

A Closed-Loop Framework for Knowledge-Based Causal Model Refinement Using Process Mining and Human-AI Governance

Eduardo Illueca Fernandez

`eduardo.illueca@ki.se`

Karolinska Institutet

Kaile Chen

Karolinska Institutet

Carlos Fernandez Llatas

Universitat Politècnica de València

Fernando Seoane

Karolinska Institutet

Farhad Abtahi

Karolinska Institutet

Research Article

Keywords: causal inference, directed acyclic graphs, process mining, human-AI collaboration, real-world evidence, disease trajectories, longitudinal data

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10371372/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Closed-Loop Framework for Knowledge-Based Causal Model Refinement Using Process Mining and Human–AI Governance

Eduardo Illueca-Fernandez^{1*}, Kaile Chen^{1,2}, Carlos Fernandez-Llatas^{1,3},
Fernando Seoane^{1,4,5,6}, Farhad Abtahi^{1,2,4}

¹Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, 17177, Sweden.

²Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Huddinge, 14157, Sweden.

³SABIEN-ITACA, Universitat Politècnica de València, Valencia, 46022, Spain.

⁴Department of Clinical Physiology, Karolinska University Hospital, Stockholm, 17176, Sweden.

⁵Department of Medical Technology, Karolinska University Hospital, Stockholm, 17176, Sweden.

⁶Department of Textile Technology, University of Borås, Borås, 50190, Sweden.

*Corresponding author(s). E-mail(s): eduardo.illueca@ki.se;

Abstract

Background: Directed acyclic graphs (DAGs) are a graph representation used in epidemiology for analysing causal inference, based on expert inputs and previous experiences. However, the lack of empirical evidence implies uncertainty in the causal structures, specially in chronic and complex disease with competing risks and temporal dynamics. This DAG uncertainty can bias the effect estimation regardless the methodological robustness. For this reason, we present a closed-loop framework that consider knowledge-based DAGs as scientific hypotheses, which are iteratively tested through governed human–AI collaboration loop based on process mining.

Methods: Our framework comprises four different phases: i) generating a knowledge-driven event log using LLMs to produce a preliminary DAG from literature and expert input; ii) applying data-driven process discovery algorithms using the longitudinal event sequences as input data for producing an empirical direct follow graph (DAG); iii) iterating over the DAG refinement by comparing the empirical patterns in the DFG against the knowledge-based DAG, with human expert adjudication of each proposed change; and iv) statistical modeling and assessment using the final DAG. We operationalised the framework through an open-source interactive tool, and demonstrated its performance using synthetic data calibrated to a published cohort, comparing proton pump inhibitor (PPI) and histamine-2 receptor blocker (H2B) users.

Results: After testing our framework against the synthetic dataset, the workflow reproduced longitudinal patterns in which CKD progression preceded cardiovascular adverse events, illustrating how the framework would surface a candidate mediator pathway for expert adjudication. Formal mediation analysis would be required to test the mediator hypothesis. Competing mortality rates illustrated how high mortality transitions would prompt expert review of competing-event handling in the analytical model. The AI tool recommended Fine–Gray competing-risk models, causal mediation analysis, and propensity-score balancing, all consistent with epidemiological best practice.

Conclusions: The closed-loop framework provides a governed, transparent, and reproducible approach for integrating AI assistance into causal inference workflows. By embedding empirical appraisal into DAG construction while preserving human authority over causal interpretation, the framework advances methodological discipline in real-world evidence generation.

Keywords: causal inference, directed acyclic graphs, process mining, human-AI collaboration, real-world evidence, disease trajectories, longitudinal data

1 Introduction

Epidemiological causal inference relies on accurately specifying causal relationships among variables [1, 2] by using regression-based methods, and guided by graphical representations of causal assumptions. In this context, directed acyclic graphs (DAGs) encode these assumptions as nodes (variables) connected by directed edges (assumed causal effects) and guide the selection of statistical analyses targeting causal estimands of interest [3–5]. In consequence, it is easy to note that a well-specified DAG can help researchers to infer causality, while a non-accurate DAG can lead to biased conclusions [6–8]. Thus, the challenge arises when causal structures are not clear, a common scenario in complex diseases where multiple interacting conditions, time-varying exposures and competing risks can be found, creating ambiguity when defining the DAG [9, 10].

Futhermore, healthcare systems generate vast longitudinal datasets through electronic health records (EHRs), administrative claims databases, disease registries, and patient-reported outcomes. While randomized controlled trials (RCTs) remain the gold standard for establishing efficacy under controlled conditions, real-world data provide complementary evidence on effectiveness, safety, long-term outcomes, and treatment patterns in diverse, unselected populations reflecting actual clinical practice [11]. However, the same characteristics that make RWD valuable also present remarkable methodological challenges, as causal uncertainty increases with data size. For instance, temporal complexity arises from irregular visit times, varying follow-up lengths, and missing episodes [12]. On the other hand, the presence of competing risks in RWD means that multiple outcomes may prevent the observation of the target effect. Last, clinical heterogeneity and trajectory variation across diverse populations may make difficult to extract disease progression by using data agregation techniques [13].

Aligned with the methodology proposed by Dr. John Snow [14], causal inference from big data is increasingly recognised as a falsifiable and iterative scientific methodology in which hypothesis are continuously tested, refined, and assessed against real-world evidence [15]. This perspective is helping to bridge the gap between clasical epidemiology and empirical discovery, emphasizing that causal understanding emerges through a continuous dialogue between conceptual models and observational data. In this approach, the initial assumptions based on domain expertise and prior literature are explicitly formulated as testable hypotheses using directed acyclic graphs (DAGs). Then, these hypotheses can be examined by applying process mining on longitudinal data to extract sequential patterns that either support or refute the proposed hypothesis [16]. These empirical patterns can be formally represented using directly-follows graphs (DFGs), which capture observed transitions between events in longitudinal data. In this framework, the DFG serves as an empirical counterpart to the knowledge-based DAG defined previously, enabling systematic comparison between hypothesized causal structures and observed temporal patterns. Discrepancies between the hypothesized and observed structures prompt systematic refinement based on biological plausibility, clinical insight, and mechanistic understanding. The

refined models are subsequently assessed using appropriate statistical techniques, following the interactive process mining approach [17]. Through this iterative cycle of hypothesis testing, refinement, and assessment, the framework provides a transparent and reproducible approach for generating robust causal evidence while maintaining the methodological rigor of causal epidemiology.

For this reason, this paper introduces a governed, closed-loop diagnostic workflow that integrates causal modelling, data-driven pattern discovery, and statistical model selection to iteratively challenge and refine pre-specified causal assumptions using real-world data, without performing causal discovery or replacing formal identification strategies. Specifically, it seeks to address these challenges systematically by treating RWD analysis as an iterative process rather than a one-time statistical application [18]. In this sense, the Tutorial Demonstration section uses synthetic data calibrated to published results for illustrative purposes only. This manuscript consolidates the iterative methodological development into a unified governed framework. Specifically, we i) formalise the four-phase refinement loop with explicit human adjudication at each transition, ii) define the role of process mining as diagnostic appraisal of causal assumptions rather than causal proof, iii) produced a web demonstrator to illustrate the closed loop over a tutorial dataset, and iv) demonstrate the governed human-AI loop can successfully recover the true clinical causal graph by rejecting observation-induced artifacts.

The rest of the paper is structured as follows. First, the framework is positioned in relation to emerging AI methods in the *Background* section. Then, our four-phase methodology is described in the *Methodology*, and the resulting web demonstrator is presented in the *Result* section, as well as the insights from the simulation study. Last, the *Discussion* describes how our framework addresses key challenges in causal epidemiology, clarifies when it provides the greatest value, and establishes governance infrastructure for integrating agentic AI systems, while preserving interpretability, biological plausibility, and clinical validity.

2 Background

Before the mathematical formalization of modern epidemiology and the release of powerful artificial intelligence models, Dr. John Snow established the foundational archetype for causal inference using real-world data during the 1854 London cholera outbreak [19]. Rather than accepting the dominant "miasma" hypothesis of his era, he systematically collected observational field data, mapping cholera mortality at the household level to uncover spatial and sequential patterns. By demonstrating that exposure to specific water sources strictly preceded disease trajectories and that removing the Broad Street pump handle halted the outbreak, Snow successfully succeed by considering causal assumptions as testable, falsifiable hypotheses.

Despite this long-standing paradigm of empirical falsification, modern workflows for specifying causal relationships have largely remained isolated from the data they intend to analyze. Traditional knowledge-based DAGs [1–5] rely on literature review and expert opinion to manually construct causal graphs. While these approaches

ensure clinical plausibility, they risk misspecification when causal structures are uncertain or complex [6, 8]. Tools such as DAGitty and ggdag support DAG construction and identification of adjustment sets, but do not provide mechanisms for empirical appraisal of the assumed structure.

Automated causal discovery algorithms, including constraint-based methods (PC, FCI) and score-based methods (GES) [10, 20], infer causal structures from conditional independence tests or statistical fit optimisation. However, these methods rely on strong assumptions (faithfulness, causal sufficiency) that are rarely satisfied in complex healthcare data, and their outputs often lack clinical interpretability [3, 10]. In the recent years, some studies have been tested large language models for causal reasoning, demonstrating that those LLMs can generate plausible DAGs from textual descriptions [21, 22]. However, these approaches based on generating DAGs without human supervision raised concerns about hallucination, citation accuracy, and the opacity of the reasoning process.

Process mining offers a complementary perspective by discovering temporal patterns directly from longitudinal event data [16]. Originally developed for business process optimisation, process mining has been increasingly applied to healthcare for pathway analysis [23–26], clinical trajectory characterisation [27–29], and care quality assessment [18]. There are several approaches using what is called causal machine learning, which instead of calculating the step probabilities, calculate the uplift, which is defined as the event contribution to the causal probability [30]. In addition, interactive process mining [17] emphasises continuous human-in-the-loop engagement, and question-driven approaches [31] centre analysis around expert-formulated research questions. However, existing process mining applications in healthcare typically focus on descriptive pathway analysis rather than causal inference, and lack systematic frameworks for translating discoveries into causal model refinements [32, 33].

The closed-loop framework presented in this paper aims to implement John Snow’s philosophy by automatizing knowledge-based DAG specification with empirical temporal pattern discovery through process mining, while embedding human-AI governance to maintain clinical validity. In consequence, it preserves expert authority over causal decisions and provides a systematic mechanism for empirical graph refinement, improving current knowledge-based approaches found in the state of the art. Table 1 summarises the different strategies reviewed in this section.

3 Methodology

Our proposed closed-loop methodology joins knowledge-based theory and empirical evidence through a structured, iterative refinement process. In this sense, the framework is designed to support the documentation of knowledge-based causal models using longitudinal observational patterns and expert review without replacing a formal identification strategy, domain knowledge, or sensitivity analysis. Thus, the main contribution of this work consists on defining the governance structure and methodological infrastructure that enables AI systems to integrate responsibly into causal inference workflows. In consequence, we have defined clear roles for AI assistance

Table 1: Comparison of approaches for causal structure specification in epidemiology.

Approach	Primary input	Method	Clinical integration	AI role	Human role	Best suited for
Knowledge-based DAGs [4]	Literature and expertise	Manual graph construction	High	Minimal	Central	Well-understood mechanisms
Constraint-based algorithms (PC, FCI) [20]	Conditional independence tests	Algorithmic structure learning	Low	Automated discovery	Limited	High-dimensional exploration
Score-based algorithms [20]	Statistical model fit	Search and optimisation	Low	Automated model comparison	Limited	Candidate-graph comparison
Agentic AI [20]	LLMs and knowledge bases	Automated causal reasoning	Moderate	Hypothesis generation	Oversight	Scalable hypothesis generation
Closed-loop framework	Domain knowledge and longitudinal data	Iterative, governed refinement	High	Checkpoint augmentation	Central	Complex and uncertain structures

tasks, while human experience is required in tasks evaluating biological plausibility, clinical interpretation, or causal adjudication.

3.1 The Four-Phase Methodology

Figure 1 summarises the closed-loop methodology. First, in Phase 1 we produce an initial knowledge-based DAG as well as a processed eventlog from literature, expert input, and a client-side RAG knowledge base. Next, Phase 2 applies process discovery to the generated event logs to generate a directly-follows graph (DFG) of observed temporal transitions. Phase 3 compares the DFG with the knowledge-based DAG and presents potential structural refinements, allowing the expert to accept or reject these changes. Last, Phase 4 links the evidence-refined DAG to the appropriate statistical model and imports the results to the knowledge base, closing the loop for the next cycle.

It is important to note that the refined DAG should be interpreted as a causal hypothesis that has been empirically produced when comparing the original DAG with the process map generated in Phase 2. In consequence, process mining remains restricted to measured and recorded events, and the resulting DAG remains subject to classical limitation found in causal inference, including unmeasured confounding variables, measurement errors, and informative censoring. Table 2 summarises each phase, its governance roles, and its main outputs.

3.1.1 Phase 1: Knowledge-Driven Event Log Construction

This phase accomplishes two tasks: (a) constructing an initial knowledge-based DAG representing current understanding of causal relationships, informed by clinical and biological mechanisms, prior epidemiological findings, and expert opinion [1–5]; and (b) using that DAG to define events, populations, and event log structure. The

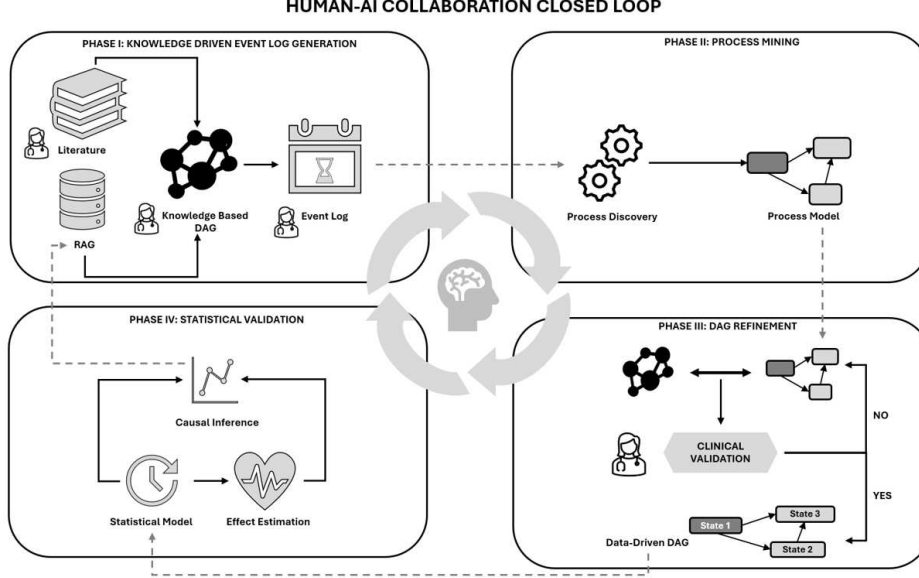


Fig. 1: The four-phase closed-loop data-driven DAG methodology. The feedback loop from Phase 4 to Phase 1 enables iterative updating of the knowledge base. The framework is specifically designed for real-world healthcare data where causal structures are uncertain, temporal patterns are complex, and competing risks are common.

DAG guides event selection (hypothesized causes, mediators, outcomes, and confounders), event definitions (using clinically meaningful thresholds), population design (using a new-user approach to avoid prevalent-user bias), and covariate measurement (distinguishing between confounders and mediators). The output is a structured event log (i.e., a patient-level, longitudinally ordered sequence of timestamped clinical events) containing case identifier, event type, timestamp, and relevant covariates. While AI could potentially assist with literature synthesis [34, 35], clinical expertise remains essential for defining clinically meaningful events and determining appropriate inclusion criteria.

The produced AI-generated initial DAG should be regarded as a syntactic synthesis of relationships described in the source literature, not as a causal map. Medical literature commonly uses associational language ("linked to," "associated with," "more frequent in") for relationships that have not been established as causal. The LLM operates on textual surface forms and cannot reliably distinguish associational from causal claims. Substantial human restructuring — including explicit reclassification of edges as confounding, mediating, collider, or selection structures — is therefore required before the AI-generated DAG can be treated as a causal hypothesis.

Table 2: Four-phase closed-loop methodology and governance roles.

Phase	Current practice	AI enhancement opportunity	Key limitation	Human responsibility
1. Event-log construction	Literature review and expert consultation	Literature synthesis, knowledge-graph organisation, and candidate DAG generation	Hallucination, citation accuracy, and contextual understanding	Define events, clinical meaning, and eligibility criteria
2. Pattern discovery	Process mining using DFG or inductive-miner algorithms	Long-range pattern recognition, clustering, and trajectory stratification	Sample-size requirements and limited interpretability	Validate patterns and assess clinical plausibility
3. DAG refinement	AI-assisted comparison followed by expert judgement	Candidate edge, node, mediator, and competing-risk suggestions	Limited clinical validation and variable explanation quality	Approve final DAG and adjudicate conflicts
4. Statistical modelling	Expert-guided model selection and assumption checking	Model-class recommendation, sensitivity-analysis templates, and result summaries	Assumption checks and effect interpretation require oversight	Interpret clinical significance and final results

3.1.2 Phase 2: Data-Driven Discovery

The objective is to discover actual transition patterns and compare them to hypothesized structures. Process mining discovery algorithms analyze the event log to identify temporal patterns, including the sequence of events, pathway frequencies, competing risks, and deviations from theoretical expectations [16, 23, 24]. This computational process transforms raw logs into a formal Process Model, visualized as states connected by transitions. Among process mining algorithms, DFG focuses on immediate transitions, while inductive miner algorithms discover hierarchical structures [16]. In this case, we used the DFG as a process discovery algorithm. The DFG is a directed graph in which nodes represent event types (e.g., CKD diagnosis, death) and edges represent directly-follows relationships between consecutive events for the same patient, annotated with frequency and median time. This format maps directly onto the DAG node-edge formalism used in Phase 1, making it well suited for systematic comparison with the knowledge-based DAG. The output is a process model representing actual temporal patterns in the population. Neural networks could potentially enhance pattern recognition through deep learning [36] or neuro-symbolic approaches [37, 38], although interpretable visualizations remain essential for clinical assessment.

3.1.3 Phase 3: DAG Refinement

This phase represents the core of the human-AI loop, where empirical discoveries meet expert knowledge to produce an evidence-refined DAG. Process-mining outputs describe trajectories of recorded clinical events, not necessarily trajectories of latent biological disease onset. Observed transition patterns are jointly determined by underlying biology, monitoring intensity, healthcare-seeking behavior, coding practices, and event-definition thresholds. In active-comparator pharmacoepidemiology specifically, differential surveillance between exposure groups is a leading source of artifactual transition asymmetry. The process model from Phase 2 is compared against the initial Knowledge-Based DAG. In this comparison, the DFG serves as an empirical baseline of observed temporal transitions, functioning as a diagnostic tool to evaluate the causal hypotheses represented by the DAG. By encoding temporal succession instead of causal direction, the DFG provides descriptive temporal evidence that can support, challenge, or prioritize hypothesized causal relationships, subject to expert adjudication and downstream formal causal analysis. Discrepancies between the data and the initial knowledge trigger a binary decision path. If the plausibility check is "NO" (the model is biologically implausible or lacks meaning), the process loops back for re-evaluation. This refinement includes adding edges (if transitions not in the original DAG are revealed), removing edges (if hypothesized relationships rarely occur), adding nodes (if important intermediate events were omitted), restructuring (if competing risks require a different architecture), or refining timing (if temporal dynamics differ from assumptions). Researchers assess whether observed patterns are biologically plausible, clinically meaningful, and distinguishable from sample-specific artifacts or confounding pathways. Once the assessment is "YES", the changes are accepted. The successful integration of these insights results in an Evidence-Refined DAG, which accurately reflects both biological mechanisms and empirical reality. Throughout, the principle holds: AI proposes, humans decide.

3.1.4 Phase 4: Statistical Model Selection

The objective is to specify statistical models aligned with the refined DAG structure and estimate causal effects. The DAG guides which variables are confounders, which are mediators and colliders, whether interaction terms are needed, and what competing risk structures exist. Regression models (Cox, logistic, linear) address straightforward structures [39]; cause-specific or subdistribution hazard models handle competing risks [40]; and multi-state models capture complex trajectories [41], as shown in Table 3. Systematic sensitivity analyses include E-value calculation [42], adverse control outcomes and negative control outcomes [43], and quantitative bias analysis [44]. While model selection could be automated, human oversight of assumptions, clinical interpretation of effect sizes, and assessment of biological meaningfulness remain essential. For this reason, we propose in our framework a model recommendation by LLMs according to the output of Phase 3 (Evidence-Refined DAG), which should be reviewed and approved by human experts. Importantly, the refined DAG also supports formal causal identification: by inspecting the DAG structure, the researcher can verify whether the back-door criterion is satisfied for the target causal

estimand and identify the minimal sufficient adjustment set, ensuring that the statistical model targets a well-defined causal effect rather than merely an associational quantity.

To make this mapping transparent and reproducible, the framework applies explicit decision rules that link structural features of the evidence-refined DAG to appropriate statistical model classes (Table 3). These rules are not heuristic recommendations but formalised mappings grounded in established epidemiological methodology: the presence of a competing-risk node triggers subdistribution or cause-specific hazard models; a mediator node on the causal pathway triggers mediation analysis; multiple intermediate states with defined transitions trigger multi-state models; and a simple confounding structure defaults to standard regression with the DAG-identified adjustment set.

Table 3: Decision rules linking evidence-refined DAG features to model classes.

DAG feature	Recommended model class	Target estimand or purpose	Required human check
Competing-risk node, such as death	Fine-Gray subdistribution model or cause-specific hazard model	Cumulative incidence or etiologic transition intensity	Confirm target estimand and censoring assumptions
Mediator on the exposure-outcome pathway	Causal mediation analysis	Direct, indirect, and total effects	Verify no unmeasured exposure-mediator or mediator-outcome confounding
Multiple ordered disease states	Multi-state model	State-transition probabilities and transition-specific hazards	Confirm clinically meaningful states and transition ordering
Measured confounder set identified by the DAG	Regression, inverse probability weighting, or propensity-score balancing	Adjusted exposure-outcome association targeting the causal estimand	Validate adjustment set and avoid collider adjustment

Systematic sensitivity analyses are critical for establishing the robustness of the statistical models derived from the evidence-refined DAG. The framework formally recommends the implementation of negative control outcomes and negative control exposures as primary falsification endpoints to detect the presence of unmeasured confounding or structural biases. By applying the selected analytical model to exposure-outcome pairs where no true causal biological relationship exists, researchers can gauge residual systemic bias.

3.1.5 Closing the loop: Iteration and Knowledge Transfer

The transition from Phase 4 back to Phase 1 is an explicitly human driven step, which aims to close the loop and include the results generated during the analysis in the knowledge base. After completing statistical model assessment, researchers can select concrete outputs from the analysis, including the reviewed DAG structure, the structured decision track documenting accepted and rejected refinements, and

the statistical model specifications and key analytical results. In subsequent cycles, the initial DAG construction assisted by LLMs is informed not only by external literature and domain expertise but also by prior project-specific outputs retrieved through the RAG system, allowing to concatenate refinements of the causal assumptions across iterations, while preserving transparency and traceability of how prior evidence helps to define new hypotheses. Moreover, we have designed this feedback mechanism as a hypothesis generating process instead of a self-reinforcing learning system. In consequence, outputs from previous iterations are considered contextual inputs subject to scientific falsifiability, rather than as validated causal knowledge. However, to mitigate risks of circular reasoning or overfitting, we recommend to perform subsequent analyses using independent datasets, temporal validation strategies, or pre-specified analytical plans. In summary, by closing the loop by a human supervised step, we transform the framework from a single-pass analytical pipeline into a governed research process, while maintaining human oversight, auditability, and methodological rigor at each stage.

3.2 Simulation Study

As the true causal structure of the SCREAM cohort is unknown, we have also evaluated the framework in a controlled continuous-time multi-state simulator with a known data-generating process. The simulator separated the latent disease process from the observation process and injected two artefacts that the framework should reject: a drug-feedback decoy and a post-event coding shift. Using this approach, we could evaluate if the framework can distinguish and reject artifacts which can affect the inference process. The experimental set up included three configurations we compared over $M = 500$ replications: i) a DAG based only in previous knowledge used as baseline, ii) a totally empirically process-mining refinement arm, and iii) a governed refinement arm with plausibility gates based on activity type, temporal directionality, and monitoring sensitivity. The simulations assessed the structural recovery of the true graph, the rejection of observation-induced artefacts, the recovery of the Death competing-risk node, and the recovery of the exposure log-hazard ratio for CKD onset. From a methodological point of view, the inclusion of the simulation experiments complements the tutorial by testing whether expert-governed refinement resists artefactual directly-follows edges when ground truth is known.

3.2.1 Latent data-generating process

We have defined four different states along the patient trajectory, namely drug initiation (S0), chronic kidney disease (S1), cardiovascular adverse event (S2), and death (S3). Between these states, we defined six different transitions which can happen, in concrete $S0 \rightarrow S1$, $S0 \rightarrow S2$, $S0 \rightarrow S3$, $S1 \rightarrow S2$, $S1 \rightarrow S3$, $S2 \rightarrow S3$. For each transition $j \rightarrow k$, we have modelled the cause-specific intensity with a Weibull distribution, according to the following equation:

$$h_{jk}(t \mid E, L) = \lambda_{jk} \rho_{jk} t^{\rho_{jk}-1} \exp(\beta_{jk}^E E + \beta_{jk}^L L)$$

where E is the exposure to the drug ($\text{PPI} = 1, \text{H2B} = 0$) and L is a baseline covariate vector (age, sex, diabetes, hypertension, baseline eGFR). In this model, exposure is computed using a logistic propensity model on L so that the arms are confounded, generating an imbalance that the proposed framework must identify. Next, patient trajectories are sampled by the standard competing-risks algorithm until death or administrative censoring happens in the 10-year horizon. In consequence, the exposure effects follow the mediation hypothesis, where we assumed β^E is positive on $\text{S0} \rightarrow \text{S1}$ (the PPI effect on CKD onset) and on the direct $\text{S0} \rightarrow \text{S2}$ path, and is modelled as zero on $\text{S1} \rightarrow \text{S2}$, as we considered mediation operates through CKD *occurrence*, not by modifying the $\text{CKD} \rightarrow \text{CVAE}$ rate. The scale parameters λ_{jk} were calibrated in a way the recorded event log approximately reproduces the Table 4 marginal transition risks. Last, the true mediated proportion of the $\text{PPI} \rightarrow \text{CVAE}$ effect that flows through CKD is estimated by high-precision g-computation on the simulator itself ($n = 50000$) counterfactual CVAE cumulative incidence under exposure overridden per transition rather than from any external estimator.

3.2.2 Observation process and injected artefacts

Each patient generates clinic visits as a Poisson process with rate $\lambda_{\text{base}} \cdot (\kappa \text{ if PPI else } 1)$, using a monitoring differential $\kappa \geq 1$ for making PPI users more intensively surveilled than H2B users. Moreover, we considered CKD and CVAE are recorded only at the first visit at or after their latent onset, with a detection delay shorter for PPI because of the higher visit rate, reproducing informative-presence bias [12]) Regarding follow-up, it is truncated at death time, so no disease event is ever recorded after death happens. We have injected two artefacts in the simulations at the next visit after a recorded CVAE, defining a probability p_{switch} for a medication change activity, producing a decoy appearing in the log as a $\text{CVAE} \rightarrow \text{drug}$ transition, DAG shorthand $\text{S2} \rightarrow \text{S0}$, and a probability p_{recode} for a false CKD re-code, producing a reverse $\text{CVAE} \rightarrow \text{CKD}$ re-detection, reflected in a DAG shorthand at $\text{S2} \rightarrow \text{S1}$. It is important to note neither artefact corresponds to a latent causal intensity, and both of them appear in the directly-follows graph and must be rejected by the loop.

3.2.3 Discovery, refinement arms, and the plausibility gate

The recorded log is summarised by a directly-follows graph (discovered with pm4py); activities are mapped to the four state nodes and each candidate state-transition is annotated with its relative-antecedent frequency, median time, and a monitoring-stratified strength (its frequency in high- versus low-monitoring strata); Fig. S1 shows the discovered process map, with the two injected artefacts highlighted. Refinement starts from the initial knowledge-based graph ($\text{S0} \rightarrow \text{S1}$, $\text{S0} \rightarrow \text{S2}$, $\text{S1} \rightarrow \text{S2}$; no Death node) and is run in three configurations:

- **Arm A:** a mechanical proposer adds every directly-follows transition whose support exceeds 10% (the Fig. 4 indicator) that is absent from the current graph; every proposal is accepted. This is a deterministic proxy for an ungoverned AI/refinement system.

- **Arm B:** the same proposer, but each candidate passes a deterministic plausibility gate. In this base-cell implementation the gate applies two trap-specific rules: (i) a directionality/acyclicity rule — a candidate that reverses the known disease-progression order is rejected (this catches the reverse coding shift $S2 \rightarrow S1$); and (ii) an activity-type rule — a transition generated by an administrative activity such as a medication change is not a biological causal edge (this catches the decoy $S2 \rightarrow S0$). The monitoring-stratified edge strength is computed and stored as falsification evidence exposed only to the gate, but is *not* used as a hard rejection rule here, because the surveillance-inflated $S0 \rightarrow S1$ edge is a true edge that such a rule would wrongly discard; a negative-control check is specified for the extended gate used in the planned sweeps.
- **Arm C:** the loop is handed the true graph; an upper bound.

3.2.4 Outcome metrics

Structural recovery is graded against the true edge set: edge precision, recall, F1, structural Hamming distance (SHD), recovery of the Death node, and the per-artefact rejection rate. The primary endpoint is the proportion of replications in which a configuration accepts at least one injected artefact edge; the primary contrast is Arm A versus Arm B. Estimand recovery is assessed by fitting cause-specific Cox models for the exposure log-HR and comparing the estimate to the simulator truth (bias, RMSE, 95% CI coverage); a small ridge penalty (0.01) is applied for numerical stability across replications, and the point estimate is unchanged without it. All proportions carry Wilson 95% intervals.

4 Results

To ensure rigorous application, the closed-loop framework operates under several key constraints. First, it distinguishes temporal succession from causal mechanism, recognizing that a directly-follows graph captures event order rather than inherent directionality. Similarly, observed transition asymmetries are treated as potential artifacts of healthcare utilization or administrative coding rather than definitive biological mediation. Because process mining is restricted to measured variables, the framework remains subject to unmeasured confounding and serves as a complement to—not a substitute for formal identification strategies and sensitivity analyses. Crucially, all structural modifications to the DAG require expert adjudication to ensure alignment with biological plausibility. This maintains a strict boundary between exploratory refinement and confirmatory estimation: DAG updates derived from process mining are strictly hypothesis-generating. By documenting every refinement in a built-in decision log—including its evidential basis and human rationale—the framework provides a robust audit trail that separates exploratory discovery from the rigorous requirements of external confirmatory analysis.

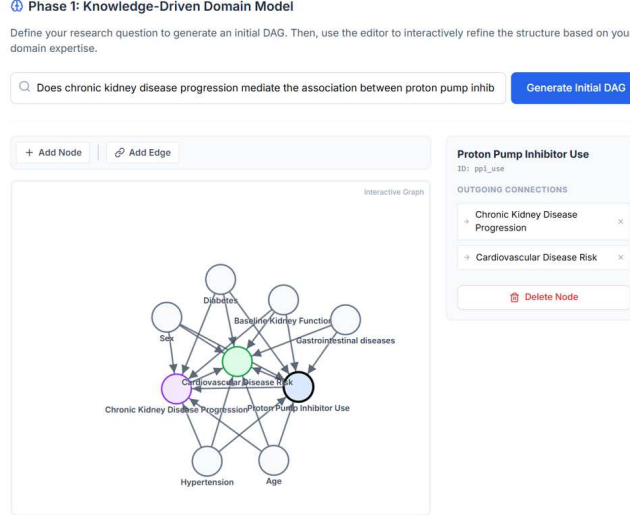


Fig. 2: Initial literature-informed causal DAG generated by the AI tool for the PPI-CKD-CVAE case study.

4.1 Tutorial Dataset

The tutorial demonstration used synthetic longitudinal data calibrated to published SCREAM cohort results comparing proton pump inhibitor (PPI) and histamine-2 receptor blocker (H2B) users [45]. In contrast, This demonstration is not intended to estimate the causal effect of PPI use on cardiovascular outcomes. It illustrates how the closed-loop framework would process longitudinal event-log data, generate process-mining summaries, propose DAG refinements, and recommend statistical models. Synthetic data were chosen to enable a controlled, transparent, and fully reproducible demonstration of the framework workflow, while avoiding privacy constraints associated with individual-level patient data. The demonstrator can be accessed in the following link: <https://closed-loop-dag.fly.dev/>

4.1.1 Case-Study Setup

The workflow was implemented in HealthProcessAI, an open-source interactive tool that supports DAG construction, event-log inspection, DFG generation, expert adjudication of proposed refinements, and model recommendations. The demonstration used GPT-4o-mini via the OpenAI API for LLM-assisted synthesis and suggestion generation; all AI outputs were treated as proposals requiring expert review.

Literature-informed grounding identified PPI exposure, H2B comparison therapy, CKD progression, cardiovascular adverse events (CVAE), death, and baseline clinical covariates as central elements of the initial causal model (Figure 2). The target estimand was framed as the effect of initiating PPI rather than H2B therapy on subsequent CVAE risk, with CKD progression considered a potential mediator and death a competing event.

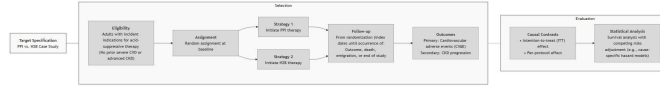


Fig. 3: Target trial emulation structure for the PPI versus H2B comparative-effectiveness question.

To ensure the causal question is well-defined and distinct from the statistical methodology used to estimate it, the analysis of proton pump inhibitors (PPIs) versus histamine-2 receptor blockers (H2Bs) is grounded in the target trial emulation framework. Figure 3 outlines the hypothetical randomized trial protocol that the observational analysis seeks to emulate.

4.1.2 Process Mining-Derived Disease Trajectories

The event log was reconstructed from relational clinical tables into chronologically ordered patient traces containing drug exposure, CKD progression, cardiovascular events, death, timestamps, and treatment group. A directly-follows graph (DFG) was then used to summarise observed temporal transitions, transition frequencies, and median transition times, as shown in Figure 4. These empirical patterns were compared with the initial literature-informed DAG to identify discrepancies requiring expert review.

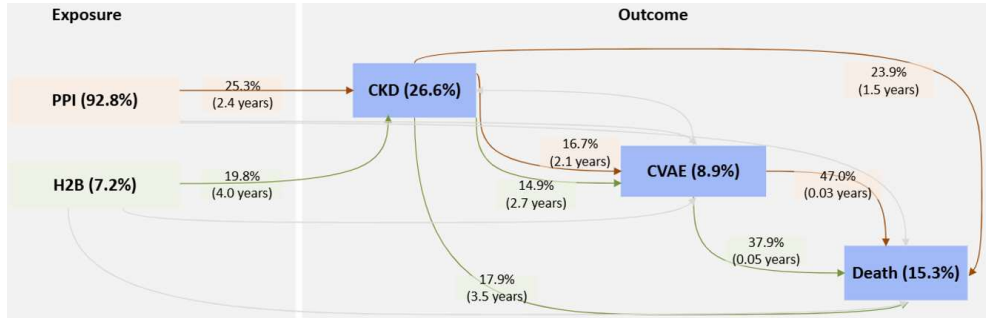


Fig. 4: Process model (directly-follows graph) of observed clinical trajectories for PPI and H2B users in the SCREAM cohort.

The governance panel translated DFG–DAG discrepancies into candidate refinements. In the tutorial (Figure 5), the tool proposed adding Death as a competing-risk node because mortality transitions from CKD and CVAE were frequent and analytically consequential. It also proposed a CVAE-to-drug feedback edge, which was rejected because post-event medication changes reflect clinical management rather than a causal pathway from CVAE to prior exposure.

Phase 3: Human-AI Governance Panel

Review data-driven proposals based on the analysis of 294734 patient trajectories.

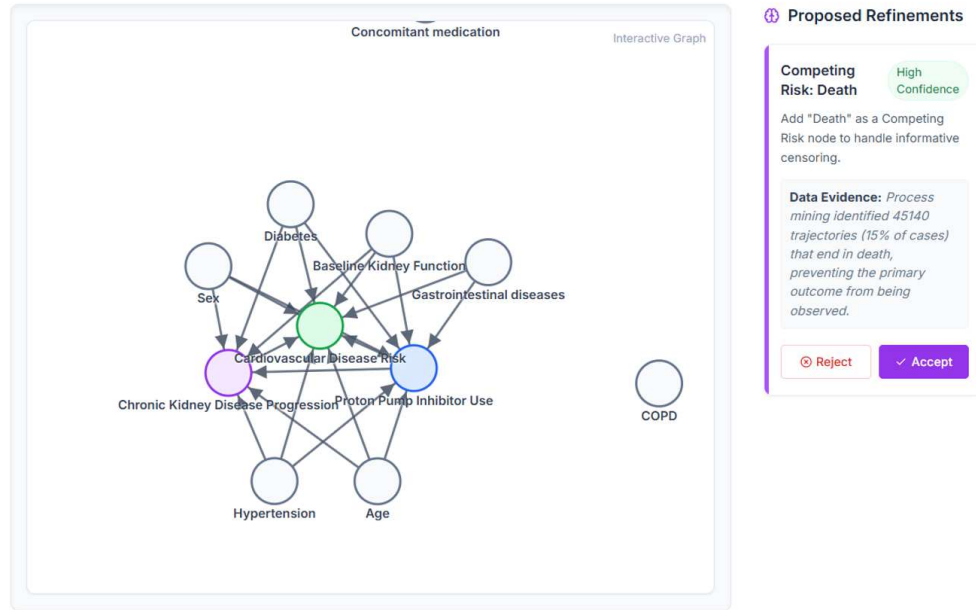


Fig. 5: Phase 3 output: the Human-AI Governance Panel for structured DAG refinement. The left panel displays the current DAG as an interactive force-directed graph, colour-coded by causal role (blue = exposure, green = outcome, red = competing risk, purple = mediator, grey = confounder). The right panel presents AI-generated candidate refinements, each showing the proposed structural change.

4.1.3 Data-Driven Refinement of the Causal Model

The evidence-refined DAG differed from the initial DAG in two main ways. First, Death was added as a competing event because high mortality transitions were observed after both CKD and CVAE. Second, CKD progression was retained as a candidate mediator on the PPI-CVAE pathway because CKD-to-CVAE transitions were frequent, while reverse CVAE-to-CKD transitions did not differ significantly between exposure groups. This pattern is diagnostic rather than confirmatory: transition asymmetry cannot prove causal direction, and formal mediation analysis is required to quantify any mediated effect. The transition rates and ratios are presented in Table 4

4.1.4 AI-Guided Analytical Strategy and Model Configuration

Using the evidence-refined DAG, the Phase 4 module recommended Fine-Gray models for cumulative-incidence estimands involving Death as a competing event, cause-specific hazard models for etiologic transition-intensity estimands, causal mediation analysis for the CKD pathway, and propensity-score balancing for baseline covariate imbalance. The recommendations were treated as structured analytical support rather than autonomous model selection, as presented in Figure 6.

Table 4: Transition rates and crude rate ratios for process-mining trajectories comparing PPI and H2B users.

Transition	PPI				H2B				PPI vs H2B	
	Events	At risk	Risk, %	Median (IQR), y	Events	At risk	Risk, %	Median (IQR), y	RR (95% CI)	<i>p</i>
<i>Key transitions used for DAG refinement</i>										
Medication→CKD	69,177	273,569	25.3	2.4 (0.7, 5.7)	4,181	21,165	19.8	4.0 (1.4, 7.7)	1.28 (1.24, 1.32)	<0.001
CKD→CVAE	12,355	73,823	16.7	2.1 (0.4, 4.7)	665	4,466	14.9	2.7 (0.6, 5.4)	1.12 (1.05, 1.21)	0.002
CKD→Death	17,659	73,823	23.9	1.1 (0.2, 3.5)	798	4,466	17.9	1.4 (0.3, 4.3)	1.34 (1.26, 1.43)	<0.001
CVAE→Death	11,610	24,707	47.0	0.03 (0, 1.2)	538	1,421	37.9	0.05 (0, 1.8)	1.24 (1.16, 1.33)	<0.001
<i>Additional observed transitions</i>										
Medication→CVAE	12,352	273,569	4.5	2.9 (1.0, 5.9)	756	21,165	3.6	3.9 (1.6, 7.3)	1.26 (1.18, 1.36)	<0.001
Medication→Death	13,967	273,569	5.1	1.5 (0.4, 4.1)	568	21,165	2.7	3.5 (1.3, 6.7)	1.90 (1.75, 2.07)	<0.001
CVAE→CKD	4,646	24,707	19.0	0.7 (0.1, 2.4)	285	1,421	20.1	1.4 (0, 1.8)	0.94 (0.84, 1.04)	0.238

Note. Risk is calculated as events divided by the number at risk. RR denotes the crude risk ratio comparing PPI with H2B; 95% confidence intervals were derived using the log method. Median time is reported in years with interquartile range. Statistics were originally published in [45] and are reproduced here to illustrate how process-mining outputs inform DAG refinement.

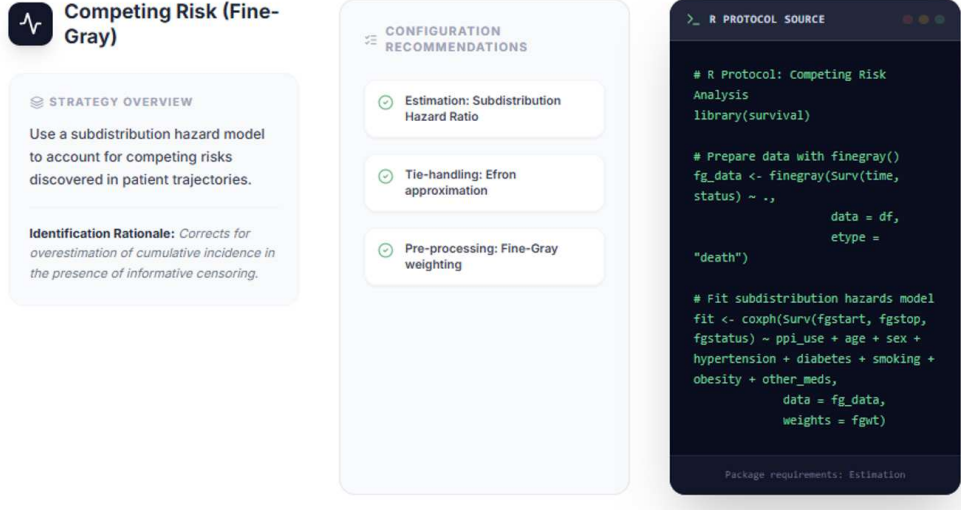


Fig. 6: ProcessAI Phase 4 output: Modelling Strategy Summary with analytical recommendations generated from the refined data-driven DAG.

4.2 Ground Truth Simulation Study

This section presents the results obtained from the simulation study described in Section 3.2, which aims to demonstrate the governed loop (a) recovers correct structure including the competing-risk node, (b) rejects edges that are artifacts of the observation process rather than the disease process, and (c) yields a refined DAG whose implied statistical model recovers known causal quantities. To guarantee reproducibility, the simulation experiment can be reproduced using the code in the following repository: <https://github.com/eduardo-illueca/closed-loop-simulation-companion>

4.2.1 Structural recovery and artefact resistance

Table 5 reports structural recovery averaged over $M = 500$ replications. The governed configuration (Arm B) recovered all high-support true transitions surfaced by process mining and added the Death competing-risk node in every replication, with perfect edge precision but one missing rare background-mortality edge (recall 0.83); the ungoverned configuration (Arm A) accepted both injected artefacts in every replication, degrading precision. Fig. S2 contrasts the refined graphs: Arm A retains both red artefact edges, whereas Arm B rejects them and adds the Death competing-risk node.

The primary contrast was unambiguous: Arm A accepted at least one artefact in 100% of replications and Arm B in 0%. The decoy ($S2 \rightarrow S0$) was rejected by the activity-type rule (a medication change is administrative, not a disease transition — the Fig. 5 rationale) and the coding shift ($S2 \rightarrow S1$) by the directionality rule (it reverses the $CKD \rightarrow CVAE$ order); the ungoverned arm, lacking these rules, accepted both because

Table 5: Structural recovery at the base cell (mean over $M = 500$; $n = 5,000$).

Arm	Precision	Recall	F1	SHD	Traps accepted	Accepts ≥ 1 trap
A — ungated proposal-only	0.71	0.83	0.77	3.00	2.00	100% (0.99–1.00)
B — governed (gate)	1.00	0.83	0.91	1.00	0.00	0% (0.00–0.01)
C — oracle	1.00	1.00	1.00	0.00	0.00	0%

their directly-follows support (≈ 0.34 and ≈ 0.18) exceeded the threshold. Recall was 0.83 for both data-driven arms, not 1.0: the rare direct background-mortality edge $S0 \rightarrow S3$ (recorded risk ≈ 0.05) falls below the 10% discovery threshold and is therefore not surfaced by process mining at this sample size — a sample-size limitation (Table 2/Table 5) — though the Death node is nonetheless recovered through the high-mortality $S1 \rightarrow S3$ and $S2 \rightarrow S3$ transitions.

4.2.2 Estimand recovery

Table 6 reports the exposure log-HR on CKD onset (true value 0.25). Fit to the underlying disease process, the cause-specific Cox model recovered the truth with negligible bias and near-nominal 95% coverage (mean $\hat{\beta} = 0.250$, 95% Monte Carlo interval 0.123–0.387), confirming that the model class the framework selects is correct. Fit to the *recorded* times, the estimate was inflated by surveillance bias and its coverage fell below nominal. Naively adjusting for the realised visit count did not repair this — it over-corrected (collider bias, because visit count is endogenous to follow-up duration); this post-hoc adjustment is a failed comparator, not the recommended model, and its failure is precisely why the framework treats surveillance bias as a **structural** diagnostic (the monitoring-attenuation check at the gate) rather than a covariate adjustment. The CKD $\rightarrow$ CVAE exposure effect, whose true value is zero, was estimated as a small non-significant positive effect on the recorded data, mirroring the manuscript’s observation that the between-group CVAE/CKD signal does not reach significance.

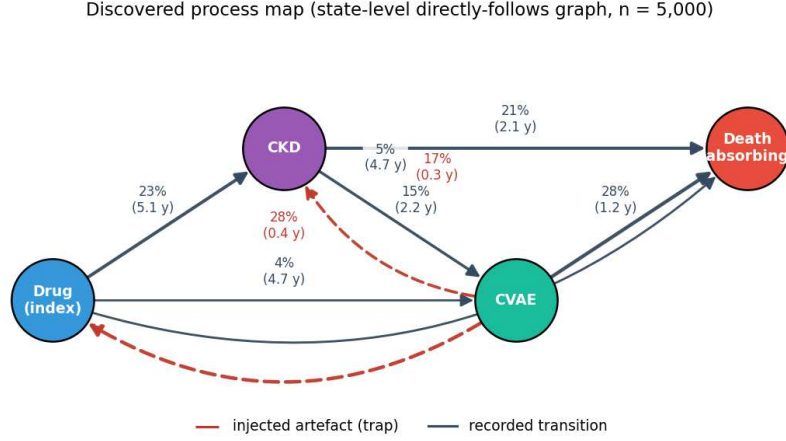
Table 6: Estimand recovery for the exposure log-HR (mean over $M = 500$).

Target	True β^E	Mean $\hat{\beta}^E$	Bias	RMSE	95% CI coverage
$S0 \rightarrow S1$, latent process (estimator validity)	0.25	0.250	-0.000	0.067	0.96
$S0 \rightarrow S1$, recorded (surveillance-inflated)	0.25	0.327	+0.077	0.105	0.83
$S0 \rightarrow S1$, post-hoc visit-count adjusted (not recommended)	0.25	0.636	+0.386	0.395	0.01
$S1 \rightarrow S2$, recorded (true effect null)	0.00	0.180	+0.180	0.273	0.91

4.2.3 Graph Comparison

Figure 7 displays a discovered process map structured as a state-level directly-follows graph that tracks patient transitions across four primary states, for the four defined phases. Solid dark grey lines represent standard recorded transitions labeled with their

relative-antecedent frequency and median time, showing pathways such as patients moving from the initial Drug state to CKD or directly to Death . Conversely, red dashed lines represent injected artifact traps, specifically highlighting a "drug-feedback decoy" and a "reverse coding shift", which represent artificial distortions or errors in the expected linear clinical pathway.



edge label = relative-antecedent frequency (median time); $S2 \rightarrow S0$ = drug-feedback decoy, $S2 \rightarrow S1$ = reverse coding shift

Fig. 7: Discovered process map (directly-follows graph) of observed transitions ($n = 5,000$). The drug-feedback decoy ($S2 \rightarrow S0$) and reverse coding shift ($S2 \rightarrow S1$) are shown as red dashed edges.

According to this stress experiment, we used the obtained process map to produce the different graphs defined in the simulation study, with the injected observation artefacts, showing that the governed loop (Arm B) recovered all high-support true transitions including the competing-risk node and rejected both artefacts in 100% of simulations, whereas the ungoverned configuration (Arm A) accepted at least one artefact in 100% of the simulations. Furthermore, the model selected by the framework recovers the true cause-specific hazard ratio at near-nominal coverage on the underlying process, as shown in Figure 8 .

DAG refinement: starting graph $\{S0 \rightarrow S1, S0 \rightarrow S2, S1 \rightarrow S2\} \rightarrow$ refined graph per arm

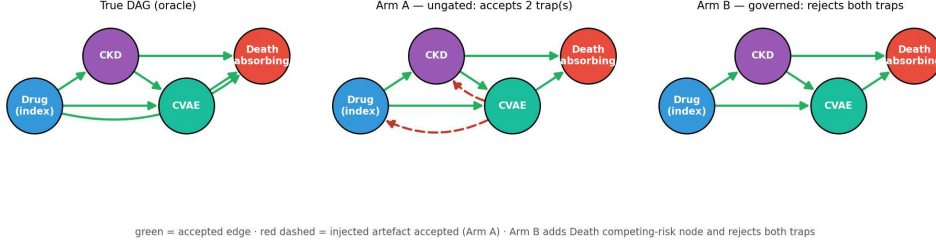


Fig. 8: DAG refinement: the true DAG, the Arm A (ungated) refined graph that accepts both artefacts, and the Arm B (governed) refined graph that rejects both and adds the Death competing-risk node.

5 Discussion

In this study, we presented a closed-loop human–AI collaboration framework which automatizes causal inference as an iterative, falsifiable, and governance-aware scientific process when using real-world data. We hypothesized that by integrating causal modelling and data-driven pattern discovery we can produce a framework for human supervised causal inference. This hypothesis has been demonstrated thanks to the simulation experiments, where we have shown that the governed framework can identify the artificats and bias introduced in the simulation data. Thus, our closed loop improves current methodologies in the state of the art, in concrete interactive process mining, where there is an interpretability gap limiting expert to recognise artifacts or noise in the process models [17]. Futhermore, rather than considering DAGs as static representations of prior knowledge or relying only on automated process discovery, the framework explicitly positions knowledge-based DAGs as testable hypotheses that can be empirically challenged according to the scientific method and refined through longitudinal data. By integrating process mining, expert oversight, and statistical model selection within a single methodological loop, the framework addresses several long-standing limitations in observational health research.

In consequence, the main contribution of our work is methodological, as we have proposed an approach based on process mining, but does not discover causality autonomously, as the event log is constructed with the information retrieved by the LLMs in the literature. This decission allowed us to avoid one of the most common limitations of process discovery applied to epidemiology, where DFGs can produce loops, backward jumps, and infinite loops. It is true there are techniques in the state of the art using trees to mitigate these risks [46], but our approach goes beyond the state of the art because the event log and the states are defined according to the outputs of Phase 1. In addition, the framework inserts an explicit diagnostic step between initial DAG specification and statistical estimation, preserving expert authority over causal interpretation, and linking the accepted refinements to model-selection

rationale. In addition, candidate refinements proposed by the AI requires biological plausibility review, documented decision criteria, an audit trail of accepted and rejected suggestions, and separation between graph refinement and effect estimation.

To illustrate how the closed-loop framework functions in practice, we have selected a case study where we compare the exposure of patients PPI and H2B in chronic kidney disease progression, as the initial literature-informed DAG shown that CKD progression could act as a mediator between PPI use and the development of cardiovascular adverse events. In the experiments performed, we have shown that transitions from CKD to cardiovascular events were consistently observed in the trajectories modelled with process mining, whereas reverse transitions (CVAE $\rightarrow$ CKD) did not differ significantly between exposure groups [45]. This pattern is coherent with prior hypotheses about CKD lying on the PPI-CVAE pathway, as it does not adjudicate causal direction nor exclude reverse biological effects of CVAE on kidney function. Furthermore, both directions of CKD-CVAE influence are biologically plausible (cardiorenal syndrome), and a non-significant between-group difference is not evidence of absence. However, a formal causal mediation analysis, as suggested in Phase 4, would be required to quantify the mediated proportion and test the mediator hypothesis statistically.

Moreover, the framework also pointed out clinically relevant competing pathways, in concrete the high transition rates to death from both CKD and CVAE states. The finding of these patterns motivated the AI recommendation to use competing risk models rather than standard survival approaches, illustrating how empirical pattern discovery can provide diagnostic evidence for analytical decisions without undermining causal rigour. It is also important to note that the AI-generated recommendations aligned with established epidemiological best practices supports the feasibility of embedding AI assistance within a governed workflow. Prior work on disease progression trajectories using event logs from real-world EHR data supports the utility of process mining for longitudinal pattern discovery in clinical cohorts [47]. In this sense, our framework extends question-driven and interactive process mining by placing temporal pattern discovery within a hypothesis-driven causal inference workflow [17, 31]. Unlike automated process discovery algorithms, it does not infer causality from data alone, as it uses empirical discrepancies to trigger structured expert review of pre-specified assumptions [3, 10, 33].

Despite the framework makes falsifiable causal assumptions explicit and empirically challengeable, it does not eliminate residual confounding, measurement error, missing data, or limitations of the underlying clinical epidemiology. In this sense, process mining is also sensitive to biased event definitions, temporal granularity, coding practices, and observation intensity. Furthermore, DFG patterns should be interpreted as diagnostic evidence rather than a proof of causality. For this reason, future work should focus on defining quantitative criteria for DAG refinement, as well as evaluating inter-rater agreement among experts by benchmarking against alternative causal modelling strategies, and testing scalability across clinical domains. In this context, Table 7 summarises key risks that require further methodological development.

Table 7: Future challenges and research gaps.

Challenge	Description
Surveillance Bias	Differential monitoring intensity between exposure groups can create artifactual transition asymmetries, leading to the false discovery of pathways that merely reflect healthcare utilization [48].
Latent Disease Processes	Process models map the trajectories of recorded clinical events, which often lag behind the latent biological onset of disease [49].
Post-Event Coding Shifts	The occurrence of a severe event may systematically alter subsequent coding practices, such as a reduction in routine CKD monitoring or coding, artificially deflating the frequency of reverse transitions [50].
Competing Risks	The DFG itself can be biased by unadjusted competing risks; the absence of a pathway might simply indicate that patients transition to an absorbing state before the target event can be observed [51].
Index-Date Artefacts	Arbitrary or misaligned index dates can distort the apparent temporal sequence of early events [52].
Locality of DFGs	DFGs strictly analyze sequential, event-to-event transitions, which can obscure long-range causal dependencies where an early exposure affects a distant outcome independently of the immediate preceding state [53].
Measurement Error	Misclassification or temporal inaccuracy in nodes can induce spurious edges or mask genuine transitions [54].
Temporal Granularity Limits	When events are recorded with coarse granularity, true temporal precedence cannot be reliably established, risking the conflation of simultaneous events with sequential pathways [55]

6 Conclusions

As conclusion, we have automatized in this work the John Snow’s enduring philosophy for making causal inference by observing real world data, within the modern big data evidence landscape, providing a governed framework for the evidence-informed refinement of knowledge-based causal DAGs. By treating causal assumptions as testable hypotheses rather than static and fixed structures, the proposed closed-loop workflow leverages data driven methodologies as a vital diagnostic tool under our proposed paradigm of structured human supervision. Through this integrated methodology, initial graphical models are built from prior knowledge using LLMs assistance, and then compared against a DFG generated using process discovery from longitudinal clinical event sequences, also modelled using the inputs from the literature. Then, we translated this information into an statistical model recommended by the AI. As a last step, the human supervisor closes the loop by deciding which information from the results is integrated into a knowledge base, which will be used for generating future hypothesis. As we have illustrated in the tutorial validation, this systematic confrontation between conceptual maps and observational trajectories successfully surfaces critical analytical hazards, including hidden mediator pathways and high competing mortality risks. Furthermore, the simulation experiments demonstrates the governed framework can detect artifacts introduced in the synthetic data. In consequence, this iterative cycle serves to complement the epidemiological domain expertise, by producing formal identification strategies and quantitative sensitivity analyses. While future

research must benchmark inter-rater reproducibility across independent expert panels and quantify the precise downstream impacts of framework-guided refinements on final effect estimates, the open-source implementation of the closed loop framework establishes a transparent, governed foundation for bridging the persistent divide between theoretical causal epidemiology and empirical discovery.

Data Availability Statement

The developed application tool is available in the project GitHub (<https://github.com/eduardo-illueca/healthprocessai-v2>), with detailed instructions for deployment. There is also an online demo accessible at <https://closed-loop-dag.fly.dev/>, including the examples shown in the paper. The synthetic dataset and the script for producing the event log are available at <https://github.com/eduardo-illueca/closed-loop-dag/tree/main/dataset>. All the scripts for reproducing all the experiments included in the *Ground Truth Simulation Study* experiment can be found at <https://github.com/eduardo-illueca/closed-loop-simulation-companion>

Ethics Statement

The current paper uses only synthetic data for its tutorial demonstration and does not involve human participants.

Funding Declaration

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

Clinical Trial Number

Clinical trial number: not applicable.

Competing Interests

The authors declare no competing interests.

Author Contributions

EI: Methodology development, formal analysis, validation, writing, original draft, code development, review and editing. KC: Methodology development, formal analysis, data curation, preparation, visualization, writing, review, and editing. CF: Methodology development, writing, review, and editing. FS: Methodology development, supervision, funding acquisition, resources, writing, review and editing. FA: Methodology development, validation, project administration, supervision, writing, original draft, code development, review and editing. All authors have read and approved the final manuscript.

Generative AI statement

The authors used AI-based tools (e.g., large language models) for English proofreading and improving the readability of the manuscript. In addition, Claude Code has been used as an agent for organising the code and the GitHub repositories. These tools were applied only to enhance clarity of expression and did not contribute to the conceptual content, data analysis, or scientific conclusions. The authors take full responsibility for the content of the manuscript.

References

- [1] VanderWeele TJ. Principles of confounder selection. *European Journal of Epidemiology*. 2019;34(3):211–219. <https://doi.org/10.1007/s10654-019-00494-6>.
- [2] Pearl J. An introduction to causal inference. *The International Journal of Biostatistics*. 2010;6(2):7. <https://doi.org/10.2202/1557-4679.1203>.
- [3] Pearl J. *Causality*. Cambridge University Press; 2009.
- [4] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48. <https://doi.org/https://pubmed.ncbi.nlm.nih.gov/9888278/>.
- [5] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615–625. <https://doi.org/10.1097/01.ede.0000135174.63482.43>.
- [6] Diggle JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*. 2022;142:264–267. <https://doi.org/10.1016/j.jclinepi.2021.08.001>.
- [7] Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, et al. Control of confounding and reporting of results in causal inference studies. *Annals of the American Thoracic Society*. 2019;16(1):22–28. <https://doi.org/10.1513/annalsats.201808-564ps>.
- [8] Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*. 2021;50(2):620–632. <https://doi.org/10.1093/ije/dyaa213>.
- [9] Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance*. 2019;32(1):42–49. <https://doi.org/10.1080/09332480.2019.1579578>.
- [10] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*. 2019;10:524. <https://doi.org/10.3389/fgene.2019.00524>.

- [11] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence — What is it and what can it tell us? *New England Journal of Medicine*. 2016;375(23):2293–2297. <https://doi.org/10.1056/nejmsb1609216>.
- [12] Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annual Review of Public Health*. 2016;37(1):61–81. <https://doi.org/10.1146/annurev-publhealth-032315-021353>.
- [13] Burns L, Le Roux N, Kalesnik-Orszulak R, Christian J, Hukkelhoven M, Rockhold F, et al. Real-world evidence for regulatory decision-making: guidance from around the world. *Clinical Therapeutics*. 2022;44(3):420–437. <https://doi.org/10.1016/j.clinthera.2022.01.012>.
- [14] Coleman T. Causality in the time of cholera: John Snow as a prototype for causal inference. Available at SSRN 3262234. 2019;<https://doi.org/10.2139/ssrn.3262234>.
- [15] Ahsen ME, Kittani R, Gerke T, Krishnan L, Rogan S, Scott ER. Leveraging Generative AI for Interpretable Clinical Decision Making Through Causal Graphs. In: *Biocomputing 2026: Proceedings of the Pacific Symposium*; 2025. p. 309–323.
- [16] Van Der Aalst W. Data science in action. In: *Process Mining: Data Science in Action*. Springer; 2016. p. 3–23.
- [17] Fernandez-Llatas C. Interactive process mining in healthcare: an introduction. In: *Interactive Process Mining in Healthcare*. Springer; 2020. p. 1–9.
- [18] Ibanez-Sanchez G, Fernandez-Llatas C, Martinez-Millana A, Celda A, Mandingorra J, Aparici-Tortajada L, et al. Toward value-based healthcare through interactive process mining in emergency rooms: the stroke case. *International Journal of Environmental Research and Public Health*. 2019;16(10):1783. <https://doi.org/10.3390/ijerph16101783>.
- [19] Dayal V, Murugesan A. John Snow and Causal Inference. In: *Demystifying Causal Inference: Public Policy Applications with R*. Springer; 2023. p. 1–17.
- [20] Chickering DM. Optimal structure identification with greedy search. *Journal of Machine Learning Research*. 2002;3:507–554. <https://doi.org/10.1162/153244303321897717>.
- [21] Kiciman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*. 2023;<https://doi.org/10.48550/arXiv.2305.00050>.
- [22] Meier S, Raut PN, Mahr F, Thielen N, Franke J, Risch F. Structured knowledge-based causal discovery: Agentic streams of thought. *Information Processing &*

- Management. 2025;62(5):104202. <https://doi.org/10.1016/j.ipm.2025.104202>.
- [23] Mans RS, Schonenberg MH, Song M, van der Aalst WM, Bakker PJ. Application of process mining in healthcare – a case study in a Dutch hospital. In: International Joint Conference on Biomedical Engineering Systems and Technologies. Springer; 2008. p. 425–438.
 - [24] Rebuge Á, Ferreira DR. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*. 2012;37(2):99–116. <https://doi.org/10.1016/j.is.2011.01.003>.
 - [25] Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*. 2016;61:224–236. <https://doi.org/10.1016/j.jbi.2016.04.007>.
 - [26] Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M, Helm E, et al. Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*. 2022;127:103994. <https://doi.org/10.1016/j.jbi.2016.04.007>.
 - [27] Fernandez-Llatas C, Lizondo A, Monton E, Benedi JM, Traver V. Process mining methodology for health process tracking using real-time indoor location systems. *Sensors*. 2015;15(12):29821–29840. <https://doi.org/10.3390/s151229769>.
 - [28] Kurniati AP, Johnson O, Hogg D, Hall G. Process mining in oncology: A literature review. In: 2016 6th International Conference on Information Communication and Management (ICICM). IEEE; 2016. p. 291–297.
 - [29] Williams R, Rojas E, Peek N, Johnson OA. Process mining in primary care: a literature review. *Studies in Health Technology and Informatics*. 2018;247:376–380. <https://doi.org/10.3233/978-1-61499-852-5-376>.
 - [30] Bozorgi ZD, Teinemaa I, Dumas M, La Rosa M, Polyvyanyy A. Process Mining Meets Causal Machine Learning: Discovering Causal Rules from Event Logs. In: 2020 2nd International Conference on Process Mining (ICPM); 2020. p. 129–136.
 - [31] Rojas E, Sepúlveda M, Munoz-Gama J, Capurro D, Traver V, Fernandez-Llatas C. Question-driven methodology for analyzing emergency room processes using process mining. *Applied Sciences*. 2017;7(3):302. <https://doi.org/10.3390/app7030302>.
 - [32] Chen K, Abtahi F, Carrero JJ, Fernandez-Llatas C, Seoane F. Process mining and data mining applications in the domain of chronic diseases: A systematic review. *Artificial Intelligence in Medicine*. 2023;144:102645. <https://doi.org/10.1016/j.artmed.2023.102645>.

- [33] Montoya F, Hidalgo M, Astudillo H. Causal Process Mining: From Interpretability to Causal Explainability. In: Proceedings of the International Conference on Process Mining; 2025. p. 1–12.
- [34] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*. 2024;18(6):186345. <https://doi.org/10.1007/s11704-024-40231-1>.
- [35] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nature Medicine*. 2023;29(8):1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>.
- [36] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*. 2018;19(6):1236–1246. <https://doi.org/10.1093/bib/bbx044>.
- [37] Garcez Ad, Lamb LC. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*. 2023;56(11):12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>.
- [38] Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh TP. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*. 2024;36(21):12809–12844. <https://doi.org/10.1007/s00521-024-09960->.
- [39] Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*. 2016;133(6):601–609. <https://doi.org/10.1161/CIRCULATIONAHA.115.017719>.
- [40] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*. 2007;26(11):2389–2430. <https://doi.org/10.1002/sim.2712>.
- [41] Haber NA, Wood ME, Wieten S, Breskin A. DAG With Omitted Objects Displayed (DAGWOOD): A framework for revealing causal assumptions in DAGs. *Annals of Epidemiology*. 2022;68:64–71. <https://doi.org/10.1016/j.annepidem.2022.01.001>.
- [42] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*. 2017;167(4):268–274. <https://doi.org/10.7326/M16-2607>.
- [43] Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*. 2013;177(4):292–298. <https://doi.org/10.1093/aje/kws412>.
- [44] Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International Journal of Epidemiology*. 2014;43(6):1969–1985. <https://doi.org/10.1093/ije/dyu149>.

- [45] Chen K, Abtahi F, Fernandez-Llatas C, Xu H, Seoane F. Longitudinal trajectories unravel the complex interplay of medication, cardiovascular events, chronic kidney disease, and mortality. *Scientific Reports*. 2025;15(1):35577. <https://doi.org/10.1038/s41598-025-23527-5>.
- [46] Fernández-Llatas C, Martínez-Salvador B, Marcos M. A declarative approach for interactive process discovery in the clinical domain. *Journal of Biomedical Informatics*. 2025;168:104862. <https://doi.org/10.1016/j.jbi.2025.104862>.
- [47] Chen K, Abtahi F, Carrero JJ, Fernandez-Llatas C, Xu H, Seoane F. Validation of an interactive process mining-driven framework for chronic disease progression. *Scientific Reports*. 2024;7(1):273. <https://doi.org/10.1038/s41598-024-79704-5>.
- [48] Uhl L, Augusto V, Dalmas B, Alexandre Y, Bercelli P, Jardinaud F, et al. Evaluating the Bias in Hospital Data: Automatic Preprocessing of Patient Pathways Algorithm Development and Validation Study. *JMIR Medical Informatics*. 2024;12:e58978. <https://doi.org/10.2196/58978>.
- [49] Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *Journal of the American Medical Informatics Association*. 2022;29(7):1191–1199. <https://doi.org/10.1093/jamia/ocac050>.
- [50] Goldstein BA, Phelan M, Pagidipati NJ, Peskoe SB. How and when informative visit processes can bias inference when using electronic health records data for clinical research. *Journal of the American Medical Informatics Association*. 2019;26(12):1609–1617. <https://doi.org/10.1093/jamia/ocz148>.
- [51] Möllenhoff K, Binder N, Dette H. Testing similarity of parametric competing risks models for identifying potentially similar pathways in healthcare. *Statistics in Medicine*. 2024;43(28):5316–5330. <https://doi.org/10.1002/sim.10243>.
- [52] Swerdel JN, Conover MM. Comparing broad and narrow phenotype algorithms: differences in performance characteristics and immortal time incurred. *Journal of Pharmacy & Pharmaceutical Sciences*. 2024;26:12095. <https://doi.org/10.3389/jpps.2023.12095>.
- [53] Pan J, Mahdavi A, Mino-Rodriguez I, Martínez-Muñoz I, Berger C, Schweiker M. The untapped potential of causal inference in cross-modal research. *Building and Environment*. 2024;248:111074. <https://doi.org/10.1016/j.buildenv.2023.111074>.
- [54] Goodwin AJ, Eytan D, Dixon W, Goodfellow SD, Doherty Z, Greer RW, et al. Timing errors and temporal uncertainty in clinical databases—A narrative review. *Frontiers in Digital Health*. 2022;4:932599. <https://doi.org/10.3389/fdgth.2022.932599>.

- [55] Schuster D, Benevento E, Aloini D, van der Aalst WM. Analyzing healthcare processes with incremental process discovery: practical insights from a real-world application. *Journal of healthcare informatics research*. 2024;8(3):523–554. <https://doi.org/10.1007/s41666-024-00165-6>.