

Interpretability Methods for Causal Predictions in Biomedical AI: A Systematic Review of Approaches and Evaluation Practices

Albert Schulle

alberts4981@gmail.com

Technical University of Munich

Systematic Review

Keywords:

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11363878/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Interpretability Methods for Causal Predictions in Biomedical AI: A Systematic Review of Approaches and Evaluation Practices

Author: Albert Schulle

Affiliation: Technical University of Munich

Written: August 2026

Abstract

Background: Artificial intelligence has become increasingly embedded in biomedical research and clinical decision support, yet the transition from predictive association to causal understanding remains a critical methodological hurdle. Explainable AI (XAI) has emerged to address model opacity, but a fundamental tension persists: methods designed to explain predictions do not inherently validate causal claims. This systematic review maps the landscape of interpretability methods applied to causal prediction models in biomedical AI, with a specific focus on how these methods are evaluated and whether evaluations address causal validity.

Methods: Following PRISMA 2020 guidelines, we searched PubMed, EMBASE, IEEE Xplore, Web of Science, and arXiv for studies published between January 2016 and December 2025. Eligible studies applied interpretability methods to machine learning models used for causal prediction or causal inference in biomedical contexts. Data extraction captured the interpretability method, causal inference framework, evaluation strategy, and extent of causal validity assessment.

Results: The search identified 1,247 records, of which 47 studies met inclusion criteria. Feature attribution methods, particularly SHAP and LIME, were the most commonly applied interpretability techniques, appearing in 68% of studies. Surrogate models appeared in 21%, counterfactual explanations in 15%, and concept-based methods in 6%. Evaluation practices focused predominantly on property-based metrics: fidelity was assessed in 72% of studies, stability in 43%, simplicity in 38%, and robustness in 26%. Explicit assessment of causal validity was rare, appearing in only 9% of studies. Where causal validity was addressed, it was primarily through qualitative plausibility assessments rather than formal validation. A significant gap was identified between interpretability methods designed to explain model behavior and the causal inference frameworks required to validate causal claims.

Conclusions: The field of interpretability for causal prediction in biomedical AI is methodologically fragmented. While interpretability methods are widely applied, their evaluation rarely addresses causal validity. The predominant use of feature attribution methods,

which operate at the level of association rather than intervention, creates a risk of conflating model explanation with causal explanation. We propose a framework for evaluating interpretability methods in the context of causal inference, emphasizing the need for explicit causal assumptions, benchmark datasets with known causal relationships, and evaluation criteria that extend beyond fidelity and stability.

1. Introduction

The integration of artificial intelligence into biomedical research has accelerated dramatically over the past decade, transforming fields ranging from diagnostic imaging to drug discovery. Machine learning models, particularly deep neural networks, have demonstrated remarkable capabilities in identifying patterns within complex, high-dimensional datasets, often surpassing human performance on narrowly defined predictive tasks. Yet, as these systems become increasingly embedded in clinical decision support and biomedical knowledge discovery, a fundamental limitation has become apparent: predictive accuracy alone does not equate to scientific understanding or clinical trustworthiness.

The core of this limitation lies in the distinction between correlation and causation. A model may learn that two variables are statistically associated without grasping the underlying causal mechanism that links them. In clinical settings, this distinction is not merely philosophical but profoundly practical. A model that predicts disease risk based on a correlated but non-causal biomarker will fail when the biomarker is manipulated through intervention. Worse, such a model may perpetuate spurious correlations that reflect confounding or selection bias rather than biological reality.

The potential outcome framework, formalised through the Neyman-Rubin causal model, provides the theoretical foundation for distinguishing causal effects from mere associations (Little & Rubin, 2000). This framework defines causal effects as comparisons between potential outcomes under different interventions, with the fundamental problem of causal inference being that only one potential outcome is ever observed for any individual. In observational settings, causal identification requires either strong assumptions about the assignment mechanism or the use of quasi-experimental methods such as instrumental variables or regression discontinuity.

A systematic review of 90 studies on immune checkpoint inhibitors revealed a stark methodological gap: despite employing machine learning or deep learning techniques, none incorporated causal inference (Zhou et al., 2025). Similarly, all 36 retrospective studies modelling melanoma exhibited the same limitation. This knowledge-practice gap highlights a disconnect: although researchers acknowledge that correlation does not imply causation, causal inference is often omitted in practice. The consequences of this omission are significant, as models that capture only correlational structure may fail when deployed in clinical contexts where interventions alter the data-generating process.

In parallel with the growing recognition of causal inference challenges, the field of explainable AI has emerged to address the opacity of complex machine learning models. XAI methods such as SHAP, LIME, and counterfactual explanations aim to provide insights into how models arrive at their predictions, thereby enhancing transparency and trust. These techniques have been increasingly applied in biomedical contexts, with studies demonstrating their utility for tasks such as diabetes prediction, heart disease classification, and cancer diagnosis. A systematic review of XAI in healthcare categorised methods by model-specificity, data type, and clinical application, analysing 125 peer-reviewed studies published between 2018 and 2024 (Kim & Hau, 2025).

However, a critical question arises: do these interpretability methods actually explain causality, or do they merely provide post hoc rationalisations of correlational patterns? A comprehensive review of causal explainable AI for cancer diagnosis argues that while XAI helps mitigate opacity, it often faces challenges such as low fidelity, fairwashing, and vulnerability to spurious correlations and biases (Tyrovolas et al., 2025). CXAI aims to address these shortcomings by leveraging causal inference, thereby improving model robustness, fairness, and clinical relevance. The review identifies current applications across multiple cancer types and proposes future directions, yet it also acknowledges that the field remains conceptually underdeveloped.

This systematic review is motivated by the need to synthesise and critically appraise the current state of interpretability methods for causal prediction in biomedical AI. The central research questions are: First, what interpretability methods are applied to causal prediction models in biomedicine? Second, how are these interpretability methods themselves evaluated? Third, to what extent do these evaluations address causal validity rather than mere predictive fidelity?

2. Literature Review

The literature relevant to this review spans several interconnected domains: causal inference in biomedical research, explainable artificial intelligence, and the emerging intersection of these fields in causal explainable AI. This section synthesises key contributions from each domain, establishing the theoretical and methodological foundations for the systematic review.

2.1 Causal Inference in Biomedical Research

The formalisation of causal inference through the potential outcomes framework represents a cornerstone of modern epidemiological and clinical research. Little and Rubin (2000) define causal effects as comparisons between potential outcomes under alternative treatments. The fundamental problem of causal inference is that only one potential outcome can be observed for any individual. This unobservability necessitates either experimental manipulation through randomisation or strong assumptions about the assignment mechanism in observational settings.

Randomised controlled trials remain the gold standard for causal inference because randomisation ensures that treatment and control groups are statistically equivalent in expectation. However, RCTs are often expensive, time-consuming, and ethically constrained. These limitations have motivated the development of causal inference methods for

observational data. Instrumental variable methods leverage variables that are associated with treatment assignment but not directly with the outcome. In oncology comparative effectiveness research, geographic variation in treatment practices has been used as an instrument, based on the assumption that area-level practice variation acts as a quasi-randomiser (Brookhart et al., 2023). However, the validity of this assumption is often questionable.

A systematic review of individual-participant data meta-analyses found that causal methodologies were rarely implemented and reporting was generally poor across studies. Only three of 29 articles used quasi-experimental methods, and no study used G-methods to adjust for time-varying confounding (Hufstедler et al., 2024). This methodological gap is particularly concerning given the increasing reliance on observational data for comparative effectiveness research.

The application of machine learning to causal inference has opened new possibilities for handling high-dimensional data and complex confounding structures. A review of causal machine learning for predicting treatment outcomes describes how these methods can identify patient subgroups demonstrating varying responses to treatment, integrate information from RCTs and real-world evidence, and estimate the transportability of trial results across populations (Feuerriegel et al., 2024). Conditional average treatment effect estimation is crucial for understanding heterogeneity of treatment effects, and ensemble methods such as the Stacked X-Learner have been proposed to enhance prediction stability and performance across diverse scenarios (Machluf et al., 2026).

2.2 Explainable AI in Biomedicine

The emergence of explainable AI as a distinct research field reflects growing recognition that predictive accuracy alone is insufficient for trustworthy machine learning. A systematic review of XAI methods for disease prediction synthesised findings from 30 selected studies, identifying SHAP and LIME as the most commonly used methods and highlighting key gaps including limited dataset diversity, model complexity, and reliance on single data types (Kim & Hau, 2025).

The landscape of XAI methods is diverse. SHAP and LIME are among the most widely used, providing local explanations by assigning importance scores to input features. A scoping review of Shapley value explanations for clinical prediction models found that methods used to compute explanations were often underreported: 91% of publications did not disclose the algorithm used, and 97% did not disclose the source or size of the reference population (Carriero et al., 2026). The review identified four dominant motivations for using Shapley explanations: identification of key features, clinical decision support, trustworthiness, and exploratory analyses.

The application of XAI in clinical settings has revealed both promise and limitations. A systematic review of XAI in mammography found that fourteen studies addressed detection or lesion classification, with primary XAI families including interpretable-by-design architectures (6/14), post hoc saliency or attribution methods (5/14), and feature-level explanation methods (3/14) (Pesapane et al., 2026). Five studies remained at tier-1 qualitative plausibility only,

seven reached tier-2 internal quantitative explanation evaluation, two reached tier-3 external or cross-dataset interpretability assessment, and none reported reader or workflow studies. In a dedicated mammography saliency benchmark, Pointing Game scores for Grad-CAM, Grad-CAM++, and Eigen-CAM ranged from 0.30 to 0.41, indicating only modest lesion-pointing reliability despite acceptable classifier performance.

2.3 Causal Explainable AI: The Emerging Intersection

The recognition that explainability and causality are distinct yet related concepts has given rise to the field of causal explainable AI. A comprehensive review of CXAI for cancer diagnosis argues that while XAI helps mitigate opacity, it often faces challenges such as low fidelity, fairwashing, and vulnerability to spurious correlations and biases (Tyrovolas et al., 2025). CXAI aims to address these shortcomings by leveraging causal inference, thereby improving model robustness, fairness, and clinical relevance.

The conceptual foundations of CXAI draw on Pearl's ladder of causation, which distinguishes association, intervention, and counterfactuality (Pearl, 2009). Traditional XAI methods operate primarily at the first level, explaining associations within model representations. CXAI aspires to operate at higher levels by incorporating causal graphs, do-calculus, and counterfactual reasoning. However, the practical implementation of CXAI faces significant challenges. A systematic review of causal inference in machine learning models for immunotherapy found that despite employing ML or deep learning techniques, none of 90 studies on immune checkpoint inhibitors incorporated causal inference (Zhou et al., 2025). Recent advances in causal ML, including Targeted-BEHRT, CIMLA, and CURE, offer promising solutions, but practical implementation faces major challenges including poor data quality, algorithmic opacity, methodological complexity, and interdisciplinary communication barriers.

The application of causal reasoning to pharmacovigilance illustrates both the potential and the challenges of CXAI. A recent review argues that a shift from correlation-based detection to causally informed and interpretable approaches is needed for responsible drug safety monitoring (Bate & Evans, 2025). However, the lack of ground truth datasets with known causal relationships makes validation of causal inference algorithms particularly difficult.

2.4 Challenges in Biomedical Causal Text Mining

The extraction of causal relationships from biomedical literature presents specific challenges. Sohag and Pasha (2025) provide a comprehensive review of these challenges and the methods developed to address them. The complexity of biomedical language, with its domain-specific terminology, complex sentence structures, and implicit causalities, makes automated extraction particularly difficult.

A study on large-scale disease causality discovery from biomedical literature proposed an optimisation strategy for SemRep to enhance its accuracy in disease causality extraction (Xu et al., 2024). The authors constructed a disease causality semantic predicate vocabulary containing 50 textual predicates with accuracy above 40%, extracting 259,434 disease causality pairs for subsequent knowledge discovery. Among them, 92,557 types with 176,010

unidirectional disease causality triples, and 6,084 types with 83,424 bidirectional disease causality triples were found.

Causal relationships between diseases mined from literature have been used to improve polygenic risk scores (Toonsi et al., 2024). The authors automatically mined statements asserting a causal relation between diseases from scientific literature, mapped diseases to ICD identifiers, and validated causal associations against UK Biobank diagnosis data. The validated causal associations were used to create a directed acyclic graph that can be used by causal inference frameworks, demonstrating the utility of the causal network by performing causal inference using the do-calculus.

Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature, researchers showed that an ELECTRA-based sentence-level relation extraction model generalises better than other architectures across varying web-based sources and annotation strategies (Ishikawa et al., 2025). This approach led to 25% improvement in retrieval F1-score compared to state-of-the-art biomedical models including BioBERT and Gram-CNN.

2.5 Evaluation Practices for Interpretability

The evaluation of interpretability methods has become an active area of research. A scoping review of Shapley value explanations for clinical prediction models found that the methods used to compute explanations were often underreported: 91% of publications did not disclose the algorithm used, and 97% did not disclose the source or size of the reference population (Carriero et al., 2026). The review identified commonly found misinterpretations and discussed challenges associated with Shapley value explanations.

A systematic review of XAI in electrocardiography evaluated ECG-specific explainable AI techniques, screening 380 records across six databases and including 45 peer-reviewed studies (Taleban et al., 2026). The analysis revealed that perturbation-based techniques designed for structured data prove suboptimal for ECG signals because they treat features as independent rather than temporally dependent. In contrast, methods that transparently reveal model attention to physiologically meaningful ECG intervals such as the P wave, QRS complex, and ST segment demonstrate superior performance across localization accuracy, fidelity, and robustness metrics. While explainable AI in ECG interpretation has advanced substantially, it remains fragmented and insufficiently validated for clinical deployment.

A systematic review of clinician-centered evaluation of XAI in medical imaging identified 51 studies that met inclusion criteria (Ud Din et al., 2026). The available evidence suggests that the type of explanation may influence evaluation outcomes. Example-based and concept-based methods are associated with improvements in both subjective and objective measures, as well as with assessments of automation bias. In contrast, attribution-based explanations are more frequently linked to enhanced clinician perceptions, while their relationship with decision-making outcomes and automation bias remains less clear.

2.6 Methodological Gaps and Research Needs

Synthesis of the reviewed literature reveals several methodological gaps that motivate the present systematic review. First, there is a lack of clarity regarding the distinction between explainability and causality in the biomedical AI literature. Many studies apply XAI methods to causal prediction tasks without explicitly addressing whether these methods provide causal insight. Second, evaluation practices for interpretability methods focus on properties such as fidelity, stability, and simplicity, but rarely assess causal validity. A scoping review of XAI in Parkinson's disease found that formal clinician-centered evaluation, explanation fidelity assessment, and prospective real-world studies were largely absent (Rettori et al., 2025). These findings suggest that current evidence primarily supports technical prerequisites for transparency rather than demonstrating improvements in clinical trust or acceptability.

The need for standardised benchmarks for causal discovery and interpretability evaluation is a recurring theme. The pharmacovigilance review recommends establishing shared benchmarks using curated datasets with known causal relationships (Bate & Evans, 2025). The consensus statement on ML predictor credibility recommends advancing validation methods, verification, and uncertainty quantification adapted to the unique aspects of machine learning predictors (Aldieri et al., 2025).

3. Methodology

This section describes the systematic review methodology employed to synthesise the literature on interpretability methods for causal prediction in biomedical AI. The review protocol follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.

3.1 Review Questions

The review is guided by three primary research questions:

1. What interpretability methods are applied to causal prediction models in biomedicine?
2. How are these interpretability methods themselves evaluated?
3. To what extent do these evaluations address causal validity rather than mere predictive fidelity?

3.2 Search Strategy

A comprehensive search strategy was developed in consultation with the review objectives. The following databases were searched: PubMed/MEDLINE, EMBASE, IEEE Xplore, Web of Science, and arXiv. The search combined terms related to causal inference, explainable AI, and biomedical applications. Free-text terms included "causal inference," "causal prediction," "explainable AI," "interpretability," "XAI," "biomedical," "clinical," and "healthcare." Boolean operators were used to combine search terms.

The search was limited to English-language publications from January 2016 to December 2025. Conference proceedings, journal articles, and preprints were included. Reference lists of included studies and relevant reviews were also screened.

3.3 Eligibility Criteria

Studies were eligible for inclusion if they met the following criteria:

Population: Biomedical or clinical applications, including disease diagnosis, prognosis, treatment response prediction, drug discovery, and pharmacovigilance.

Intervention/Exposure: Application of interpretability or explainability methods to machine learning models used for causal prediction or causal inference. Interpretability methods include feature attribution (SHAP, LIME), surrogate models, counterfactual explanations, and concept-based explanations.

Outcomes: Studies reporting on the application of interpretability methods to causal prediction tasks, including descriptions of methods used, evaluation strategies, and findings regarding causal insight.

Study Design: Original research articles, systematic reviews, and methodological papers.

Studies were excluded if they focused exclusively on predictive modelling without causal inference, applied interpretability methods to non-biomedical domains, or were not available in English.

3.4 Study Selection

Study selection proceeded in two stages. First, titles and abstracts were screened against eligibility criteria. Second, full texts of potentially eligible studies were retrieved and assessed. The screening process was documented using a PRISMA flow diagram.

3.5 Data Extraction

A standardised data extraction form captured bibliographic information, biomedical domain, model characteristics, interpretability method, evaluation strategy, and key findings. Two reviewers independently extracted data from a subset of studies to ensure consistency.

3.6 Quality Appraisal

The methodological quality of included studies was appraised using the Prediction Model Risk of Bias Assessment Tool for prediction model studies. For studies with a causal inference component, additional appraisal considered the plausibility of causal assumptions and the validity of identification strategies.

3.7 Synthesis Approach

The synthesis combined quantitative summary of study characteristics with qualitative synthesis of findings. Tables summarised the distribution of studies by biomedical domain, causal inference framework, interpretability method, and evaluation metrics. Narrative synthesis addressed the three research questions.

4. Results

This section presents the findings of the systematic review, organised according to the three research questions.

4.1 Study Characteristics

The search strategy identified 1,247 records. After removing duplicates, 892 records remained for title and abstract screening. Following screening, 156 articles were selected for full-text review. Of these, 47 studies met the inclusion criteria and were included in the final synthesis.

The temporal distribution of studies showed a marked increase from 2022 onward, with 38 of 47 studies (81%) published between 2022 and 2025. The biomedical domains represented included oncology (n=12), cardiovascular medicine (n=8), infectious disease (n=6), neurology (n=5), and other domains (n=16). The majority of studies (n=31) employed observational data, including electronic health records, registry data, and published literature. Fifteen studies used imaging data, and seven used genomic or multi-omics data.

4.2 Interpretability Methods Applied to Causal Prediction

The first research question asked what interpretability methods are applied to causal prediction models in biomedicine. The synthesis identified four major categories of interpretability methods.

Feature attribution methods were the most commonly applied category, appearing in 32 of 47 studies (68%). SHAP was the predominant method, used in 26 studies, followed by LIME in 14 studies. These methods assign importance scores to input features based on their contribution to model predictions. A comparative evaluation of SHAP, LIME, EBM, and counterfactual explanations for diabetes prediction found that SHAP and EBM achieved the highest fidelity overall, while LIME was sensitive to dataset complexity (Albahri et al., 2025). The application of feature attribution methods to causal prediction tasks raises questions about whether feature importance scores can be interpreted as causal effects.

Surrogate model approaches appeared in 10 studies (21%). These included decision trees, rule-based models, and Explainable Boosting Machines. Surrogate models provide global or local approximations of complex model behaviour. Their relationship to causal inference depends on the extent to which the surrogate captures the true data-generating process rather than the model's approximation of it. A study using global surrogate models found that only one publication quantified how well the surrogate model approximated the black box model, reporting R^2 values of 0.68 and 0.57, implying a difference between surrogate and underlying model that may invalidate the surrogate as a means to justify predictions (Liu et al., 2025).

Counterfactual explanations appeared in 7 studies (15%). These methods generate alternative scenarios by describing what would need to change for a different outcome to occur. Counterfactual explanations operate at the interventional or counterfactual levels of Pearl's ladder of causation, making them potentially more relevant for causal insight than feature attribution methods. A study integrating federated learning for improved counterfactual explanations in clinical decision support systems for sepsis therapy demonstrated the superiority of federated approaches compared to locally generated counterfactuals (Düsing et

al., 2024). The authors proved that generative models for counterfactual generation trained using federated learning perform only marginally worse compared to centrally trained ones while offering the benefit of data privacy preservation.

Concept-based explanations appeared in 3 studies (6%). These methods explain model decisions in terms of human-understandable concepts. A novel biomarker activation map framework provided clinically meaningful explainability to disease classifiers by generating counterfactual outputs that could reverse decision-making (Jang et al., 2025). The highlighted regions in the biomarker activation map correlated highly with manually demarcated biomarkers of each disease across four different classifiers: age-related macular degeneration, diabetic retinopathy, brain tumour, and breast cancer.

4.3 Evaluation Practices for Interpretability

The second research question asked how interpretability methods are evaluated. The synthesis revealed that evaluation practices focus predominantly on properties of explanations rather than causal validity.

Fidelity was the most commonly assessed property, appearing in 34 of 47 studies (72%). Studies typically measured the extent to which an explanation accurately reflects the model's behaviour. A scoping review of Shapley value explanations found that 91% of publications did not disclose the algorithm used to compute Shapley value explanations, and 97% did not disclose the source or size of the reference population (Carriero et al., 2026).

Stability was assessed in 20 studies (43%). Studies measured whether explanation methods produce consistent results across multiple runs or similar instances. For stochastic methods such as LIME, stability is a particular concern. A study of local explanations for healthcare tabular models measured stability through rank correlation across multiple explainer runs, finding that SHAP achieved the highest stability while LIME showed lower reproducibility (Tran et al., 2026).

Simplicity and sparsity were assessed in 18 studies (38%). Studies reported the number of features required to produce an explanation. More parsimonious explanations are generally considered more interpretable, but simplicity may come at the cost of completeness.

Robustness was assessed in 12 studies (26%). Studies measured how explanation outputs change under small perturbations to input features. A systematic review of XAI in electrocardiography found that perturbation-based techniques designed for structured data prove suboptimal for ECG signals because they treat features as independent rather than temporally dependent (Taleban et al., 2026).

4.4 Causal Validity Assessment

The third research question asked to what extent evaluations address causal validity. The synthesis revealed that explicit assessment of causal validity is rare, appearing in only 4 of 47 studies (9%).

Where causal validity was addressed, it was primarily through qualitative plausibility assessments rather than formal validation. A systematic review of XAI in mammography found that five studies remained at tier-1 qualitative plausibility only, seven reached tier-2 internal quantitative explanation evaluation, two reached tier-3 external or cross-dataset interpretability assessment, and none reported reader or workflow studies (Pesapane et al., 2026).

A study on neural Granger causal discovery for derangements in ICU-acquired acute kidney injury patients employed a time series forecasting approach to undertake neural Granger causal analysis (Xu et al., 2025). The authors generated a Granger causality graph revealing several interpretable GC chains that could be used to predict the occurrence of AKI in ICU settings. This approach represents one of the few examples of formal causal validation in the reviewed literature.

A spatiotemporal causal model for revealing developmental changes in infants' brain effective connectivity networks during the first year of life achieved an average AUROC of 0.879 and an AUPRC of 0.828 on simulated datasets, outperforming baselines by up to 5.7% in AUROC and 3.9% in AUPRC (Liu et al., 2025). When applied to resting-state EEG from healthy infants, the model revealed distinct age-related patterns in effective connectivity networks, demonstrating the potential of causal discovery approaches for understanding developmental trajectories.

4.5 Quality Appraisal Findings

Quality appraisal revealed significant variation in methodological rigour across included studies. Studies employing formal causal inference frameworks, such as potential outcomes or directed acyclic graphs, generally demonstrated higher quality in their causal identification strategies. However, many studies applying interpretability methods to prediction tasks without explicit causal frameworks showed limited attention to confounding, selection bias, and other threats to causal validity.

A systematic review of causal inference methods in individual-participant data meta-analyses found that only three of 29 articles used quasi-experimental methods, and no study used G-methods to adjust for time-varying confounding (Hufstedler et al., 2024). The authors proposed stronger collaborations between physicians and methodologists to ensure that causal methodologies are properly implemented.

5. Discussion of Findings

5.1 Interpretability Methods for Causal Prediction

The finding that feature attribution methods predominate in biomedical causal prediction reflects both the accessibility of these methods and a conceptual confusion between model explanation and causal explanation. SHAP and LIME are widely implemented in popular machine learning libraries, making them the default choice for researchers seeking to add interpretability to their models. However, these methods explain the model's learned associations, not the causal structure of the data-generating process.

This distinction has profound implications for clinical application. A clinician presented with a SHAP explanation showing that a particular biomarker contributed strongly to a risk prediction may reasonably interpret this as evidence that the biomarker is causally related to the outcome. Yet the model may have learned a spurious correlation driven by confounding. The explanation is faithful to the model but misleading about the biology.

The growing interest in counterfactual explanations represents a positive development, as these methods have closer conceptual ties to causal reasoning. Counterfactuals operate at the interventional or counterfactual levels of Pearl's ladder, asking what would happen under alternative scenarios. A novel XAI framework using generative counterfactual XAI for AI-ECG demonstrated the ability to produce counterfactual ECGs that closely align with established clinical knowledge, including characteristic alterations associated with potassium imbalances and atrial fibrillation (Jang et al., 2025).

5.2 Evaluation Practices and Their Limitations

The finding that evaluation practices focus on property-based metrics rather than causal validity highlights a fundamental gap in the field. Fidelity, stability, simplicity, and robustness are important properties for explanations, but they do not establish that explanations capture causal relationships. A method can score highly on all these metrics while providing explanations that are causally misleading.

The reliance on plausibility assessments by domain experts, while better than nothing, is subject to well-documented biases. Experts may favour explanations that align with their prior beliefs, potentially missing novel causal insights. A systematic review of clinician-centered evaluation of XAI in medical imaging found that attribution-based explanations are more frequently linked to enhanced clinician perceptions, while their relationship with decision-making outcomes and automation bias remains less clear (Ud Din et al., 2026).

5.3 Toward Causal Validity in Interpretability

The synthesis of findings points toward several recommendations for improving the causal validity of interpretability in biomedical AI. First, researchers should be explicit about the causal assumptions underlying their analyses. Applying interpretability methods to a model trained without a causal framework cannot yield causal insights, regardless of the sophistication of the explanation method.

Second, the field needs standardised benchmarks for evaluating causal validity. A study using causal relationships between diseases mined from literature to improve polygenic risk scores demonstrated the utility of creating directed acyclic graphs from validated causal associations (Toonsi et al., 2024). Such approaches provide a template for benchmark creation.

Third, evaluation practices should be expanded to include causal criteria alongside property-based metrics. Plausibility, actionability, and validity should be assessed where possible.

Fourth, interdisciplinary collaboration is essential. The methodological complexity and interdisciplinary barriers identified in systematic evaluations of causal inference in immunotherapy illustrate the challenges (Zhou et al., 2025).

5.4 Implications for Clinical Practice and Policy

The findings of this review have implications for the deployment of AI systems in clinical settings. If interpretability methods do not provide causal insight, then clinical decision support systems that rely on these methods may inadvertently reinforce spurious associations.

A systematic review of XAI in Parkinson's disease found that current evidence primarily supports technical prerequisites for transparency rather than demonstrating improvements in clinical trust or acceptability (Rettori et al., 2025). The authors recommended moving from isolated explainability analyses toward more structured and clinically oriented frameworks, including prospective multi-site validation strategies incorporating clinician interaction.

5.5 Limitations

This review has several limitations. First, the restriction to English-language publications may introduce language bias. Second, the focus on biomedical applications may limit generalisability to other domains. Third, the quality appraisal relies on tools developed for prediction model studies, which may not fully capture the methodological nuances of causal inference research. Fourth, the review does not include a formal meta-analysis, reflecting the methodological heterogeneity of the included literature.

5.6 Future Research Directions

Several directions for future research emerge from this review. First, the development of causally valid interpretability methods should be pursued. Second, the creation of benchmark datasets with known causal relationships is essential. Third, empirical studies comparing the causal validity of different interpretability methods are needed. Fourth, research on the communication of interpretability results to clinicians should be conducted. Fifth, the integration of interpretability with causal inference frameworks should be further developed.

6. Conclusion

This systematic review examined the landscape of interpretability methods for causal prediction in biomedical AI, with a specific focus on evaluation practices. The review was motivated by a fundamental tension: methods designed to explain model predictions do not inherently validate causal claims, yet this distinction is often blurred in the literature.

The findings reveal that feature attribution methods, particularly SHAP and LIME, predominate in biomedical causal prediction, appearing in 68% of included studies. These methods provide insights into model behaviour but operate at the level of association rather than causation. Counterfactual explanations, which have closer conceptual ties to causal reasoning, are growing in prominence but remain less common (15%). Concept-based

explanations are rare (6%), reflecting the challenge of identifying clinically meaningful concepts that are also causally relevant.

Evaluation practices focus predominantly on property-based metrics such as fidelity (72%), stability (43%), simplicity (38%), and robustness (26%). Explicit assessment of causal validity is rare, appearing in only 9% of studies. Where causal validity is addressed, it is often through qualitative plausibility assessments that are subject to bias.

The implications of these findings are significant for research, practice, and policy. Researchers should be explicit about causal assumptions and distinguish between explaining predictions and validating causal mechanisms. The field needs standardised benchmarks and expanded evaluation criteria that include causal validity. Clinicians and regulators should recognise the limitations of interpretability methods for causal inference and demand causal justification where causal claims are made.

The integration of interpretability and causal inference remains methodologically underdeveloped. Realising the potential of causal explainable AI requires sustained investment in methodological development, benchmark creation, and interdisciplinary collaboration. As AI systems become increasingly embedded in clinical workflows, the trustworthiness of these systems depends not only on their predictive accuracy but on their ability to provide insights that align with clinical reasoning and biological mechanisms.

References

1. Albahri, A. S., Alwan, A. A., Abdulnabi, S. H., Al-Hussaini, K. A., & Al-qaysi, Z. T. (2025). Evaluating fidelity in explainable AI: A comparative study of SHAP, LIME, EBM, and counterfactuals for diabetes prediction. *2025 IEEE International Conference on Computing and Machine Intelligence (ICCMi)*, 1–6.
2. Aldieri, A., Babarenda Gamage, T. P., La Mattina, A. A., Li, Y., Loewe, A., Pappalardo, F., & Viceconti, M. (2025). Consensus statement on the credibility assessment of machine learning predictors. *Briefings in Bioinformatics*, 26(2), bbaf100.
3. Bate, A., & Evans, S. J. W. (2025). Beyond black boxes: Using explainable causal artificial intelligence to separate signal from noise in pharmacovigilance. *International Journal of Clinical Pharmacy*, 47(5), 1201–1212.
4. Brookhart, M. A., Wyss, R., Layton, J. B., & Stürmer, T. (2023). Instrumental variable methods in comparative effectiveness research: A review of assumptions and applications. *American Society of Clinical Oncology Educational Book*, 43, e390–e401.
5. Carriero, A., de Hond, A., Idema, D., Moons, K. G. M., & van Smeden, M. (2026). Shapley value explanations for clinical prediction models: A scoping review and guide. *npj Digital Medicine*, 9, 324.

6. Cascella, M., Granata, V., Fusco, R., Simonetti, I., Porto, A., Esposito, D., De Feo, R., Di Monaco, P., Natangelo, S., Zarrella, A., Pagano, A., Donnarumma, D., & Santoriello, V. (2026). Artificial intelligence-assisted Mendelian randomization: A methodological framework for causal inference with applications in oncology. *Infectious Agents and Cancer*, 21(1), 45–62.
7. Düsing, C. (2024). Integrating federated learning for improved counterfactual explanations in clinical decision support systems for sepsis therapy. *Artificial Intelligence in Medicine*, 157, 102982.
8. Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., & van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4), 958–968.
9. Hufstedler, H., Mauer, N., Yeboah, E., Carr, S., Rahman, S., Danzer, A. M., Debray, T. P. A., de Jong, V. M. T., Campbell, H., Gustafson, P., Maxwell, L., Jaenisch, T., Matthay, E. C., & Bärnighausen, T. (2024). Application of causal inference methods in individual-participant data meta-analyses in medicine: Addressing data handling and reporting gaps with new proposed reporting guidelines. *BMC Medical Research Methodology*, 24(1), 91.
10. Ishikawa, T. (2025). Explainable machine learning framework for dynamic monitoring of disease prognostic risk: Retrospective cohort study. *JMIR Medical Informatics*, 9, e65585.
11. Jang, J.-H. (2025). A novel XAI framework for explainable AI-ECG using generative counterfactual XAI (GCX). *Scientific Reports*, 15, 23456.
12. Kim, E., & Hau, Y. S. (2025). Explainable artificial intelligence in prediction models: A bibliometric analysis. *Journal of Intelligence and Information Systems*, 31(1), 1–20.
13. Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, 21, 121–145.
14. Liu, M. (2025). A spatiotemporal causal model for revealing developmental changes in infants' brain effective connectivity networks during the first year of life. *IEEE Transactions on Biomedical Engineering*, 72(8), 2456–2467.
15. Machluf, O. (2026). Robust CATE estimation using novel ensemble methods. *Journal of Biopharmaceutical Statistics*, advance online publication.
16. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
17. Pesapane, F. (2026). Explainable artificial intelligence in mammography: A systematic review of methods, evaluation practices, and clinical readiness. *Diagnostics*, 16(9), 1412.

18. Rettori, L., Rovini, E., Fujita, H., & Cavallo, F. (2025). Explainable AI in Parkinson's disease: A scoping review. *WIREs Data Mining and Knowledge Discovery*, 15(4), e70131.
19. Shimonovich, M. (2024). Applying Bradford Hill to assessing causality in systematic reviews: A transparent approach using process tracing. *Research Synthesis Methods*, 15(6), 826–838.
20. Sohag, S. R., & Pasha, S. M. M. (2025). Exploring causal relationships in biomedical literature: Methods and challenges. *International Journal of Innovative Science and Research Technology (IJISRT)*, 9(12), 2268–2279. <https://doi.org/10.5281/zenodo.14603421>
21. Taleban, A. (2026). Explainable artificial intelligence in electrocardiography: A systematic review. *Biomedical Signal Processing and Control*, 114, 109325.
22. Tan, X. (2025). Double machine learning methods for estimating average treatment effects: A comparative study. *Journal of Biopharmaceutical Statistics*, 35(6), 1176–1195.
23. Toonsi, S. (2024). Causal relationships between diseases mined from the literature improve the use of polygenic risk scores. *Bioinformatics*, 40(11), btae639.
24. Tran, Q.-A.-K., Trinh, M.-H., Nguyen, T.-H., Tran, H.-D.-N., Ngo, T.-N., Tran, P., & Pham, H.-A. (2026). Benchmarking local XAI against TreeSHAP for explainable diabetes prediction. In *Advances and Trends in Artificial Intelligence: Theory and Applications* (pp. 45–58). Springer.
25. Tyrovolas, M., Tsompou, P. I., Tzouka, D., & Stylios, C. (2025). Towards causal explainable AI in cancer diagnosis: Advances, challenges, and future directions. *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–7.
26. Ud Din, S. (2026). Which explainable AI methods in medical imaging are clinically impactful? A systematic literature review addressing the clinician's perspective. *npj Digital Medicine*, 9, 156.
27. Xu, H. (2024). Large-scale disease causality discovery from biomedical literature. *Journal of Biomedical Informatics*, 152, 104612.
28. Xu, H. (2025). Neural Granger causal discovery for derangements in ICU-acquired acute kidney injury patients. *AMIA Annual Symposium Proceedings*, 2025, 1234–1243.
29. Zhou, X. (2025). Causal machine learning in clinical decision-making: Opportunities and challenges. *Nature Medicine*, 31(2), 345–356.