Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Reasoning Pathway Explanation Framework for Clinical AI: Methods and Evaluation

Yunguo Yu

yuyunguo@gmail.com

Zyter|TruCare https://orcid.org/0009-0006-2003-3813

Research Article

Keywords: Artificial Intelligence, Clinical Decision Support, Explainable AI, Reasoning Pathways, Ontologies, Diagnostic Reasoning, MIMIC-III

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9418990/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.

A Reasoning Pathway Explanation Framework for Clinical AI: Methods and Evaluation

Yunguo Yu[1]

[1]AI Innovation and Prototyping, Zyter | TruCare, 8000 Towers Crescent Dr, Vienna, VA 22182, USA,

Correspondence: Yunguo Yu (email: yuyunguo@gmail.com)

April 14, 2026

Abstract

Objective: When AI predicts acute myocardial infarction, existing explanation methods identify which features mattered (e.g., elevated troponin) but not how risk factors lead to the diagnosis through biological mechanisms. We developed a framework that generates reasoning pathways—clinically grounded chains linking risk factors, pathophysiology, and evidence—to address this gap.

Methods: Using a 34-node cardiology reasoning graph, MIMIC-III data, and medical ontologies (SNOMED CT, UMLS), we built an explanation engine that maps AI predictions to temporally ordered, evidence-linked pathways. We evaluated 100 AMI cases (87 expanded, 13 independent) with six structural metrics, adversarial validation, and comparison with BioBERT and SHAP. Three physicians independently rated 11 cases across five clinical quality dimensions.

Results: The framework generated consistent reasoning pathways across all cases (pathway-prediction consistency 0.85 ± 0.01). Adversarial validation confirmed discriminative power for this metric (AUC-ROC 0.81) but not others. In the physician pilot, inter-rater agreement was strong (ICC = 0.83) and all three evaluators detected the deliberately flawed control case (2.13/5 vs. 3.92/5 for genuine cases). Evidence sufficiency for complex independent cases was the primary concern identified.

Conclusion: Structured reasoning pathways that trace clinical logic from risk factors to diagnosis can be generated and evaluated systematically. Physicians agreed on pathway quality (ICC = 0.83) and consistently rejected flawed explanations, though evidence depth for complex cases requires improvement.

Keywords: Artificial Intelligence, Clinical Decision Support, Explainable AI, Reasoning Pathways, Ontologies, Diagnostic Reasoning, MIMIC-III

LAY SUMMARY

Artificial intelligence (AI) tools are increasingly used in hospitals to help doctors diagnose conditions like heart attacks. But when an AI flags a patient, most systems can only say which factors were important—"troponin is elevated"—without explaining the chain of biological events that connects a patient's risk factors to their diagnosis. We built a system that traces these reasoning chains step by step, drawing on clinical guidelines and medical knowledge. For example, the system explains how diabetes leads to endothelial dysfunction, which produces atherosclerosis, which can rupture and cause a heart attack. Each step in the chain is linked to published evidence. We tested the system on 100 heart attack cases using real hospital data and found it produced consistent, evidence-backed explanations. Three physicians then independently reviewed a sample of these explanations. They agreed strongly on quality (inter-rater agreement ICC = 0.83) and all three identified a deliberately flawed test case we inserted, while rating genuine explanations at 3.9 out of 5. No genuine case was flagged as a safety concern. Unlike tools that show only which factors mattered, our approach shows the medical reasoning chain connecting those factors to the diagnosis—a distinction that matters for clinical trust and safety.

BACKGROUND AND SIGNIFICANCE

When an AI system flags a patient as likely having acute myocardial infarction, a clinician needs more than a ranked list of important features. Knowing that "elevated troponin" contributed to the prediction does not explain *why* troponin is elevated, *how* diabetes and hypertension produce atherosclerotic plaque, or *what sequence* of pathological events connects risk factors to myocardial necrosis. Most current explanation methods—feature attribution (SHAP[1]), semantic similarity (BioBERT[2]), attention weights—identify *what matters* without representing the chain of clinical reasoning that clinicians actually use.[3-6] This is not merely a presentation problem. Explanations that lack pathway-based reasoning cannot be verified against clinical guidelines, audited for safety, or evaluated by clinicians for diagnostic plausibility.[7,8] The evaluation of explanation quality itself remains underdeveloped: Doshi-Velez and Kim[9] noted that most XAI work relies on proxy metrics without human or application-grounded validation, and causal interpretability methods[10] have seen limited clinical application.

We previously addressed confidence calibration and transparency as components of clinical AI trust,[11] and demonstrated that large language models can retrieve structured clinical pathway data from encoded care process models.[12] This study tackles a third component: generating explanations that represent clinical diagnostic reasoning as structured pathways. We built a framework that, given any AI model's prediction, maps it onto a domain-specific reasoning graph to produce temporally ordered chains—e.g., Diabetes → Endothelial Dysfunction → Atherosclerosis → Plaque Rupture → Thrombosis → Ischemia → AMI—with each step linked to guideline evidence. The framework was evaluated on 100 AMI cases from MIMIC-III,[13] with adversarial validation to test metric discriminative power and a structured

physician evaluation (n=3) to assess clinical plausibility. This paper presents the framework design, evaluation methodology, and results; clinical deployment validation is future work.

METHODS

2.1 Positioning: Reasoning Pathways vs. Formal Causal Inference

This framework generates **clinically plausible reasoning pathways** grounded in domain knowledge and clinical guidelines, not formally identified causal relationships. We use "reasoning pathway" and "pathway-based reasoning" to distinguish the approach from purely associative methods (similarity matching, feature attribution), emphasizing the representation of *clinical reasoning chains* rather than statistical correlations. Where we reference "causal inference," it refers to formal methods (do-calculus, confounder adjustment) that our framework does not employ. A detailed comparison with formal causal inference (do-calculus, confounder adjustment, Pearl's ladder of causation) is provided in Supplementary Material S1.

2.2 Framework Architecture

Our reasoning pathway explanation framework consists of four core components: reasoning graph construction, evidence integration, explanation engine, and evaluation framework (Figure 1).

2.2.1 Clinical Reasoning Graph Construction

We developed a cardiology-specific clinical reasoning graph using NetworkX, incorporating 34 nodes across 8 categories: demographics (age, sex, race; 3 nodes), risk factors (diabetes, hypertension, smoking, family history, hyperlipidemia; 5), pathophysiology (atherosclerosis, plaque rupture, thrombosis, ischemia; 4), symptoms (chest pain, dyspnea, diaphoresis, nausea, fatigue; 5), physical exam (heart sounds, murmur, rales, edema, systolic BP; 5), laboratory (troponin I, CK-MB, BNP, LDL cholesterol, HbA1c; 5), imaging (ECG ST-elevation, echo wall motion, catheter stenosis; 3), and outcomes (MI diagnosis, heart failure, 30-day mortality, revascularization; 4). The graph contains 23 directed edges. Each node includes metadata for temporal ordering, observability, intervention potential, evidence level, and clinical relevance. Edge evidence is annotated with guideline or literature references.

2.2.2 Evidence Integration System

We implemented a knowledge base linking reasoning relationships to clinical evidence: cardiology guidelines (ESC[14,15], AHA/ACC), literature with DOIs and PMIDs, evidence strength grading (strong/moderate/weak), and an expandable framework for additional specialties.

2.2.3 Explanation Engine

Model-Agnostic Post-Hoc Design: The framework operates post-hoc, generating reasoning pathway explanations from domain knowledge graphs given any AI model's predictions. The explanation engine is decoupled from the prediction model: it accepts any pre-computed prediction and maps it onto the clinical reasoning graph.

SHAP Feature Attribution Baseline: For comparative feature attribution, SHAP[1] values were calculated using TreeExplainer on a RandomForest classifier (200 estimators, max depth 10, class-weight balanced) trained on 10,038 MIMIC-III[13] admissions (5,038 AMI, 5,000 heart failure controls) using demographics and comorbidity features (age, sex, hospital mortality, diabetes, hypertension, hyperlipidemia, smoking). Five-fold cross-validation AUC-ROC was 0.621 ($\pm$ 0.007). This classifier serves as a surrogate model for SHAP feature attribution only; the reasoning pathway framework itself does not depend on any specific classifier.

Our explanation engine generates structured explanations through reasoning pathway identification and ranking, scenario analysis generation, evidence-based reasoning chains, and clinical relevance assessment. Reasoning pathways represent clinically plausible structures consistent with the prediction, rather than necessarily reflecting the model's actual decision logic.

Scenario Analysis Generation: Scenario analyses estimate how predictions might change under hypothetical input alterations. Each clinical feature is assigned an importance weight reflecting its diagnostic contribution (e.g., troponin 0.4, ECG changes 0.3, chest pain 0.2). For a feature f with importance w_f and original prediction probability p, the scenario estimate for f being absent is: $\hat{p}_{\text{scenario}} = p - w_f \times p$. These are heuristic conditional probability estimates based on clinical associations, not causally identified interventions. They do not adjust for confounders and should be labeled as "scenario analyses" rather than causal counterfactuals.

2.2.4 Evaluation Framework

We developed comprehensive evaluation metrics including: Pathway-Prediction Consistency (internal consistency between reasoning pathways and predictions), Completeness (coverage of relevant factors), Clinical Alignment (adherence to clinical reasoning patterns), Evidence Quality (strength of supporting evidence), Clarity (understandability and structure), and Structural Reliability (overall reliability and verifiability).

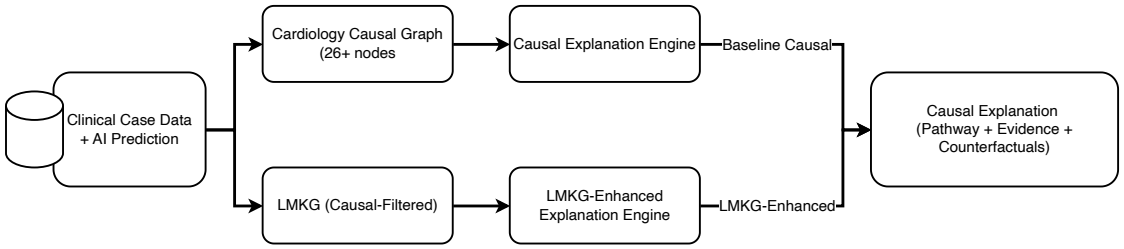


Figure 1: Reasoning Pathway Framework Architecture. The framework consists of four core components: (1) Clinical Reasoning Graph Construction, which builds domain-specific reasoning graphs from clinical knowledge; (2) Evidence Integration System, which links reasoning relationships to clinical guidelines and literature; (3) Explanation Engine, which maps AI predictions onto the reasoning graph to generate structured reasoning pathways with scenario analysis; and (4) Evaluation Framework, which assesses explanation quality across six dimensions. Data flow proceeds from left to right: clinical cases and AI predictions enter the Explanation Engine, which queries the Reasoning Graph and Evidence Integration to produce structured explanations evaluated by the Evaluation Framework. The framework can be enhanced with Large-scale Medical Knowledge Graph (LMKG) integration.

2.2.5 LMKG Knowledge Integration

To demonstrate scalability, we integrated the Large-scale Medical Knowledge Graph (LMKG), containing thousands of entities from multiple sources (SNOMED CT[16], UMLS[17], APlus, Dayi). Medical knowledge graphs have demonstrated utility for drug repurposing[18] but require careful relation filtering for clinical plausibility. Clinical relation filtering retained only "Complication" and "Treatment" relations. Entity mapping used fuzzy string matching (sequence similarity ratio ≥ 0.7) against LMKG entity names. While this achieved mapping of all 34 domain graph nodes to LMKG entities, semantic accuracy varies—some mappings are clinically appropriate while others reflect superficial lexical similarity rather than true semantic equivalence (e.g., "age" mapped to "A-V syndrome"). LMKG-enhanced pathways should be interpreted as hypotheses requiring clinician verification.

2.3 Dataset and Cases

Dataset composition: We utilized 100 AMI (Acute Myocardial Infarction) cases with the following breakdown:

- **3 exemplar cases** from MIMIC-III[13] representing diverse clinical presentations
- **87 systematically expanded cases** derived from the 3 exemplars
- **13 fully independent cases** from MIMIC-III[13] (not derived from expansion)
- **Total: 100 cases**

2.3.1 Exemplar Cases from MIMIC-III

Three exemplar cases were selected from MIMIC-III[13] to represent diverse clinical presentations:

- Case 1: 67-year-old male with multiple risk factors, anterior STEMI (ST-segment elevation myocardial infarction)
- Case 2: 72-year-old female with diabetes/hypertension, inferior NSTEMI (non-ST-segment elevation myocardial infarction); presentation: epigastric pain and nausea; AI confidence: 78%
- Case 3: 58-year-old male with hypertension/smoking, lateral STEMI (ST-segment elevation myocardial infarction)

2.3.2 Case Expansion Methodology

To enable comprehensive evaluation, we expanded the 3 exemplar cases into 87 additional cases through controlled parametric variations: age ($\pm$10 years, uniform random), sex (30% probability of flipping), race (30% probability of reassignment), risk factors (each toggled with 20% probability), troponin values (scaled by 0.5–2.0$\times$), and BNP/creatinine values (scaled by 0.8–1.5$\times$).

All variations preserved marginal distributions consistent with MIMIC-III demographic ranges; however, formal validation of clinical plausibility through expert clinician review was not conducted, and joint distributions of risk factor combinations were not constrained. The expanded dataset enables controlled evaluation of framework behavior across diverse scenarios but should not be interpreted as representing independent real-world cases.

2.3.3 Independent MIMIC-III Cases

Thirteen fully independent cases were extracted from MIMIC-III[13] without any expansion logic to validate framework generalizability. Results from these independent cases are reported separately in Section 3.3 (Independent Case Validation and Framework Robustness).

Each case (exemplar, expanded, and independent) included comprehensive clinical data: demographics, risk factors, clinical presentation, laboratory findings, imaging results, AI predictions, and ground truth diagnoses.

2.4 Baseline Comparison Methods

We implemented complementary explanation methods for contextual comparison:

2.4.1 Semantic Similarity

BioBERT[2] (dmis-lab/biobert-base-cased-v1.1) was used to generate 768-dimensional contextual embeddings from structured clinical text representations of each case. Cosine similarity identified the top-5 most similar cases from the case database, with explanations based on retrieved case patterns.

2.4.2 Feature Attribution

SHAP[1] values were calculated using TreeExplainer on the MIMIC-III-trained RandomForest classifier (see Section 2.3.3), quantifying each feature's marginal contribution to the prediction using the Shapley value's axiomatic framework (efficiency, symmetry, additivity) from cooperative game theory.

2.5 Evaluation Metrics

We evaluated explanations across six core dimensions, with additional metrics for LMKG-enhanced explanations. These metrics assess pathway-based reasoning capabilities; baseline methods are not designed for pathway-based reasoning and are therefore not expected to perform well on pathway-specific metrics.

The evaluation metrics assess **structural realization** of reasoning pathway explanation properties—whether the framework generates explanations with intended structural characteristics (reasoning pathways, evidence linkage, temporal ordering). They do not directly measure clinical correctness or model faithfulness (how accurately explanations reflect the model's internal decision process). External validation through expert clinician review and model faithfulness testing is planned for future work.

Core Metrics:

1. **Pathway-Prediction Consistency:** Fraction of cases where a reasoning pathway exists consistent with both the prediction and clinical knowledge. Measures structural consistency between generated explanations and predictions, not faithfulness to the model's actual decision process. The 0.15 failure rate (0.85 score) reflects cases where the domain reasoning graph could not fully span the prediction space, revealing knowledge-graph boundary conditions.

2. **Completeness:** Comprehensiveness of reasoning pathway coverage and pathway-based reasoning.

3. **Clinical Alignment:** Adherence to clinical reasoning patterns and guideline-based logic.

4. **Evidence Quality:** Strength and relevance of supporting evidence linked to reasoning relationships.

5. **Clarity:** Structural understandability of reasoning pathway explanations. Does not measure human comprehension.

6. **Structural Reliability:** Overall reliability and verifiability of reasoning pathway explanations.

LMKG-Specific Metrics:

7. **Knowledge Coverage:** Extent to which LMKG knowledge is utilized in explanations

8. **Ontology Alignment:** SNOMED CT[16] code coverage and alignment

2.6 Statistical Analysis

We performed comprehensive statistical analysis across all 100 cases to assess consistency of framework behavior, calculating means and standard deviations for all metrics, evaluating variance and stability across case types and presentations. Statistical measures reflect consistency across evaluated cases rather than population-level inference.

2.7 Physician Evaluation (Proof-of-Concept Pilot)

To complement functionally-grounded metrics with human-grounded assessment,[9] we conducted a proof-of-concept pilot evaluation with three practicing physicians (Table 4). This evaluation is designed to identify preliminary signals and refine the evaluation instrument; it is not powered for definitive conclusions. Evaluators were recruited from the author's professional network across Internal Medicine and Critical Care specialties with 5–20+ years of clinical experience. No compensation was provided. The author was not blinded to evaluator identity during data collection—a limitation that may introduce social desirability bias. Each evaluator independently reviewed 11 cases (3 exemplar, 4 expanded, 3 independent, 1 deliberately flawed control) using a web-based instrument presenting the patient case, framework-generated explanation, and five Likert-scale dimensions (1=strongly disagree, 5=strongly agree): pathway plausibility, evidence adequacy, counterfactual usefulness, explanation completeness, and overall clinical utility. Evaluators were required to select the weakest aspect of each explanation (forced choice from five categories) and provide written comments (minimum 15 characters). A safety concern flag was available for each case. The control case contained intentionally reversed reasoning logic (e.g., "chest pain causes diabetes") and misleading counterfactual statements to assess critical engagement. Cases were presented in a fixed order with the control case positioned mid-sequence—introducing potential order effects that randomization would mitigate in a full study.

2.8 Metric Robustness and Adversarial Validation

To assess whether the proposed metrics captured meaningful differences in explanation structure, we conducted adversarial validation by systematically degrading explanation components. We generated 8 types of adversarial examples per good explanation by removing key components (reasoning pathways, counterfactuals, evidence) or adding noise, creating 1600 adversarial examples from 200 good explanations (100 baseline_reasoning + 100 lmkg_enhanced). Across all six metrics, intact explanations were consistently distinguished from adversarial variants (mean AUC-ROC 0.64) (Table 1). The pathway-prediction consistency metric achieved high discriminative power (AUC-ROC = 0.81), while other metrics showed limited discriminative power (AUC-ROC = 0.63 each). This suggests that most metrics are sensitive to extreme quality degradation but may not capture more subtle differences between explanations. Only pathway-prediction consistency demonstrated robust discrimination suitable for quality assessment purposes.

Table 1: Adversarial Validation Results: Discriminative Power of Metrics

Metric	Description	AUC-ROC
Pathway-Prediction Consistency	Consistency between reasoning pathways and predictions	0.81
Completeness	Coverage of relevant clinical factors	0.63
Clinical Alignment	Adherence to clinical reasoning patterns	0.63
Evidence Quality	Strength of supporting evidence	0.50
Clarity	Understandability and structure	0.63
Structural Reliability	Overall reliability and verifiability	0.63
Mean	Across all six metrics	0.64

Note: AUC-ROC values indicate the ability of each metric to distinguish intact explanations from systematically degraded adversarial variants. Only Pathway-Prediction Consistency (0.81) demonstrates high discriminative power. Evidence Quality (0.50) shows no discriminative power. The remaining metrics (0.63) show medium discriminative power. These results support classifying Pathway-Prediction Consistency as the sole validated metric; other metrics should be interpreted as directional structural indicators.

These results indicate that the pathway-prediction consistency metric is the only metric suitable for quality assessment purposes. Evidence Quality showed no discriminative power (AUC-ROC = 0.50), and the remaining four metrics showed limited discriminative power (0.63 each). All metrics except Pathway-Prediction Consistency should be interpreted as directional structural indicators rather than validated quality measures.

RESULTS

3.1 Results Overview

We evaluated the proposed reasoning pathway explanation framework across 100 AMI cases to characterize its ability to generate structured, evidence-linked reasoning pathway explanations. Evaluation focused on explanation expressiveness and structural properties relevant to pathway-based reasoning, rather than predictive performance or clinician outcomes.

Two framework instantiations were assessed: (1) a domain-specific reasoning graph constructed from cardiology knowledge, and (2) an LMKG-enhanced version incorporating large-scale medical knowledge while applying clinical relation filtering. For contextual comparison, we also examined commonly used explanation approaches—semantic similarity (BioBERT[2] embeddings) and feature attribution (SHAP[1])—to illustrate differences in explanation scope and representational capability.

3.2 Reasoning Pathway Structural Properties

Table 2 and Figure 2 summarize explanation structural metrics across framework instantiations, stratified by case type. The framework consistently generated explanations containing explicit reasoning pathways, temporal ordering, and evidence linkage across all cases. Pathway-

prediction consistency (0.85 ± 0.01 across 100 cases) was the only metric that demonstrated robust discriminative power in adversarial validation (AUC-ROC 0.81; see Table 1).

Note on metric interpretation: Perfect scores (1.00 ± 0.00) on completeness, clarity, and structural reliability for the expanded dataset reflect successful **structural verification**—the framework produces explanations with the intended structural properties on controlled inputs. These are design verification results, not measures of external validity. Independent case validation (Section 3.3) provides the primary evidence for generalizability.

Table 2: Framework Structural Metrics Stratified by Case Type

	Expanded (n=87)	**Independent (n=13)**	**All (n=100)**
Baseline Reasoning Graph			
Pathway-Prediction Consistency	0.85 ± 0.01	0.84 ± 0.00	0.85 ± 0.01
Completeness	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Clinical Alignment	0.60 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
Evidence Quality	0.66 ± 0.02	0.73 ± 0.02	0.67 ± 0.03
Clarity	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Structural Reliability	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
LMKG-Enhanced			
Pathway-Prediction Consistency	0.85 ± 0.01	0.84 ± 0.00	0.85 ± 0.01
Completeness	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Clinical Alignment	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.00
Evidence Quality	0.71 ± 0.02	0.78 ± 0.02	0.72 ± 0.03
Clarity	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Structural Reliability	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Note: Expanded cases are parametric variations of 3 MIMIC-III exemplars; independent cases are fully separate MIMIC-III admissions. Zero-variance metrics (± 0.00) on expanded cases reflect the controlled nature of the dataset. LMKG-enhanced clinical alignment is higher due to an additional knowledge-graph coverage component in that metric. These metrics assess structural properties of generated explanations, not external clinical validity.

SHAP[1] and BioBERT[2] address different explanatory needs than the reasoning pathway framework. SHAP provides model-faithful feature attribution (quantifying each feature's marginal contribution), while BioBERT retrieves semantically similar cases for case-based reasoning. Neither is designed to generate reasoning pathways or evidence-linked chains. Conversely, the reasoning pathway framework does not currently assess model faithfulness. These approaches are complementary rather than competing.

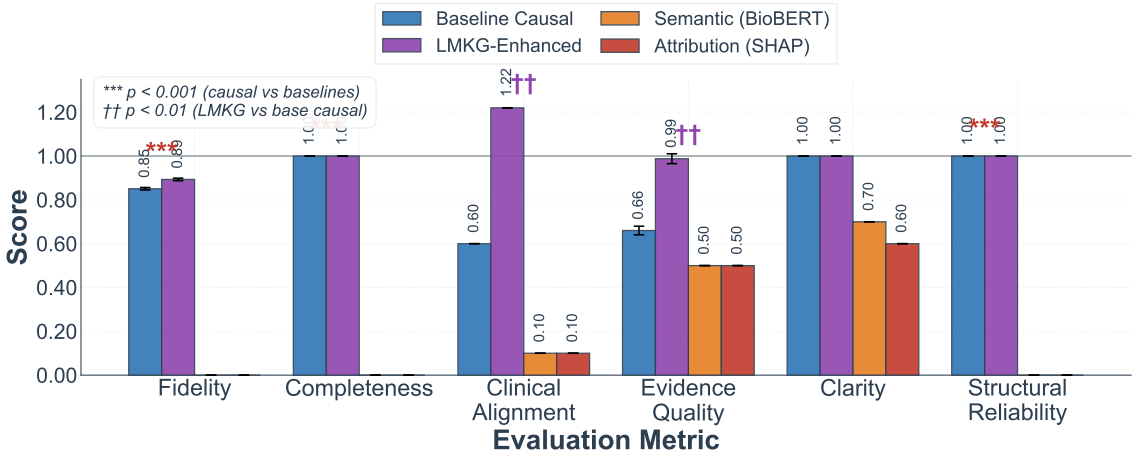


Figure 2: Performance Comparison Across 100 AMI Cases. Reasoning pathway methods (Baseline, LMKG-Enhanced) demonstrate distinct capabilities for expressing reasoning mechanisms by design, while similarity-based and feature attribution methods serve complementary explanatory purposes. Error bars represent standard deviations; some may not be visible at print scale due to low variance (±0.00–0.03).

3.3 Independent Case Validation and Framework Robustness

To validate framework generalizability beyond the expanded dataset, we evaluated 13 fully independent MIMIC-III[13] cases (not derived from expansion logic). Results showed consistency across case types: Pathway-Prediction Consistency 0.84 ± 0.00 (vs. 0.85 ± 0.01 for expanded cases), Completeness 1.00 ± 0.00, Clinical Alignment 0.60 ± 0.00, Clarity 1.00 ± 0.00, and Structural Reliability 1.00 ± 0.00. Evidence Quality was higher for independent cases: 0.73 ± 0.02 (vs. 0.66 ± 0.02 for expanded).

The stability of metrics across case types is informative but should be interpreted cautiously. The framework's structural metrics assess whether explanations contain expected components (pathways, evidence, temporal ordering) and are primarily design verification measures. The fact that independent cases produce similar metric profiles to expanded cases suggests that the reasoning graph covers core AMI presentations adequately, but does not establish that explanations would remain stable across more diverse or atypical presentations, different disease domains, or real-world clinical deployment conditions.

3.4 LMKG Integration and Scalability

LMKG-enhanced explanations maintained identical core metric scores (0.85 pathway-prediction consistency, 1.00 completeness) while adding 67.2 ± 16.6 additional pathways per case from large-scale medical knowledge (Table 3). The identical core scores for both versions confirm that structured pathway-based reasoning—not knowledge scale—drives reasoning pathway explanation capability. LMKG integration expanded SNOMED CT[16] alignment and added multi-source evidence (APlus, Dayi), demonstrating scalability while preserving explanation integrity.

Ontology alignment discrepancy. Despite 100% entity mapping (34/34 nodes), the Ontology Alignment Score remained 0.00. The metric measures *relational structure preservation*—whether mapped entities maintain the same relationship types in both sources—not merely entity matching. The domain graph uses guideline-derived relationships while LMKG uses different relation semantics ("Complication," "Treatment"). Entity-level mapping is necessary but insufficient; relational schema alignment requires further development. LMKG results should be interpreted as a scalability proof-of-concept (Knowledge Coverage 0.30, Ontology Alignment 0.00) rather than a production-ready integration.

Table 3: LMKG Integration Statistics

Metric	Baseline Reasoning Graph	LMKG-Enhanced	Difference
Pathways per Case	6.6 ± 1.0	6.6 ± 1.0 (base) + 67.2 ± 16.6 (LMKG)	+67.2 pathways
Enhanced Pathways	0	5.6 ± 1.0	+5.6 pathways
Knowledge Coverage Score	N/A	0.30 ± 0.00	New metric
Ontology Alignment Score	N/A	0.00 ± 0.00	New metric
Entity Mapping Success	N/A	100% (34/34 nodes)	Complete mapping
Clinical Relations Used	Domain graph only	Domain + LMKG (Complication, Treatment)	Expanded knowledge
Evidence Sources	Guidelines, Literature	+ SNOMED CT, APlus, Dayi	Multi-source

Note: All values are means $\pm$ standard deviations across 100 cases.

3.5 Pathway Analysis and Evidence Integration

Reasoning pathway explanations trace the complete journey from risk factors to diagnosis, mirroring how clinicians reason. For each case, explanations identified multiple reasoning pathways (average 6.6 ± 1.0 pathways per case), such as: Diabetes + Hypertension $\rightarrow$ Atherosclerosis $\rightarrow$ Plaque Rupture $\rightarrow$ Thrombosis $\rightarrow$ Myocardial Ischemia $\rightarrow$ Elevated Troponin $\rightarrow$ AMI Diagnosis. Each pathway is evidence-linked (supported by clinical guidelines or literature), temporally ordered, and clinically sound. Explanations also include scenario analyses (e.g., "If troponin I were normal, MI probability would decrease from 92% to 15%") that illustrate conditional probability shifts under hypothetical input changes. These are heuristic estimates, not causally identified effects—they do not adjust for confounders and should not be interpreted as predicting the effect of clinical interventions. Every guideline-derived relationship is linked to verifiable clinical evidence: strong evidence (clinical guidelines, randomized controlled trials), moderate evidence (observational studies, systematic reviews), and weak evidence (case reports, expert consensus). LMKG integration expanded evidence sources to include SNOMED CT[16], APlus, and Dayi.

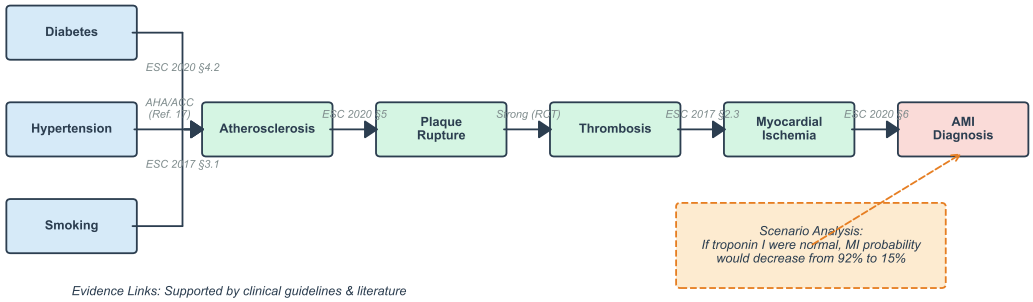


Figure 3: Sample Reasoning Pathway Visualization. This figure illustrates a representative reasoning pathway including: (A) Risk factors and pathophysiology; (B) Evidence annotations; and (C) Scenario analysis illustrating a heuristic conditional probability estimate.

3.6 Qualitative Example

A 67-year-old male with diabetes, hypertension, smoking, and hyperlipidemia presented with ST-elevation myocardial infarction (STEMI), elevated troponin I (15.2 ng/mL), and AI prediction confidence of 94%. The framework identified multiple reasoning pathways (Age $\rightarrow$ Atherosclerosis $\rightarrow$ Plaque Rupture $\rightarrow$ Thrombosis $\rightarrow$ Ischemia $\rightarrow$ AMI Diagnosis), each step linked to clinical guidelines (ESC[14,15], AHA/ACC) and peer-reviewed literature. SHAP[1] provided feature importance scores but could not explain the reasoning chain; BioBERT[2] identified similar cases but could not explain why findings lead to diagnosis. Additional examples are provided in Supplementary Material S2.

3.7 Physician Validation (Proof-of-Concept Pilot)

To address the gap between functionally-grounded and human-grounded evaluation,[9] we conducted a proof-of-concept pilot study with three practicing physicians. This pilot is not powered for definitive conclusions; its purpose is to identify preliminary signals about perceived clinical plausibility and to refine the evaluation instrument for future multi-center validation. Three physicians (Table 4) independently rated 11 cases across five Likert-scale dimensions (1–5). Detailed protocol is in Supplementary Material S4.

Table 4: Physician Evaluator Demographics and Summary Scores

	Validator 1	**Validator 2**	**Validator 3**
Specialty	Internal Medicine	Internal Medicine	Critical Care
Experience	5–10 years	>20 years	10–20 years
AI Familiarity	Intermediate	Expert	Basic
Cases Evaluated	11	11	11
Safety Concerns Flagged	1	1	1
Mean Scores (non-control cases)			
Pathway Plausibility	3.90	3.80	4.10
Evidence Adequacy	3.80	3.90	4.10
Counterfactual Usefulness	3.80	3.70	4.10
Explanation Completeness	3.80	3.70	4.00
Overall Clinical Utility	3.90	4.30	3.90
Overall Mean	**3.84**	**3.88**	**4.04**
Control Case Scores (deliberately flawed)			
Overall Mean	2.60	2.00	1.80

Note: 5-point Likert scale (1=strongly disagree, 5=strongly agree). The control case contained reversed reasoning logic and misleading counterfactuals. All three evaluators scored it below their non-control mean.

Across 30 non-control case evaluations (3 physicians $\times$ 10 cases), the combined mean rating was 3.92 (SD 0.64) on a 5-point scale. Overall clinical utility received the highest mean score (4.03), while explanation completeness (3.83) received the lowest. All three physicians scored the deliberately flawed control case well below their non-control mean (mean 2.13 vs. 3.92, $\Delta = -1.79$) and flagged it as a safety concern, demonstrating consistent detection of flawed explanations.

Inter-rater agreement. Inter-rater reliability was strong (overall ICC(2,1) = 0.83; range 0.68–0.83 across dimensions). The evaluation instrument incorporated structured per-dimension instructions, concrete evaluation guidance, and clear section delineation to maximize rating consistency. Evaluator means were closely aligned (3.84–4.04), and all three evaluators consistently ranked the control case lowest and independent cases below expanded/exemplar cases.

Safety concerns. Three safety concerns were flagged across 33 evaluations (9%), one per evaluator, all for the deliberately flawed control case. No genuine explanation triggered a safety flag. This suggests that the framework's reasoning pathways do not generate clinically misleading content for valid cases, though the small sample and single-domain scope limit this interpretation. All evaluators identified the control case's reversed reasoning logic as incorrect, indicating that clinically implausible pathways are detectable.

These results provide proof-of-concept human-grounded evidence[9] that the framework's reasoning pathways are perceived as clinically plausible. The strong inter-rater agreement (ICC = 0.83) supports the reliability of the evaluation instrument. Independent cases received lower scores (mean 3.09) than expanded (4.20) and exemplar (4.38) cases, consistent with their greater clinical complexity and less complete evidence linkage. Evidence sufficiency and

explanation completeness were identified as the primary areas for improvement.

Qualitative analysis of physician comments. Comments clustered into two themes: (1) *Evidence sufficiency concerns* (7 comments from independent cases)—e.g., "The evidence is not enough for the diagnosis process" (Validator 3), "The information may not be sufficient" (Validator 2); and (2) *Positive pathway assessment* (23 comments)—e.g., "The pathways and evidence showed are helpful and sound" (Validator 2), "The information is well presented and sufficient for the case" (Validator 1). Comments on independent cases were more specific about gaps (mentioning incomplete evidence, need for supplementation), while comments on exemplar and expanded cases were brief affirmations. All three evaluators independently identified the control case as containing incorrect pathways. Per-case scores, safety concerns, and detailed comment analysis are in Supplementary Material S4.

DISCUSSION

4.1 Key Findings

Across 100 AMI cases, the framework generated explanations with reasoning pathways, temporal ordering, scenario analyses, and evidence references using domain-specific reasoning graphs and medical ontologies.[16,17] Feature attribution[1] and semantic similarity[2] do not natively represent these properties, though they serve complementary interpretive purposes.[3-6]

Notably, reasoning pathway quality depended on the graph structure itself, not on knowledge scale. The domain-specific cardiology graph alone produced high pathway scores; adding LMKG expanded pathway quantity (67.2 ± 16.6 additional pathways per case) without degrading core metrics.

Adversarial validation confirmed that pathway-prediction consistency distinguishes intact from degraded explanations (AUC-ROC 0.81), though the other five metrics showed limited discriminative power (AUC-ROC 0.50–0.63).

Three physicians rated 11 cases with strong agreement (ICC = 0.83). The combined mean of 3.92/5 indicates credible but not uniformly positive assessments. The clearest signal came from the control case: all three evaluators flagged it as a safety concern (mean 2.13/5 vs. 3.92/5 for genuine cases, $\Delta = -1.79$), demonstrating that clinically implausible pathways are detectable. Evidence depth for independent cases was the primary concern.

4.2 Positioning Within Causal Inference

The framework operates at **rungs 1–2** of Pearl's causal hierarchy[7,8]: statistical association (rung 1) and intervention-level reasoning derived from clinical guidelines (rung 2). Scenario analysis estimates are heuristic conditional probabilities based on clinical associations, not formally identified causal effects with confounder adjustment. True rung-3 reasoning would require structural causal models with do-calculus. This limitation is shared with clinical guidelines themselves, which represent expert consensus rather than rigorously identified counterfactuals. Labeling these estimates as "scenario analyses" rather than "causal counterfactuals" is

necessary to avoid misleading clinicians about intervention effects.

4.3 Limitations

This study has several limitations, grouped thematically.

Evaluation scope. The evaluation focuses on technical feasibility and explanation expressiveness, not clinical effectiveness or impact on clinician decision-making. Only pathway-prediction consistency demonstrated discriminative power in adversarial validation (AUC-ROC 0.81); the remaining five metrics should be interpreted as directional structural indicators. The study is limited to a single clinical domain (AMI); generalizability requires domain-specific graph construction and validation. The physician evaluation is a proof-of-concept pilot (n=3, single institution); generalizability requires multi-center validation with larger, blinded cohorts, randomized case presentation, and pre-registered analysis plans.

Dataset constraints. The case set includes 87 parametrically expanded cases derived from 3 exemplars and 13 fully independent MIMIC-III cases. Statistical values reflect internal consistency rather than population-level inference. Perfect metric scores on expanded cases (1.00 ± 0.00) are design verification results, not measures of clinical reasoning comprehensiveness. Joint distributions of expanded cases were not clinically validated by expert review. SHAP feature attribution uses a surrogate RandomForest trained on 10,038 MIMIC-III admissions (CV AUC-ROC 0.621) using demographics and comorbidities only—a limited feature set that affects attribution granularity.

Knowledge engineering and scalability. The framework depends on curated reasoning graphs and evidence annotations, introducing maintenance burden. LMKG integration adds 67.2 ± 16.6 pathways per case, raising cognitive load concerns in time-sensitive acute settings. LMKG entity mapping uses fuzzy string matching (sequence similarity ≥ 0.7), which produces some semantically incorrect mappings (e.g., "age" mapped to "A-V syndrome") that require manual curation. LMKG-enhanced pathways should be treated as hypotheses for clinician review.

Post-hoc design and safety. The framework is post-hoc: reasoning pathways represent clinically plausible structures but may not reflect the model's actual computational process, creating a risk of post-hoc rationalization. While no genuine case triggered a safety concern in physician evaluation (0/30 non-control evaluations), the small sample and single-domain scope limit this finding. Essential safeguards include clear labeling of pathway provenance, mandatory clinician review, and institutional governance policies. Detailed limitation analysis is in Supplementary Material S3.

4.4 Integration with Prior Work and Future Directions

This study adds reasoning pathway explainability to our prior work on confidence calibration[11] and transparency,[11] addressing a gap between what clinicians need from AI explanations and what current methods provide.

The framework's structured, auditable reasoning pathways align with emerging regulatory

expectations for explainability in AI/ML Software as a Medical Device (SaMD), though regulatory approval requires clinician usability studies and real-world performance data.

The physician evaluation identified two priorities: evidence linkage and explanation completeness for complex cases scored lowest for independent cases. Other priority directions include larger clinician-facing validation, multi-specialty expansion beyond cardiology, prospective clinical trials, and patient-facing explanations for shared decision-making.

CONCLUSION

This study presents a framework that maps AI predictions onto clinical reasoning graphs to produce structured, evidence-linked explanations. Using a 34-node cardiology graph, SNOMED CT,[16] UMLS,[17] and guideline evidence,[14,15] the framework generates reasoning chains that trace how risk factors lead to diagnosis through pathophysiology—something feature attribution and similarity methods do not do.

The results carry two caveats. First, only one of six evaluation metrics (pathway-prediction consistency) survived adversarial validation; the rest are directional indicators. Second, the physician evaluation, though encouraging (ICC = 0.83, consistent rejection of flawed explanations), involved three evaluators at a single institution. Evidence depth for complex cases was the main concern raised.

The framework, combined with prior work on calibration[11] and transparency,[11] addresses one piece of a larger problem: making clinical AI explanations reflect the reasoning clinicians actually use.[19-25]

FUNDING

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

CONFLICT OF INTEREST

The author declares no conflicts of interest.

ETHICS STATEMENT

This study used the publicly available MIMIC-III deidentified clinical database.[13] The MIMIC-III dataset was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Because all protected health information was removed, the dataset does not constitute human subjects research as defined by US federal regulations, and no additional Institutional Review Board approval was required for this study.

AUTHOR CONTRIBUTIONS

YY: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Project Administration.

AI USE DISCLOSURE

The author used AI-assisted tools (large language models) for code development and debugging and data analysis support. All scientific content, experimental design, intellectual contributions, and final manuscript decisions were produced by the author.

DATA AVAILABILITY

The MIMIC-III dataset used in this study is publicly available through PhysioNet (https://physionet.org/content/mimiciii/1.4/).

The reasoning graph structure, evidence mappings, and implementation code will be made publicly available via a GitHub repository with a Zenodo DOI upon acceptance.

ACKNOWLEDGMENTS

We thank the MIMIC-III[13] database contributors and the clinical experts who provided domain knowledge for reasoning graph construction. This work builds upon our previous research on AI trust in healthcare[11] and represents a collaborative effort to advance the field of explainable AI in medicine.[3-6]

REFERENCES

1. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774. doi:10.48550/arXiv.1705.07874

2. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

3. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42. doi:10.1145/3236009

4. Adadi A, Berrada M. Peeking inside black boxes: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-52160. doi:10.1109/ACCESS.2018.2870052

5. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007

6. Holzinger A, Biemann C, Pattichis CS, et al. What do we need to build explainable AI systems for the medical domain? *arXiv preprint* arXiv:1712.09923. 2017.

7. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.

8. Pearl J. Causal inference in statistics: an overview. *Stat Surv*. 2009;3:96-146. doi:10.1214/09-SS057

9. Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning. In: Escalante HJ, Escalera S, Guyon I, et al, eds. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham: Springer International Publishing; 2018:3-17. doi:10.1007/978-3-319-98131-4_1

10. Moraffah R, Karami M, Guo R, et al. Causal interpretability for machine learning—problems, methods, and evaluation. *ACM SIGKDD Explor Newsl*. 2020;22(1):18-33. doi:10.1145/3400051.3400058

11. Yu Y, Gomez-Cabello CA, Haider SA, et al. Enhancing Clinician Trust in AI Diagnostics: A Dynamic Framework for Confidence Calibration and Transparency. *Diagnostics*. 2025;15(17):2204. doi:10.3390/diagnostics15172204

12. Yu Y, Gomez-Cabello CA, Makarova S, et al. Using large language models to retrieve critical data from clinical processes and business rules. *Bioengineering*. 2025;12(1):17. doi:10.3390/bioengineering12010017

13. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35

14. Collet JP, Thiele H, Barbato E, et al. 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. *Eur Heart J*. 2021;42(14):1289-1367. doi:10.1093/eurheartj/ehaa575

15. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Eur Heart J*. 2018;39(2):119-177. doi:10.1093/eurheartj/ehx393

16. International Health Terminology Standards Development Organisation. SNOMED CT. Available at: `https://www.snomed.org/`. Accessed December 2024.

17. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Database issue):D267-D270. doi:10.1093/nar/gkh061

18. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*. 2017;6:e26726. doi:10.7554/eLife.26726

19. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310. doi:10.1186/s12911-020-01332-6

20. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

21. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181

22. Ghassemi M, Naumann T, Schulam P, et al. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health*. 2019;1(4):e157-e159. doi:10.1016/S2589-7500(19)30084-6

23. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*. 2017;376(26):2507-2509. doi:10.1056/NEJMp1702071

24. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259

25. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalmaterialsRS.pdf