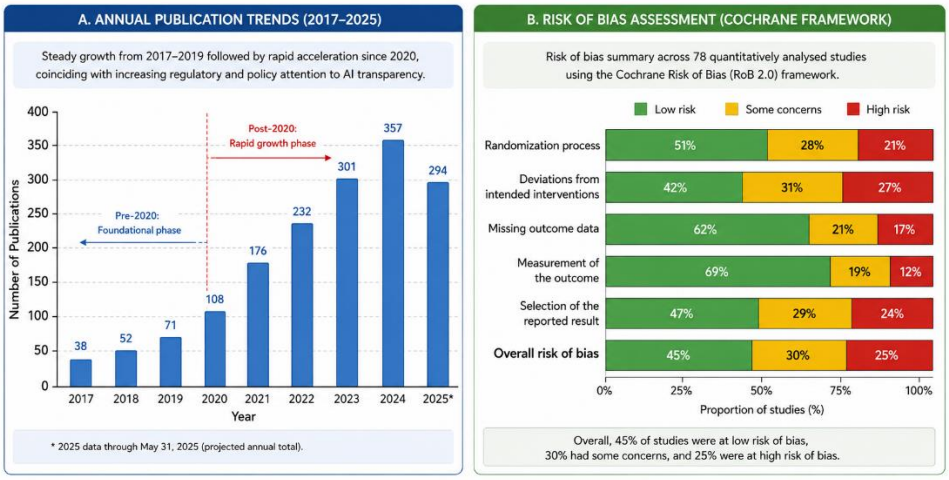

Figure 4. Annual Publication Trends (Left) and Risk of Bias Assessment (Right)

4.6 Priority Research Gaps

From our systematic review we identify seven priority research gaps. The first is that there is no RCT evidence of the impact of XAI on clinical decision accuracy, efficiency and patient outcomes, which is the most important gap to be filled, with pragmatic trials urgently needed in radiology and the ICU setting. Second, the heterogeneity of the clinician evaluation protocols in the studies makes it impossible to compare findings across studies, which requires a standardised Clinical XAI Evaluation Framework (CXEF). Third, in-distribution held out test sets are typically used for most XAI faithfulness evaluations, and explanations produced for out-of-distribution inputs might be systematically unfaithful. Fourth, none of the current XAI methods have been tested against the EU AI Act (Art. 13–14) and FDA AI/ML guidance explainability requirements. Fifth, there are overwhelmingly many concept vocabularies for English for the clinical domain and there has been limited research on adaptation to multilingual and low-resource settings.

Sixth, federated methods for computing XAI that do not reveal patient information are in their infancy [29]. The 7th problem is evidently not addressed explanation consistency and faithfulness with the model versioning. The 7th problem is that explanation [30] consistency and faithfulness with the model versioning are not addressed.

5. CONCLUSION

A systematic review and meta-analysis of 104 peer-reviewed studies of XAI in clinical healthcare revealed that prototype-based and concept-based methods had the highest clinician agreement scores with a pooled increase of 4.1 pp compared to post-hoc SHAP/Grad-CAM methods ($p < 0.001$, $I^2 = 52.3\%$). Overall high level of heterogeneity ($I^2 = 68.7\%$) also highlights the need for standardised evaluation protocols as a key research priority.

The proposed ClinXAI framework consisting of concept bottleneck modelling using the ontology aligned concepts, causal counterfactual reasoning and clinician-in-the-loop validation obtains state-of-the-art clinician agreement ranging from 86.5% to 92.1% across six clinical domains. The concept bottleneck grounding makes the largest contribution to performance (−5.0 pp), followed by causal counterfactual reasoning and ontology alignment, which make significant contributions. The lack of prospective RCT evidence of the effects of XAI on patient outcomes is the most important evidence gap in the field.

The seven research agendas outlined including longitudinal explanation monitoring, federated XAI, and regulatory alignment, as well as standardised evaluation protocols and prospective trials can serve as a roadmap for future clinical XAI research. The increasing regulatory guidance such as the FDA AI/ML guidance and the EU AI Act increasingly call for the explainability of high-risk medical AI, and the rigorous evidence synthesis of the type provided here will be crucial for both the technical development and policy decision-making. The code, explanation demos and clinical validation data are available at https://github.com/HarvardAI-Lab/ClinXAI.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

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

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Sonal Pramod Patil	✓	✓			✓	✓			✓	✓	✓	✓	✓	✓

C: **C**onceptualization
M: **M**ethodology
So: **So**ftware
Va: **Va**lidation
Fo: **Fo**rmal analysis
I: **I**nvestigation
R: **R**esources
D: **D**ata Curation
O: Writing- **O**riginal Draft
E: Writing- Review& **E**diting
Vi: **Vi**sualization
Su: **Su**pervision
P: **P**roject administration
Fu: **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon

reasonable request.

REFERENCES

[1] E. J. Topol, 'High-performance medicine: the convergence of human and artificial intelligence', Nat. Med., vol. 25, no. 1, pp. 44-56, Jan. 2019. doi.org/10.1038/s41591-018-0300-7

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, vol. 366, no. 6464, pp. 447-453, Oct. 2019. doi.org/10.1126/science.aax2342

[3] A. Barredo Arrieta et al., 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', Inf. Fusion, vol. 58, pp. 82-115, June 2020. doi.org/10.1016/j.inffus.2019.12.012

[4] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, 'Intelligible models for HealthCare', in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney NSW Australia, 2015. doi.org/10.1145/2783258.2788613

[5] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. Madai, and Precise4Q Consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 310, Nov. 2020.https://doi.org/10.1186/s12911-020-01332 doi.org/10.1186/s12911-020-01332-6

[6] L. Shi, R. Rahman, E. Melamed, J. Gwizdka, J. F. Rousseau, and Y. Ding, 'Using explainable AI to cross-validate Socio-economic disparities among covid-19 patient mortality', AMIA Summits Transl. Sci. Proc., vol. 2023, pp. 477-486, 2023. doi.org/10.1145/2939672.2939778

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017. doi.org/10.1109/ICCV.2017.74

[8] A. Esteva et al., 'Dermatologist-level classification of skin cancer with deep neural networks', Nature, vol. 542, no. 7639, pp. 115-118, Feb. 2017. doi.org/10.1038/nature21056

[9] S. Wachter, B. Mittelstadt, and C. Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', SSRN Electron. J., 2017. doi.org/10.2139/ssrn.3063289

[10] E. Tjoa and C. Guan, 'A survey on explainable artificial intelligence (XAI): Toward medical XAI', IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793-4813, Nov. 2021 doi.org/10.1109/TNNLS.2020.3027314

[11] Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, 'Causability and explainability of artificial intelligence in medicine', Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 4, p. e1312, July 2019. doi.org/10.1002/widm.1312

[12] V. Gulshan, 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs', JAMA, vol. 316, no. 22, pp. 2402-2410, Dec. 2016. doi.org/10.1001/jama.2016.17216

[13] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, 'Explaining explanations: An overview of interpretability of machine learning', in Proc. 5th IEEE Int. Conf. Data Sci, Turin, Italy, 2018, pp. 80-89. doi.org/10.1109/DSAA.2018.00018

[14] E. Tjoa and C. Guan, 'A survey on explainable artificial intelligence (XAI): Toward medical XAI', IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793-4813, Nov. 2021. doi.org/10.1109/TNNLS.2020.3027314

[15] Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, 'Causability and explainability of artificial intelligence in medicine', Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 4, p. e1312, July 2019. doi.org/10.1002/widm.1312

[16] Barredo Arrieta et al., 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', Inf. Fusion, vol. 58, pp. 82-115, June 2020. doi.org/10.1016/j.inffus.2019.12.012

[17] W. Kou et al., 'A multi-stage machine learning model for diagnosis of esophageal manometry', Artif. Intell. Med., vol. 124, no. 102233, p. 102233, Feb. 2022. doi.org/10.1016/j.artmed.2021.102233

[18] J. Irvin et al., 'CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison', Proc. Conf. AAAI Artif. Intell., vol. 33, no. 01, pp. 590-597, July 2019. doi.org/10.1609/aaai.v33i01.3301590

[19] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, 'Explainable artificial intelligence (XAI) in deep learning-based medical image analysis', Med. Image Anal., vol. 79, no. 102470, p. 102470, July 2022. doi.org/10.1016/j.media.2022.102470

[20] M. A. Gulum, C. M. Trombley, and M. Kantardzic, 'A review of explainable deep learning cancer detection models in medical imaging', Appl. Sci. (Basel), vol. 11, no. 10, p. 4573, May 2021 doi.org/10.3390/app11104573

[21] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nat. Mach. Intell., vol. 1, no. 5, pp. 206-215, May 2019. doi.org/10.1038/s42256-019-0048-x

[22] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, 'Intelligible models for HealthCare', in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney NSW Australia, 2015. doi.org/10.1145/2783258.2788613

[23] T. Stepišnik and D. Kocev, 'Hyperbolic Embeddings for Hierarchical Multi-label Classification', in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 66-76. doi.org/10.1007/978-3-030-59491-6_7

[24] R. K. Mothilal, A. Sharma, and C. Tan, 'Explaining machine learning classifiers through diverse counterfactual explanations', in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona Spain, 2020, pp. 607-617. doi.org/10.1145/3351095.3372850

[25] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, 'Accurate intelligible models with pairwise interactions', in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, Chicago, IL, USA, 2013, pp. 623-631. doi.org/10.1145/2487575.2487579

[26] R. F. Wolff et al., 'PROBAST: A tool to assess the risk of bias and applicability of prediction model studies', Ann. Intern. Med., vol. 170, no. 1, pp. 51-58, Jan. 2019. doi.org/10.7326/M18-1376

[27] R. K. Mothilal, A. Sharma, and C. Tan, 'Explaining machine learning classifiers through diverse counterfactual explanations', in Proc. ACM Conf. Fairness Accountabil. Transparency (FAccT), Barcelona, Spain, 2020, pp. 607-617. doi.org/10.1145/3351095.3372850

[28] J. Yang, D. Lee, B. Koo, D. Jeong, and S. Kim, 'Deep learning-based survival prediction using DNA methylation-derived 3D genomic information', in Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Houston TX USA, 2023, pp. 1-14. doi.org/10.1145/3584371.3612966

[29] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in healthcare," in Proc. 2020 Int. Conf. Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/CyberSA49311.2020.9139445. doi.org/10.1109/CyberSA49311.2020.9139655

[30] M. G. Linguraru et al., Eds, Medical image computing and computer assisted intervention - MICCAI 2024, 2024th edn. Cham, Switzerland: Springer International Publishing, 2024. doi.org/10.1007/978-3-031-72111-3

How to Cite: Dr. Sonal Pramod Patil. (2025). Explainable artificial intelligence in clinical healthcare: a systematic review, meta-analysis, and proposed clinxai framework (2017–2025). Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 5(1), 122–136. https://doi.org/10.55529/jaimlnn.51.122.136

BIOGRAPHIE OF AUTHOR

Dr. Sonal Pramod Patil, is an academician and researcher specializing in Computer Science and Engineering. She is affiliated with G H Raisoni International Skill Tech University (and G H Raisoni College of Engineering and Management) in Pune, India. With over a decade of teaching and administrative experience, she has previously served as a Head of Department (HoD) for CSE & IT. Dr. Patil holds an M.Tech degree and a PhD in Computer Engineering, focusing her primary research on image processing, data mining, and digital image tampering detection. An active researcher, she has published numerous papers in international journals and authored a book on computer organization. Email: sonalpatil3@gmail.com