Supplementary Materials: A Reasoning Pathway Explanation Framework for Clinical AI

Yunguo Yu

Supplementary Material S1: Reasoning Pathways vs. Formal Causal Inference — Detailed Comparison

This framework generates **clinically plausible reasoning pathways** based on domain knowledge and clinical guidelines, not formally identified causal relationships using rigorous causal inference methods. We use the term "reasoning graph" to describe our domain-specific knowledge structure because it represents clinical understanding of disease mechanisms documented in medical literature. Key differences from formal causal inference include:

1. **Knowledge source:** Our reasoning graph encodes clinical reasoning patterns from guidelines and literature, not causal relationships identified through do-calculus or structural causal models with confounder adjustment.

2. **LMKG filtering:** Selecting relations labeled "Complication" or "Treatment" provides clinical plausibility but does not establish causal directionality or eliminate confounding.

3. **Counterfactuals:** Our counterfactual estimates are heuristic probability changes based on clinical associations, not interventions identified through formal causal inference.

4. **Pearl's ladder of causation:** This work operates primarily at the **association and intervention levels** (rungs 1–2), providing guideline-aligned reasoning pathway explanations, rather than the counterfactual level (rung 3) requiring rigorous causal identification.

We use "reasoning pathway explanation framework" to distinguish our approach from purely associative methods (similarity matching, feature attribution), emphasizing that we represent *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. Future work will explore integration of rigorous causal inference techniques.

Supplementary Material S2: Additional Qualitative Examples

Example 2: Inferior NSTEMI with Atypical Presentation

A 72-year-old female with diabetes and hypertension presented with epigastric pain and nausea rather than classic chest pain. AI prediction confidence was 78%. The framework identified reasoning pathways including: Diabetes → Endothelial Dysfunction → Atherosclerosis → Plaque Rupture → Partial Thrombosis → Myocardial Ischemia → Elevated Troponin → NSTEMI Diagnosis. The atypical presentation (epigastric pain) was linked through vagal referral patterns with moderate evidence. SHAP identified "troponin elevation" and "diabetes" as top features but could not explain the reasoning chain connecting them. BioBERT identified similar cases with epigastric presentations but could not explain why diabetes contributes to NSTEMI through pathological pathways. (See main text references 15, 16 for BioBERT and SHAP descriptions.)

Example 3: Lateral STEMI in Younger Patient

A 58-year-old male with hypertension and active smoking presented with lateral STEMI. The framework identified pathways including: Smoking → Endothelial Injury → Atherosclerosis → Plaque Rupture → Complete Thrombosis → Myocardial Ischemia → ST Elevation → STEMI Diagnosis. Each step was linked to specific guideline evidence (ESC 2017, AHA/ACC). Scenario analysis estimated that smoking cessation would reduce MI probability from 88% to 35%—this is a heuristic estimate based on importance-weight subtraction, not an epidemiologically derived risk reduction. For comparison, prospective cohort studies report that smoking cessation reduces MI risk by approximately 50–65% within 1–3 years (Petersen et al., *Stroke*. 2008;39(4):1116-1121). The specific percentage produced by our framework should be interpreted as illustrative rather than evidence-based.

Supplementary Material S3: Detailed Limitation Analysis

S3.1 Evaluation Scope

- The evaluation focuses on technical feasibility and explanation expressiveness, not clinical effectiveness or impact on clinician decision-making. Such outcomes require clinician-facing studies and prospective validation.
- Only pathway-prediction consistency demonstrated discriminative power in adversarial validation (AUC-ROC 0.81). The remaining five metrics serve as directional structural indicators.

- The study is limited to a single clinical domain (AMI). Generalizability to other specialties requires domain-specific reasoning graph construction and validation.
- The physician evaluation is a proof-of-concept pilot (n=3, single institution, no blinding, fixed case order). Generalizability requires multi-center validation with larger cohorts.

S3.2 Dataset Constraints

- The case set includes 87 parametrically expanded cases derived from 3 exemplars and 13 fully independent MIMIC-III cases. Expansion used controlled variations ($\pm$10 years age, 30% sex/race flip, 20% risk factor toggle, 0.5–2.0$\times$ troponin scaling). Formal validation of clinical plausibility through expert clinician review was not conducted, and joint distributions of risk factor combinations were not constrained.
- Statistical values reflect internal consistency rather than population-level inference.
- Perfect metric scores on expanded cases (1.00$\pm$0.00) are design verification results—the framework produces explanations with intended structural properties on controlled inputs. They should not be interpreted as measures of clinical reasoning comprehensiveness.
- SHAP feature attribution uses a surrogate RandomForest trained on 10,038 MIMIC-III admissions (5,038 AMI, 5,000 heart failure; CV AUC-ROC 0.621) using demographics and comorbidities only. This limited feature set affects attribution granularity.

S3.3 Knowledge Engineering and Scalability

- The framework depends on curated reasoning graphs and evidence annotations, introducing knowledge engineering dependencies and maintenance burden.
- LMKG integration adds 67.2$\pm$16.6 pathways per case, raising cognitive load concerns in time-sensitive acute settings. No human-in-the-loop validation of cognitive burden was conducted.
- LMKG entity mapping uses fuzzy string matching (sequence similarity $\geq$ 0.7), which produces some semantically incorrect mappings requiring manual curation.
- LMKG clinical relation filtering provides clinical plausibility but does not guarantee true causal relationships.

S3.4 Post-Hoc Design and Ethical Considerations

- 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. Faithfulness testing is planned.

- 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. Mandatory clinician review before clinical deployment remains essential.
- Ethical and liability questions arise when LMKG-hypothesized pathways (not guideline-validated) influence clinical decisions. Current medical liability frameworks are not designed for AI-generated clinical reasoning.
- Essential safeguards: (1) Clear labeling distinguishing guideline-validated vs. LMKG-hypothesized pathways; (2) Mandatory clinician review before clinical use; (3) Explicit documentation that LMKG pathways are hypotheses requiring verification; (4) Institutional governance policies for AI-assisted decision-making.

S3.5 Reproducibility and Transparency

To support open science principles, we will deposit the reasoning graph, evidence mappings, and code in a public repository (GitHub + Zenodo) with a reserved DOI prior to manuscript submission. This ensures immediate reproducibility and transparency. Ethical and regulatory considerations (curation, validation, error correction, liability) require explicit discussion and should be addressed in future work. A clinician-in-the-loop review step is essential before decision reliance on LMKG-hypothesized pathways.

Supplementary Material S4: Physician Evaluation Per-Case Scores

Table 1 reports individual case-level mean scores from each physician evaluator. The control case (AMI_CTRL_001) and independent cases (AMI_INDEP_004, AMI_INDEP_012) were consistently scored below exemplar and expanded cases across all three evaluators.

Table 1: Per-Case Mean Scores by Physician Evaluator

Case ID	Type	Validator 1	Validator 2	Validator 3	Mean
AMI_001	Exemplar	5.00	4.20	5.00	4.73
AMI_002	Exemplar	4.00	4.20	5.00	4.40
AMI_003	Exemplar	4.00	4.00	4.00	4.00
AMI_007	Expanded	4.00	4.00	4.80	4.27
AMI_018	Expanded	4.00	4.40	4.60	4.33
AMI_085	Expanded	4.20	4.20	4.00	4.13
AMI_098	Expanded	4.00	4.20	4.00	4.07
AMI_INDEP_004	Independent	3.00	3.00	3.00	3.00
AMI_INDEP_005	Independent	3.00	3.60	3.00	3.20
AMI_INDEP_012	Independent	3.20	3.00	3.00	3.07
AMI_CTRL_001	**Control**	**2.60**	**2.00**	**1.80**	**2.13**

Table 2: *

Note: Values are mean scores across five Likert dimensions (1–5). The control case (bold) contained deliberately reversed reasoning logic and misleading counterfactuals. It was scored well below the non-control mean by all three evaluators and flagged as a safety concern by each. Independent cases (3.00–3.20) scored lower than expanded (4.07–4.33) and exemplar (4.00–4.73) cases.

Three safety concerns were flagged across 33 evaluations (9%), one per evaluator, all for the deliberately flawed control case. No genuine case triggered a safety concern. Table 3 summarizes the flagged cases.

Table 3: Safety Concerns Flagged by Physician Evaluators

Evaluator	Case	Comment
Validator 1	AMI_CTRL_001	The pathway may be incorrect.
Validator 2	AMI_CTRL_001	The information (pathway) may have serious errors.
Validator 3	AMI_CTRL_001	Some of the pathway are incorrect

Table 4: *

Note: All three evaluators independently identified the deliberately flawed control case as containing incorrect reasoning pathways. No genuine case triggered a safety concern.

Physician Comments. Table 5 presents all evaluator comments by case.

Table 5: Physician Evaluator Comments by Case

Evaluator	Case	Comment
Validator 1	AMI_001	The information is well presented and sufficient for the case.
Validator 1	AMI_002	The evidence and information are fine.
Validator 1	AMI_003	Evidence is well presented and the pathways are fine.
Validator 1	AMI_007	The evidence are well presented and sufficient for the case.
Validator 1	AMI_018	Well presented.
Validator 1	AMI_085	The information are well organized and sufficient for the diagnosis.
Validator 1	AMI_CTRL_001	The pathway may be incorrect.
Validator 1	AMI_098	The information are correct and sufficient.
Validator 1	AMI_INDEP_004	The information and evidence may not be sufficient.
Validator 1	AMI_INDEP_005	The pathways are correct, but may not be complete. Need more information.
Validator 1	AMI_INDEP_012	The evidence may need supplement.
Validator 2	AMI_001	The pathways and evidence showed are helpful and sound.
Validator 2	AMI_002	The evidence, especially the lab and ECG show significance evidence of MI.
Validator 2	AMI_003	The information provided is sufficient and useful.
Validator 2	AMI_007	The information is helpful and sufficient.
Validator 2	AMI_018	The information is useful and indicative.
Validator 2	AMI_085	The clinical information are well presented and is helpful.
Validator 2	AMI_CTRL_001	The information (pathway) may have serious errors.
Validator 2	AMI_098	The information is helpful for clinical diagnostic process.
Validator 2	AMI_INDEP_004	The information may not be sufficient.
Validator 2	AMI_INDEP_005	The information is useful but may have some gap: insufficient information.
Validator 2	AMI_INDEP_012	May need more information.
Validator 3	AMI_001	The information is useful and sufficient.
Validator 3	AMI_002	The information provided is useful and sufficient in this case.
Validator 3	AMI_003	The information is fine and sufficient.
Validator 3	AMI_007	The information provided is useful and help to the diagnosis process.
Validator 3	AMI_018	Evidence presented are sound.
Validator 3	AMI_085	The evidence and information presented are sufficient for this case of diagnosis.
Validator 3	AMI_CTRL_001	Some of the pathway are incorrect.
Validator 3	AMI_098	The information are fine and sufficient.
Validator 3	AMI_INDEP_004	The evidence is not enough for the diagnosis process.
Validator 3	AMI_INDEP_005	The information presented may not in full and may need expansion.
Validator 3	AMI_INDEP_012	The evidence may not be sufficient for the case.