DoAtlas-1: A Causal Compilation Paradigm for Clinical AI

Yulong Li[1,2] **Jianxu Chen**[1,2] **Xiwei Liu**[1] **Chuanyue Suo**[3]
Rong Xia[1,2] **Yichen Li**[1] **Xinlin Zhuang**[1] **Niranjana Arun Menon**[1,2]
Jionglong Su[3] **Eran Segal**[1] **Imran Razzak**[1,2†]

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE [2]MedOmni
[3]Xi'an Jiaotong-Liverpool University, Suzhou, China
[†]Corresponding author: imran.razzak@mbzuai.ac.ae

Abstract

Medical foundation models generate narrative explanations but cannot quantify intervention effects, detect evidence conflicts, or validate literature claims, limiting clinical auditability. We propose **causal compilation**, a paradigm that transforms medical evidence from narrative text into executable code. The paradigm standardizes heterogeneous research evidence into structured estimand objects, each explicitly specifying intervention contrast, effect scale, time horizon, and target population, supporting six executable causal queries: do-calculus, counterfactual reasoning, temporal trajectories, heterogeneous effects, mechanistic decomposition, and joint interventions. We instantiate this paradigm in DoAtlas-1, compiling 1,445 effect kernels from 754 studies through effect standardization, conflict-aware graph construction, and real-world validation (Human Phenotype Project, 10,000 participants). The system achieves 98.5% canonicalization accuracy and 80.3% query executability. This paradigm shifts medical AI from text generation to executable, auditable, and verifiable causal reasoning.

1 Introduction

Medical foundation models are emerging as the universal interface layer of clinical decision support systems (CDSS), delivering textual answers for clinical questions with evidence grounding by combining generative reasoning with retrieval augmentation [37]. However, as shown in Table 1, what ultimately matters for clinical deployment is not generating plausible-sounding knowledge narratives, but supporting accountable intervention decisions: quantifying interventional causal effects on outcomes for a specified target population under explicit constraints [37]. Existing paradigms constructed around generative reasoning and citation alignment typically lack a principled mechanism for compiling heterogeneous research findings into executable interventional estimands [17, 18]. In particular, contrast definition, effect scale/unit, time window, and target population are often left implicit; uncertainty and evidential conflicts are difficult to represent and adjudicate in a structured manner, which limits auditability in high-stakes clinical settings. To address this, we propose DoAtlas, a new paradigm for clinical causal inference that compiles heterogeneous studies into structured interventional estimands, explicitly models effect contrasts, scale, time window, and target population, and adjudicates evidential conflicts using real-world data validation signals.

Despite the importance of clinically actionable and auditable decision support, current medical foundation model paradigms still exhibit three fundamental gaps for clinical causal inference.

1) Missing interventional estimand. Narrative explanations with citation alignment do not yield recomputable estimand-level outputs. Without explicit contrasts and effect scales, explanations remain readable but not actionable, as they are neither recomputable, auditable, nor verifiable [18, 26].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Paradigm shift from narrative generation to causal compilation for clinical decision support.

	Narrative Generation	**Causal Compilation (DoAtlas)**
Representation	Text with citations	Executable estimand objects
Query Interface	Natural language questions	Typed queries (Q_{do}, Q_{med}, Q_{CATE}, ...)
Output Format	"Studies suggest..." + citations	$\hat{\theta} = 0.73$ [0.60–0.88] ✓validated
Effect Quantification	✗ Qualitative descriptions only	✓ Quantitative estimates with CI
Conflict Detection	✗ Implicit ("results vary")	✓ Explicit (directional, interval, I^2)
External Validation	✗ Cannot verify claims	✓ Statistical test vs. real-world data
Causal Reasoning	✗ Narrative descriptions only	✓ 6 executable query types (Figure 1)
Composability	✗ Each query independent	✓ Graph-based query composition
Reproducibility	Stochastic (varies across runs)	Deterministic with full provenance
Evidence Update	Requires model retraining (weeks)	Incremental compilation (minutes)
Example Systems	Med-PaLM, GPT-4-Med, RAG systems	DoAtlas (ours)

2) Absence of effect estimation. Existing medical foundation model paradigms typically do not include causal effect estimation pipelines, lacking estimand specification, scale transformation, and uncertainty propagation [37, 13]. Consequently, lacking comparable and composable effect objects, they cannot reliably answer *"How large is the effect?"*, *"Which is stronger?"*, or *"Does consensus exist?"*

3) Absence of real-world external validation. Existing medical foundation models lack mechanisms to assess the external validity of research claims, preventing the generation of empirically grounded, updatable reliability signals needed to assess credibility, adjudicate conflicts, and guard against unsafe recommendations. Clinical deployment demands not better narrative generation, but foundational infrastructure comprising interventional estimand objects, conflict governance, and real-world validation that renders evidence recomputable, auditable, and verifiable [13].

To address this, we introduce a causal compilation paradigm that systematically compiles heterogeneous research evidence into executable causal atlas objects. ***i)*** **Effect standardization.** Unifying effects from heterogeneous models (linear, logistic, Cox, MR, mediation) within an SCM framework, compiling them into unified interventional estimand objects; each object explicitly specifies intervention contrast, scale/unit, time window, and target population, and retains evidence provenance and key assumptions to support auditability. ***ii)*** **Conflict-aware graph construction.** Detecting evidential discrepancies (directional conflicts, interval incompatibility, heterogeneity), preserving conflict states as structured metadata, and using validation signals to calibrate evidence selection and conflict adjudication when available. ***iii)*** **External validation.** Aligning research claims to Human Phenotype Project (HPP) [1] on mappable subsets, generating reliability signals (directional, effect-size, and subgroup/effect-modification consistency) to inform evidence selection and knowledge evolution. Unlike medical foundation models centered on narrative generation and citation alignment, causal compilation optimizes for executability: claims specify computable bounds and execute only when information is complete. DoAtlas compiles heterogeneous evidence into queryable, auditable causal atlas objects with real-world validation signals.

As illustrated in Figure 1, DoAtlas supports six typed, executable causal inference capabilities:

- **Do-calculus queries:** Answering interventional queries such as "What is the causal effect on Y under $do(X)$?", and returning quantitative effect estimates with confidence intervals and explicit intervention contrasts *(e.g., initiate vs. not initiate; drug A vs. drug B)*, rather than correlational predictions or narrative rationales.
- **Mechanistic causal pathway decomposition:** When mediator information is available and identifiability conditions hold, decomposing total effects into constituent causal pathways and quantifying their relative contributions. *For example, the effect of SGLT2 inhibitors on heart failure hospitalization may be attributed ∼45% to diuretic pathways, ∼30% to myocardial metabolic reprogramming, and ∼25% to renoprotective pathways, enabling mechanism-level interpretation.*
- **Counterfactual reasoning:** Quantifying counterfactual outcomes under alternative interventions to support clinical "what-if" analysis. *For example, "If a patient had quit smoking five years earlier, their current cardiovascular risk would be reduced to X%," providing causal grounding for retrospective analysis and decision-making.*

[1] `https://knowledgebase.pheno.ai/`

Type	Query (NL + sketch)	Output (sketch)
QDO	**NL**: At τ=P2W, what is the effect of Sleep Extension vs Control on Change in Energy Intake? **Query**: Qdo(Π=adult_general, ι=SleepExt vs Ctrl, o=EnergyIntakeChange, τ=P2W, μ=mean_diff)	**Answer**: θ = -270 (mean_diff) **CI**: [-459, -81] **Witness**: EV-2022-SLEEP-EI, EV-2022-TASALI (2)
QCF	**NL**: What is the counterfactual effect of switching sleep from Normal to Short sleep on CAD? **Query**: Qcf(Π=adult_general, ι=Short vs Normal, o=CAD, τ=Lifetime, μ=logOR)	**Answer**: θcf = 0.215 (logOR) **CI**: [0.140, 0.293] **Witness**: EV-2022-930000-MR-CAD, EV-2022-930000-CAD (2)
QTRAJ	**NL**: How does the effect of low Delta entropy on incident CHD vary over time (τ=P11Y,P15Y6M)? **Query**: Qtraj(Π=cvd, ι=DeltaEntropy Low vs Mid, o=Incident CHD, τ∈{P11Y,P15Y6M}, μ=logHR)	**Answer**:{(P11Y,0.378),(P15Y6M,0.582)} (logHR) **CI**: P11Y [0.020, 0.723]; P15Y6M [0.157, 1.004] **Witness**: 5 evidences (e.g., EV-2024-SHHS-DELTA-ENTROPY, EV-2024-MROS-DELTA-ENTROPY)
QCATE	**NL**: Is the effect of Finerenone on CV death different in diabetes vs adult_general? **Query**: QCATE(Π∈{diabetes, adult_general}, ι=Finerenone, o=CV death, τ={P3.4Y,P32M}, μ=logHR)	**Answer**: θ(diabetes)=-0.105, θ(adult_general)=-0.073, Δ=-0.033 (logHR) **CI**: diabetes [-0.301, 0.086]; adult_general [-0.248, 0.104] **Witness**: EV-2021-FIGARO, EV-2024-FINEARTS-HF (2)
QMED	**NL**: What are the direct and indirect effects of BMI on CVD mediated by KDM-BA acceleration? **Query**: Qmed(Π=adult_general, ι=BMI, M=KDM-BA Accel, o=CVD, τ=P9Y, μ=logOR)	**Answer**: direct = 0.199, indirect = 0.020 (logOR) **CI**: direct [0.113, 0.285]; indirect [0.010, 0.030] **Witness**: EV-1-OBESITY-BAA-CVD (1)
QJOINT	**NL**: What is the joint effect of dapagliflozin and calorie restriction on diabetes remission? **Query**:Qjoint(do(Dapagliflozin,CalorieRestriction); Π=diabetes, o=Diabetes remission, τ=P12M, μ=logRR)	**Answer**: θ = 0.445 (logRR) **CI**: [0.157, 0.737] **Witness**: EV-2025-000001 (1)

Figure 1: **Query Gallery.** DoAtlas supports six executable query types: QDO (interventional effect), QCF (counterfactual effect under an alternative exposure), QTRAJ (time-indexed effect trajectory), QCATE (heterogeneous effects across populations/strata), QMED (direct/indirect effects via a mediator), and QJOINT (joint intervention effect). Each answer is computed from a *witness* subset of evidence objects (EV-*); when a query is not executable, the system returns diagnostic *flags* instead of a numeric answer.

- **Individualized treatment effect estimation:** Incorporating effect modifiers such as age, sex, comorbidities, and baseline risk to move from population-average to subgroup- or patient-specific expected benefits, quantitatively informing precision treatment decisions.
- **Combination intervention modeling:** Modeling joint intervention effects and identifying causal pathway interactions to characterize whether combinations of therapies are synergistic, additive, or antagonistic, providing quantitative support for combination treatment decisions.
- **Dynamic prognostic trajectory modeling:** Predicting disease trajectories over time under alternative interventions, answering *"what is the effect?"*, *"when do benefits begin?"*, and *"how long do benefits persist?"*, and offering temporal causal insights to guide long-term treatment and follow-up.

DoAtlas can be integrated with foundation models (FMs) to form a layered architecture. The FMs layer provides clinical semantic understanding and natural language interfaces, while the causal inference layer, implemented by DoAtlas, supplies executable causal reasoning and auditable inference traces. Together, these layers enable a paradigm shift in medical AI from narrative generation toward accountable decision support. Overall, we propose a causal compilation paradigm, instantiated as a foundation for clinical causal inference, that achieves systematic validation against real-world data. This paradigm enables auditable and verifiable causal reasoning in medical AI.

2 Related Work

Paradigm evolution of medical AI. Early medical AI progressed from rule-based expert systems to task-specific models, including regression-based risk scores and large-scale EHR-driven prediction [9, 10, 33, 35]. More recently, general-purpose large language models (e.g., GPT-style models, Claude, Baichuan) and domain-adapted variants such as Med-PaLM have been positioned as interface layers for clinical systems [1, 5, 44, 31, 38]. With retrieval augmentation, these models generate literature-grounded explanations that improve interpretability [27]. However, medical foundation models remain confined to narrative assistance rather than auditable estimates of intervention effects [31, 38, 39]. Elements for clinical decision-making—intervention contrasts, effect magnitudes, and uncertainty—are left implicit, limiting suitability for high-stakes clinical use [24, 2].

Literature grounding and causal modeling. To address these limitations, recent work has pursued two complementary directions: literature-grounded generation to reduce hallucination [27], and causal inference frameworks for modeling interventional effects [32, 17, 40]. Despite these advances, the two lines of work operate in isolation: biomedical literature is treated as narrative justification, while causal methods remain confined to individual studies [35, 24]. Consequently, existing systems

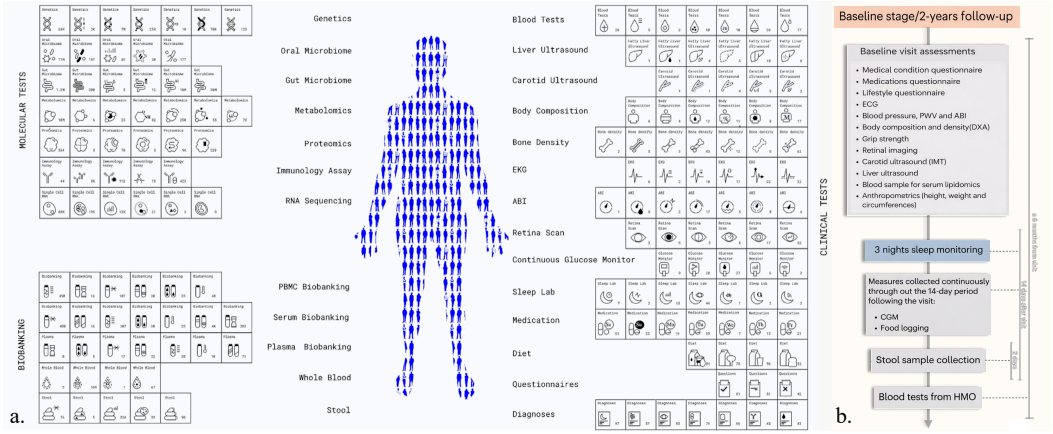

Figure 2: **Illustration of the HPP study data. (a)** Multi-modal and multi-omic data architecture of HPP. Molecular tests include genetics, microbiome metagenomics, metabolomics, proteomics, and single-cell RNA sequencing. Clinical tests capture deep physiological states via retinal imaging, liver and carotid ultrasound, DXA body composition, and continuous monitoring through CGM and multi-night sleep monitoring. Biobanking involves the systematic preservation of blood and stool samples for longitudinal discovery. **(b)** Timing of sleep monitoring with respect to all other phenotypes. The body system characteristics were measured within a period of ±6 months from the visit, with the sleep monitoring performed in three nights within a 2-week time period after that visit. ABI, ankle–brachial index; CGM, continuous glucose monitoring; DXA, dual-energy X-ray absorptiometry; ECG, electrocardiogram; HMO, health maintenance organization; IMT, intimamedia thickness; MS, mass spectrometry; PWV, pulse wave velocity; RDS, recent depressive symptoms.

fail to produce standardized interventional estimates, leaving a critical disconnect between narrative evidence and formal causal inference that limits accountable clinical decision support [39, 2].

3 HPP Dataset

The **Human Phenotype Project (HPP)** is a prospective deep-phenotyping cohort study that is currently enrolling 10,000 participants to support systematic, multi-modal health assessments. As illustrated in Figure 2, the project provides unprecedented phenotypic granularity by integrating data from **17 physiological systems**. Each participant undergoes evaluations spanning cardiovascular, metabolic, immune, neurocognitive, musculoskeletal, digestive, and respiratory systems. These assessments include laboratory biochemistry (e.g., lipid panels, glucose metabolism, renal function, and inflammatory markers), imaging modalities (carotid ultrasound, retinal imaging, and hepatic ultrasound), body composition analysis using dual-energy X-ray absorptiometry (DXA), functional testing, and multi-omics profiling, including genomics, microbiome metagenomics, and metabolomics.

Participants receive comprehensive reassessments on a biennial basis, while key biomarkers are continuously monitored through wearable devices, such as continuous glucose monitoring (CGM) and multi-night home sleep apnea testing. This design enables the dataset to capture dynamic physiological trajectories over time rather than static cross-sectional snapshots. In particular, the longitudinal tracking of biological age (BA) scores across multiple systems allows DoAtlas to reconcile conflicting evidence by projecting heterogeneous research findings onto a consistent real-world temporal axis. This capability further supports the identification of sex-specific dynamics that may be overlooked in less granular datasets, including sigmoid-shaped aging patterns in blood lipid profiles that are associated with menopausal status. Additional details are provided in Appendix D.6.

4 Causal Compilation Paradigm

4.1 Core Abstractions

Medical evidence is reported in highly heterogeneous forms. Across studies, researchers employ different statistical models, report effects using different measures, and target distinct populations and time horizons. This heterogeneity poses a fundamental challenge for systematic comparison, synthesis, and downstream causal reasoning. To address this challenge, we introduce an intermediate

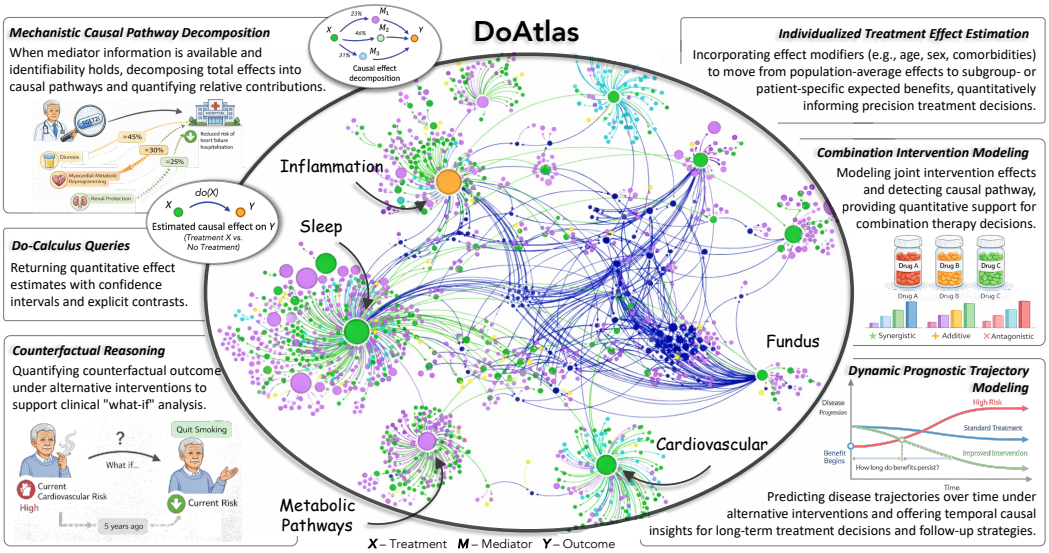

Figure 3: **Overview of DoAtlas.** DoAtlas compiles heterogeneous clinical evidence into standardized interventional estimand objects with explicit contrasts, effect scales, time horizons, and target populations. It organizes comparable claims into conflict-aware causal graphs, supports executable causal queries, including do-calculus estimation, mechanistic pathway decomposition, counterfactual reasoning, individualized treatment effect estimation, combination intervention modeling, and dynamic prognostic trajectory modeling. External validation signals from HPP are used to assess reliability, adjudicate evidential conflicts, and support auditable clinical decision support.

representation (IR) that explicitly separates the semantic content of an interventional estimand from its numeric reporting form, while supporting deterministic normalization across studies.

Definition 1 (Estimand Intermediate Representation)**.** *An interventional estimand is represented by a five-tuple* $\varepsilon = (\Pi, \iota, o, \tau, \mu)$*, where* Π *specifies the target population,* ι *the intervention contrast,* o *the outcome variable,* τ *the time horizon, and* μ *the measure functional. A complete causal claim further includes numeric information, represented by a claim object* $c = (\theta, \mathrm{CI})$*, consisting of a point estimate* θ *and its confidence interval* CI*. Full object specification with provenance metadata is provided in Appendix B.1.1.*

When studies target the same semantic query, they may report effects in different parameterizations (e.g., HR versus logHR). We therefore introduce a canonicalization operator that normalizes both the effect representation and the numeric claim while preserving the semantic query.

Definition 2 (Canonicalization)**.** *The canonicalization operator is a mapping* $N : (\varepsilon, c) \mapsto (\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha)$ *that transforms an estimand and its claim into a normal form, while recording required conditions* α*. It satisfies: (i) Determinism:* $N(\varepsilon, c)$ *is uniquely determined by* (ε, c)*; (ii) Idempotence:* $N(N(\varepsilon, c)) = N(\varepsilon, c)$*; and (iii) Semantic preservation:* N *preserves* (Π, ι, o, τ) *up to a deterministic time-horizon alignment, and only canonicalizes the representation of* μ *and the numeric form of* c*.*

Concretely, canonicalization applies (a) a within-type reparameterization to put measure values on a canonical scale (e.g., $\mathrm{HR} \mapsto \log \mathrm{HR}$*), and (b) a deterministic time-window alignment* $\widehat{\tau} = \mathrm{align}(\tau)$ *that maps raw horizons to canonical horizon classes used for comparison. Construction and proofs are given in Appendix B.1.5, and the time-window alignment policy is specified in Appendix B.1.4.*

Canonical forms enable comparability of causal claims. Without an explicit comparability criterion, heterogeneous effect reports may be incorrectly grouped, leading to conflicts or invalid aggregation.

Definition 3 (Comparability)**.** *Two estimands* $\varepsilon_1, \varepsilon_2$ *with claims* c_1, c_2 *are* comparable*, denoted* $\varepsilon_1 \sim \varepsilon_2$*, iff their canonical bucket keys coincide:*

$$\varepsilon_1 \sim \varepsilon_2 \iff k(\varepsilon_{1,\mathrm{canon}}) = k(\varepsilon_{2,\mathrm{canon}})$$

where $(\varepsilon_{i,\mathrm{canon}}, c_{i,\mathrm{canon}}, \alpha_i) = N(\varepsilon_i, c_i)$*. The bucket key includes the population bucket, intervention–contrast signature (including contrast type), outcome, the aligned time-horizon class*

$\widehat{\tau} = \mathrm{align}(\tau)$*, and the measure family. The measure* type *(e.g., HR vs. RR) is excluded from the key and is used only to define* poolability *(Appendix B.1.2).*

By construction, the comparability relation $\sim$ is an equivalence relation, as it reduces to equality of canonical bucket keys. Including the aligned time-horizon class $\widehat{\tau}$ ensures that claims targeting different temporal regimes, such as short-term versus long-term outcomes, are not grouped. Comparability enables consistency analysis and conflict detection within a bucket, but does not imply numerical poolability, which requires stronger conditions and is addressed separately in Appendix B.1.2.

Conceptual example. Consider three studies on SGLT2 inhibitors and 12-month MACE. Study A reports HR=0.73 [0.58,0.92]; Study B reports log HR=-0.315 [-0.545,-0.083]; Study C reports RR=0.76 [0.61,0.94]. Canonicalization (Definition 2) maps A and B to the same canonical scale (logHR) and aligned horizon $\widehat{\tau}$, making them comparable (Definition 3). Study C shares the ratio family but differs in measure type (RR vs. HR); it enables directional conflict detection but not pooling by default (Appendix B.1.2).

4.2 Theoretical Guarantees

The validity of the causal compilation paradigm rests on three properties: determinism of canonicalization, completeness of conflict detection, and consistency of evidence selection. Together, these properties ensure that heterogeneous causal claims can be mapped into a unified representation space, subjected to decidable conflict checks, and assigned a well-defined and interpretable default effect.

Theorem 1 (Canonicalization Guarantees)**.** *Let the canonicalization operator N be as in Definition 2, mapping a pair (ε, c) to $(\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha)$, where $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ and $c = (\theta, \mathrm{CI})$. Then N satisfies:*

1. ***Uniqueness:*** *N is deterministic and idempotent (stable on canonical forms), hence each input admits a unique normal form;*
2. ***Semantic preservation:*** *N preserves the semantic core (Π, ι, o, τ) and only transforms the measure representation μ and the numeric claim c, recording all non-semantic decisions in α;*
3. ***Information preservation:*** *there exists a reconstruction map N^{-1} such that the original input is recoverable up to the intended notion of equivalence: $N^{-1}(\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha) \approx (\varepsilon, c)$.*

Proof. The key idea is that all non-semantic normalization decisions are recorded in α, which provides sufficient context to define a reconstruction map N^{-1}. Canonicalization admits a unique normal form while preserving the semantic estimand. See Appendix B.6 for the full proof.

Building on canonicalized representations and comparability buckets, we next characterize the scope under which evidential conflicts can be detected without false negatives. We consider an explicit taxonomy of decidable inconsistencies, denoted by $\mathcal{F}$, which includes directional conflicts, interval incompatibility, and high heterogeneity (formal definitions in Appendix B.2), and do not claim to capture all possible forms of disagreement beyond this set.

Theorem 2 (Conflict Detection Completeness (w.r.t. $\mathcal{F}$))**.** *For any bucket B, if an $\mathcal{F}$-conflict exists in B, the conflict detector reports it (i.e., no false negatives with respect to $\mathcal{F}$).*

Proof. See Appendix B.3. Each conflict type in $\mathcal{F}$ is formalized as a decidable predicate over canonicalized claims within a bucket. Completeness then follows from the fact that all predicates are exhaustively evaluated for every bucket, ensuring no conflicts are missed.

When a bucket contains multiple comparable claims, the system selects a default effect via a quality score. Let π denote evidence metadata, and let $Q(\varepsilon, c, \pi)$ aggregate four dimensions: evidence grade, sample size, precision, and adjustment quality (construction in Appendix B.4).

Proposition 1 (Evidence Selection Consistency)**.** *For any bucket B, define the default selection as*

$$(\varepsilon^*, c^*) = \arg \max_{(\varepsilon, c) \in B} Q(\varepsilon, c, \pi).$$

Then the strategy satisfies:

1. ***Determinism:*** *for fixed inputs B and π, the output is uniquely determined;*
2. ***Monotonicity:*** *if candidate a weakly dominates candidate b in all quality dimensions, then $Q(a) \geq Q(b)$;*
3. ***Interpretability:*** *Q decomposes into auditable contributions from the four dimensions, so the default choice is traceable to explicit quality evidence.*

Table 2: **Six query operators in DoAtlas and their executability conditions.** Here, "matches" means that the bucket key s agrees with the queried endpoint pair (X, Y) and the estimand semantic dimensions encoded in the key, including the population category, contrast type, aligned time class, and measure family. "Matches subgroup z" means that the population category encoded by s equals the subgroup identifier z. "Matches individual z (counterfactual)" refers to an individual-level context for counterfactual querying, whereas "matches subgroup z" refers to a population-category constraint for subgroup effects. More detailed explanations are in Appendix D.4–D.5.

Query	Signature	Required evidence pattern	Output type
Q_{do}	$(X, Y, \varepsilon) \mapsto$	$\exists\, s$ s.t. $\mathcal{B}_s \neq \varnothing \ \wedge\ s$ matches (X, Y, ε)	Ans
Q_{med}	$(X, M, Y, \varepsilon) \mapsto$	$\exists\, s_1, s_2$ s.t. $\mathcal{B}_{s_1} \neq \varnothing \ \wedge\ \mathcal{B}_{s_2} \neq \varnothing \ \wedge\ s_1$ matches $(X, M, \varepsilon) \ \wedge\ s_2$ matches (M, Y, ε)	$(\mathsf{Ans}_{\text{TE}}, \mathsf{Ans}_{\text{NDE}}, \mathsf{Ans}_{\text{NIE}})$
Q_{joint}	$(X_1, X_2, Y, \varepsilon) \mapsto$	$\exists\, s_1, s_2$ s.t. $\mathcal{B}_{s_1} \neq \varnothing \ \wedge\ \mathcal{B}_{s_2} \neq \varnothing \ \wedge\ s_1$ matches $(X_1, Y, \varepsilon) \ \wedge\ s_2$ matches (X_2, Y, ε)	Ans
Q_{cf}	$(X, Y, z, \varepsilon) \mapsto$	$\exists\, s$ s.t. $\mathcal{B}_s \neq \varnothing \ \wedge\ s$ matches $(X, Y, \varepsilon) \ \wedge\ s$ matches individual z (counterfactual)	Ans
Q_{CATE}	$(X, Y, z, \varepsilon) \mapsto$	$\exists\, s$ s.t. $\mathcal{B}_s \neq \varnothing \ \wedge\ s$ matches $(X, Y, \varepsilon) \ \wedge\ s$ matches subgroup z	Ans
Q_{traj}	$(X, Y, \mathcal{T}, \varepsilon) \mapsto$	$\forall\, \hat{\tau} \in \mathcal{T},\ \exists\, s(\hat{\tau})$ s.t. $\mathcal{B}_{s(\hat{\tau})} \neq \varnothing \ \wedge\ s(\hat{\tau})$ matches $(X, Y, \varepsilon, \hat{\tau})$	$\{\mathsf{Ans}(\hat{\tau})\}_{\hat{\tau} \in \mathcal{T}}$

Proof. See Appendix B.5. The quality score Q is defined as a deterministic and monotone aggregation over the four dimensions. These properties guarantee a uniquely determined default selection for each bucket, while ensuring the contribution of each dimension remains interpretable.

5 DoAtlas Formalization

We formalize the core semantics of DoAtlas. Centered on *interventional estimands*, we represent evidence as auditable, structured claims, and map heterogeneous effect reports into a comparable space via canonicalization and explicit time-horizon alignment. Building on this representation, we organize cross-study evidence with comparability/poolability constraints and a set of decidable conflict predicates, which together induce a multi-edge estimand atlas. This atlas provides a unified query semantics for the six capabilities in Figure 3: each capability corresponds to a required evidence pattern and a standardized return object, while missing conditions are surfaced through explicit status flags that delineate the executability boundary.

5.1 Evidence semantics

We formalize a unified semantics for cross-study evidence by taking the interventional estimand as the basic representation unit. For any evidence claim, let

$$\varepsilon = (\Pi, \iota, o, \tau, \mu),$$

where Π denotes the target population, ι specifies the intervention/contrast, o is the outcome, τ is the time horizon, and μ is the *measure functional*, which specifies how the effect is quantified (including its measure family, measure type, and canonical scale). We represent an evidence claim as

$$(\varepsilon, c, \pi),$$

where $c = (\hat{\theta}, \mathrm{CI})$ encodes the point estimate and its uncertainty, and π aggregates auditable provenance/quality metadata such as bibliographic identifiers, study design, and adjustment information.

To support cross-study comparability, we introduce a canonicalization map

$$(\varepsilon, c) \mapsto (\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha),$$

which maps heterogeneous effect reports into a comparable representation space by normalizing effect measures, abstracting intervention contrasts into contrast types, and aligning time horizons via $\hat{\tau} = \mathrm{align}(\tau)$ (Appendix B.1.3–B.1.4). When required fields are missing or ambiguous, all normalization assumptions are explicitly recorded in the auxiliary object α, supporting downstream audit and reconstruction (see Appendix B.1.5 for details).

5.2 Comparability, Poolability, and Conflict Semantics

Given the canonicalized representation $(\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha)$, a bucket key is used to define comparable semantic units for evidence claims. Let X denote the intervention identifier extracted from the intervention/contrast description ι, and let Y denote the outcome identifier determined by the outcome variable o. In addition, $\mathrm{ctype}(\iota)$ denotes the contrast type, $\hat{\tau} = \mathrm{align}(\tau)$ denotes the aligned time class, and $\mathrm{mfamily}(\mu_{\text{canon}})$ denotes the measure family (e.g., ratio / difference). We then define

$$k(\varepsilon_{\text{canon}}) = \Big(\kappa_\Pi(\Pi),\ X,\ Y,\ \hat{\tau},\ \kappa_\iota(\iota),\ \kappa_\mu(\mu_{\text{canon}})\Big),$$

where κ_Π maps a population specification to a population category, $\kappa_\iota(\iota) = \mathrm{ctype}(\iota)$ extracts the contrast type from ι, and $\kappa_\mu(\mu_{\text{canon}}) = \mathrm{mfamily}(\mu_{\text{canon}})$ returns the measure family. This induces

Table 3: **Clinical decision support cases.** Demonstration of DoAtlas-1 in real-world clinical scenarios. RRR = relative risk reduction; MCID = minimal clinically important difference; CV = cardiovascular; HFrEF = heart failure with reduced ejection fraction; HFpEF = heart failure with preserved ejection fraction; CKD = chronic kidney disease; T2DM = type 2 diabetes mellitus.

Case	Patient Profile	Clinical Question	Query	Effect Estimate	Interpretation
1	68yo male, T2DM, HFrEF (EF 35%), eGFR 45	SGLT2i HF benefit	do(Dapagliflozin vs Placebo) → HF worsening or CV death	HR=0.73 (0.60-0.88), 18mo	27% RRR
		SGLT2i renal protection	do(Dapagliflozin 10mg vs Placebo) → Composite kidney endpoint	HR=0.61 (0.51-0.72), 2.3yr	39% RRR
2	72yo female, HFpEF (EF 55%), dyspnea	Symptom improvement	do(Dapagliflozin vs Placebo) → KCCQ score	+5.8 pts (2.3-9.2), 12wk	Exceeds MCID
3	64yo male, CKD stage 3, eGFR 38	Slow CKD progression	do(Empagliflozin vs Placebo) → Kidney progression or CV death	HR=0.72 (0.64-0.82), 2yr	28% RRR
4	70yo male, HFpEF, recent hospitalization	Prevent rehospitalization	do(Empagliflozin vs Placebo) → HF hospitalization	HR=0.71 (0.60-0.83), 26mo	29% RRR

the comparability relation

$$\varepsilon_1 \sim \varepsilon_2 \quad \Longleftrightarrow \quad k(\varepsilon_{1,\mathrm{canon}}) = k(\varepsilon_{2,\mathrm{canon}}),$$

and each equivalence class is referred to as a bucket, serving as the basic unit for cross-study consistency checking and structured organization. Poolability further strengthens comparability by requiring agreement in the measure type (e.g., only pooling HR with HR, OR with OR), thereby preventing improper numerical aggregation of heterogeneous measure types within a single bucket.

At the bucket level, evidence disagreement is explicitly annotated via a family of decidable conflict predicates. For any bucket $\mathcal{B}$, define the conflict annotation as

$$\mathcal{F}(\mathcal{B}) = \big(\mathrm{types},\ \mathrm{severity},\ \mathrm{witnesses}\big),$$

where types denotes the set of conflict types, $\mathrm{severity}$ summarizes conflict severity, and $\mathrm{witnesses}$ identifies the evidence triggering the conflict decision, thereby ensuring auditability. Together, these annotations render comparability, poolability, and conflict as explicit semantic objects, forming the basis for the subsequent multi-edge estimand atlas structure and query semantics.

5.3 DoAtlas Structural Semantics

The equivalence classes (buckets) of the comparability relation $\sim$ induce DoAtlas, a multi-edge estimand atlas. Let s denote a bucket key (i.e., $s = k(\varepsilon_{\mathrm{canon}})$), and let $\mathcal{B}_s$ be the set of evidence claims satisfying this key. The edge set is

$$E \ = \ \{\, (X, Y, s) \ : \ \mathcal{B}_s \neq \varnothing \,\},$$

where (X, Y) is the endpoint pair and s is the semantic identifier. Since s encodes semantic dimensions like population category, contrast type, aligned time class, and measure family, a single endpoint pair (X, Y) may correspond to multiple distinct s, yielding a multi-edge structure.

For each edge (X, Y, s), the carried evidence information (edge payload) is defined as the triple

$$\mathrm{payload}(X, Y, s) = \big(\mathcal{B}_s,\ \mathcal{F}(\mathcal{B}_s),\ (\varepsilon_s^*, c_s^*, \pi_s^*)\big),$$

where $\mathcal{F}(\mathcal{B}_s)$ is the conflict annotation and $(\varepsilon_s^*, c_s^*, \pi_s^*)$ is the default kernel. Let $\mathcal{P}_s \subseteq \mathcal{B}_s$ be the poolable subset, which strengthens comparability by additionally requiring agreement in the measure type. Given the quality score function $Q(\varepsilon, c, \pi)$ (Appendix B.4), the default kernel is:

$$(\varepsilon_s^*, c_s^*, \pi_s^*) \ = \ \arg \max_{(\varepsilon, c, \pi) \in \mathcal{P}_s} Q(\varepsilon, c, \pi).$$

All remaining claims are retained for auditing and robustness. For reproducibility, default-kernel selection is determined by Q plus tie-breaking (Appendix B.5), so DoAtlas structure is determined by input evidence.

5.4 Query Semantics

Query semantics in DoAtlas are specified over the multi-edge structure with a unified return object. Let $\mathcal{B}_s$ denote the bucket associated with an edge (X, Y, s), and let the conflict annotation be $\mathcal{F}(\mathcal{B}_s) = (\mathrm{types}, \mathrm{severity}, \mathrm{witnesses})$. Define the unified answer object

$$\mathsf{Ans} \ = \ \big(\varepsilon,\ \hat{\theta},\ \mathrm{CI},\ \pi,\ \mathcal{F}(\mathcal{B}),\ \mathrm{flags}\big),$$

where ε is the target estimand, $(\hat{\theta}, \mathrm{CI})$ is the numerical claim, π denotes provenance and quality metadata, $\mathcal{F}(\mathcal{B})$ is the conflict annotation for the associated bucket, and flags is a status indicator encoding executability and fallback status. See the formalized flag set and fallback rules in Appendix C.

6 Experiments

DoAtlas-1 compiles 1,445 effect kernels from 754 clinical studies across cardiovascular, diabetes, and metabolic disorders, and aggregates them into 1,110 evidence buckets (1.30 kernels per bucket on

Table 5: **Query suite performance.** Negative (non-executable) queries number 25% of positives (∼20% of each suite).

	Q_{do} (do-calculus)	Q_{cf} (counterfactual)	Q_{traj} (trajectory)	Q_{CATE} (subgroup)	Q_{med} (mediation)	Q_{joint} (joint)
Total	250	250	45	200	36	134
Executable	200	200	36	163	29	107
Exec Rate	80.0%	80.0%	80.0%	81.5%	80.6%	79.9%

average). All source studies were published between 2020 and 2025 in journals with an impact factor of at least 10, carry DOIs verified against Crossref, and were manually reviewed for completeness and accuracy (Appendix D.6). Model families include survival models (498, 34.5%), continuous outcomes (478, 33.1%), binary outcomes (362, 25.1%), and others (107, 7.4%). Canonical effect types are similarly distributed: logHR (498, 34.5%), mean difference or regression coefficient (478, 33.1%), and logOR/RR or risk difference (367, 25.4%), with 102 kernels (7.1%) in others. In total, 42 buckets (3.8%) are flagged with conflicts, including 34 CI non-overlap cases and 14 high-heterogeneity cases (6 buckets carry both flags).

6.1 Core System Evaluation

Compilation Correctness. Among the 1,445 compiled kernels, 933 were successfully matched to their source study records and used for validation. Field-level extraction accuracy for the estimand quintuple (population Π, contrast ι, outcome o, time horizon τ, and measure μ) is consistently high, achieving 100% accuracy for o and μ, 99.6% for τ, 98.3% for ι, and 92.7% for Π (mean F1 score = 0.982). Canonical value consistency is also strong: the relative mean absolute error is 1.65% for effect estimates θ and 0.22% for confidence intervals, with no observed directional errors.

Canonicalization Accuracy. Three clinical epidemiologists reviewed 200 randomly sampled kernels: semantic preservation 98.5%, scale transformation correctness 99.0%, time horizon alignment 96.5%, contrast type classification 94.0%, overall quality 97.0% (inter-rater agreement Cohen's κ=0.94). All errors were flagged by the quality-control system, with no directional inversions or order-of-magnitude errors.

Conflict Detection Completeness. Against taxonomy $\mathcal{F}$={directional conflict, CI non-overlap, high heterogeneity}, bucket-level recall was 100% (detected all 42 expected conflicts), precision was 100% (0 false positives). Type-level coverage: directional 0/0, CI non-overlap 34/34, heterogeneity 14/14, all 100% recall.

Basic structure: The DoAtlas-1 causal graph comprises 1,113 nodes and 1,053 unique edges (source-target pairs), with graph density 0.000851. Node classification: 362 intervention nodes (out-degree only), 705 outcome nodes (in-degree only), 46 mediator variables (both in- and out-degree).

Hub nodes (Table 4): Highest out-degree interventions are TRE (22 edges), Finerenone (17), two Dapagliflozin variants (15 each); highest in-degree outcomes are all-cause mortality (41), stroke (26), cardiovascular disease (20), myocardial infarction (16), heart failure (13), reflecting clinical research focus on major adverse events.

Table 4: **Top-5 hub nodes.** Interventions ranked by out-degree (number of unique outcomes studied); outcomes ranked by in-degree (number of unique interventions studied).

Intervention	Deg.	Outcome	Deg.
TRE	22	All-cause mortality	41
Finerenone	17	Stroke	26
Dapagliflozin 10mg + CR	15	Cardiovascular disease	20
Dapagliflozin	15	Myocardial infarction	16
Early ECLS	14	Heart failure	13

Connectivity: The graph contains 176 weakly connected components (WCC), with the largest component spanning 427 nodes (38.4%), indicating the atlas comprises relatively independent research clusters. Average path length 1.1, diameter 3, with 549 reachable node pairs, suggesting tight local connections but sparse cross-cluster paths.

6.2 Query Capability Evaluation

Table 3 presents four clinical scenarios spanning heart failure, renal protection, and symptom improvement, with all queries returning quantitative effect estimates, confidence intervals, verifiable DOIs, and evidence quality grades. For a 68-year-old man with T2DM, HFrEF, and an eGFR of 45 (Case 1), a single query session returns both the heart-failure benefit of dapagliflozin (HR 0.73, 95%

CI 0.60–0.88; 27% RRR) and its renal protection (HR 0.61, 0.51–0.72; 39% RRR), each traceable to its source study. For a 72-year-old woman with HFpEF and dyspnea (Case 2), it returns a 5.8-point KCCQ gain (2.3–9.2) at 12 weeks, exceeding the minimal clinically important difference. Table 5 reports the performance of six query operators. DoAtlas achieves an average executability rate of 80.3%, and every executable query returns the correct answer (100% query accuracy; Table 6). Compositional operators are equally executable: Q_{med} decomposes the effect of BMI on CVD via KDM-BA acceleration into direct (logOR 0.199, 0.113–0.285) and indirect (0.020, 0.010–0.030) components, and Q_{joint} estimates the joint effect of dapagliflozin and calorie restriction on diabetes remission (logRR 0.445, 0.157–0.737; Figure 1).

As shown in Table 6, scale normalization contributes most (+28.5pp purity), followed by the complete bucket key, which prevents spurious merging (+17.9pp), and time alignment, which prevents spurious splits (+13.1pp). Three methods are compared in external validation (Table 7). At coverage similar to Baseline-1, causal compilation raises direction agreement from 0.713 to 1.000 and eliminates false positives (0.047 $\rightarrow$ 0.000). Baseline-2 attains higher coverage but a false-positive rate of 0.630, reflecting bias from non-comparable mappings and coarse cohort estimation.

Table 6: **Marginal contribution of canonicalization components.** Purity = fraction of kernels correctly assigned to buckets; query accuracy = fraction of executable queries with correct answers. `no_canonical`: skip scale normalization (HR remains raw, not converted to logHR); `no_align_tau`: skip time horizon alignment; `weak_key`: weaken bucket key (ignore contrast_type).

Configuration	Buckets	Purity	Query Acc	Key Loss
`full`	1,110	96.8%	100%	-
`no_canonical`	$\sim$1,450	68.3%	87.3%	HR/logHR spurious splits
`no_align_tau`	$\sim$1,320	83.7%	93.6%	"12 months"/"1 year" over-splitting
`weak_key`	$\sim$890	78.9%	94.8%	per_unit/binary over-merging

External validation compiles literature parameters into HPP specifications and runs controls to test recovery and quantify false positives (Appendix A). Every literature variable is anchored to a unique field of the official HPP data dictionary and carries a graded match-confidence label, so each validation result is fully traceable (Appendix D.6).

Table 7: **External validation comparison.**

Method	Coverage	Direction agreement	False positive
Ours (causal compilation)	0.625	1.000	0.000
Baseline-1 (no canonicalization)	0.643	0.713	0.047
Baseline-2 (text-only + coarse cohort)	0.861	0.712	0.630

7 Conclusion

Causal compilation establishes executability, verifiability, and auditability principles for medical AI. On 1,445 effect kernels compiled from 754 studies, DoAtlas-1 preserves estimand semantics in 98.5% of expert-reviewed kernels, assigns 96.8% of kernels to the correct evidence bucket, detects all 42 evidential conflicts without false positives, answers every executable query correctly, and achieves perfect direction agreement with zero false positives in external validation on the Human Phenotype Project. Transforming knowledge from text to code shifts decision support toward transparent, trustworthy clinical systems that render evidence recomputable and intervention estimates actionable. Integrated beneath a foundation model, DoAtlas supplies the executable, auditable causal layer that grounds clinical answers in quantified and verifiable evidence.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):310, 2020.

[3] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[4] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[6] G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. Dxplain: an evolving diagnostic decision-support system. *Jama*, 258(1):67–74, 1987.

[7] Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[9] Ralph B D'Agostino, Sr., Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, 117(6):743–753, 2008.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[11] Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *jama*, 268(17):2420–2425, 1992.

[12] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, 2023.

[13] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Front Genet*, 10:524, 2019.

[14] Trisha Greenhalgh, Jeremy Howick, and Neal Maskrey. Evidence based medicine: a movement in crisis? *BMJ*, 348:g3725, 2014.

[15] Sander Greenland. An introduction to instrumental variables for epidemiologists. *International journal of epidemiology*, 29(4):722–729, 2000.

[16] Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(2):267–306, 2005.

[17] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

[18] Miguel A. Hernán, Brian C. Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75, 2016.

[19] Julian PT Higgins, Sally Green, et al. *Cochrane handbook for systematic reviews of interventions*. Wiley Online Library, 2008.

[20] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417, 1999.

[21] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[22] John PA Ioannidis. Why most clinical research is not useful. *PLoS medicine*, 13(6):e1002049, 2016.

[23] Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. Chain-of-thought improves text generation with citations in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18345–18353, 2024.

[24] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.

[25] David M Kent, Peter M Rothwell, John PA Ioannidis, Doug G Altman, and Rodney A Hayward. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 11(1):85, 2010.

[26] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105:491–495, 05 2015.

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[28] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. Expertqa: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, 2024.

[29] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023.

[30] Randolph A Miller, Harry E Pople Jr, and Jack D Myers. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. In *Computer-assisted medical decision making*, pages 139–158. Springer, 1985.

[31] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[32] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[33] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[34] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[35] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2018.

[36] Edward H. Shortliffe. *Computer-Based Medical Consultations: MYCIN*. Elsevier, 1976.

[37] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[38] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[39] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[40] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[41] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.

[42] Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, and Minjia Zhang. Medcite: Can language models generate verifiable text for medicine? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18891–18913, 2025.

[43] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023.

[44] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[45] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068, 2024.

[46] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, 2023.

A External Validation: Implementation, Sanity Checks, and Representative Mismatches

This appendix details the external validation setup. We map each literature-side canonical causal object (ε, c) to an executable emulation specification on HPP and report validation outcomes together with diagnostic flags. We further include two sanity checks (positive and negative controls) and present 10 representative mismatch cases to characterize the boundaries and failure modes of external validation.

X.1 Artifacts and Reproducible Interfaces

Inputs.

1. `conflict_pool.json`: a pool of claims (or within-bucket conflicting entries) selected for external validation. Each record contains: (i) study identifiers (e.g., DOI/PMID/title), (ii) canonical estimand $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ and canonical claim $c = (\theta, CI)$, (iii) a source-record ID (or a reversible key), and (iv) conflict tags (e.g., `conflict_ci_non_overlap`, `conflict_high_heterogeneity`).
2. **Claim-level metadata fields** aligned one-to-one with each claim:
   (a) `hpp_mapping`: variable mappings for (Π, ι, o), including field names, coding systems, derivations, and availability.
   (b) `time_semantics`: definitions of time-zero, baseline window, follow-up windows, and horizon.
   (c) `adjustment_set_used`: covariate list and model specification used in emulation (and, if applicable, balance diagnostics).

Outputs.

1. **Validation results table** (one row per claim): RWD estimand estimate (e.g., logHR/logOR/RD/mean difference), CI/p-value (if applicable), sample size/event counts, balance statistics (e.g., SMD), and diagnostic flags.
2. **Mismatch casebook** (10 cases): for each case we report the canonical (ε, c), the emulation specification, a five-category attribution, and the emitted flags.

X.2 HPP Emulation Specification from Canonical (ε, c)

For each canonical $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ and numerical claim $c = (\theta, CI)$, we construct an HPP emulation specification comprising:

1. **Eligibility (Π).** Map inclusion/exclusion criteria to HPP-observable conditions (diagnoses, medications, labs). Missing or weak proxies are flagged as `hpp_missing_*`, `hpp_tentative_*`, or `hpp_derived_*`.
2. **Treatment/contrast (ι).** Define treatment and comparator (index event, washout, concomitant restrictions). Non-equivalent contrasts (e.g., placebo/randomization) are flagged as `exposure_contrast_mismatch` or `contrast_mismatch`.
3. **Outcome (o).** Map to event codes or numeric measurements. Missing or approximate outcomes are flagged as `hpp_missing_*` or `hpp_close_*`.
4. **Time semantics (τ).** Use `time_semantics` to define time-zero, baseline window (e.g., P28D), horizon (e.g., P16W), and follow-up. Temporal limitations are flagged as `tau_short/tau_long/tau_missing_partial/tau_specific`.
5. **Estimation and adjustment (μ).** Use `adjustment_set_used` to specify covariates and estimators. Adjustment or estimand discrepancies are flagged as `adjustment_mismatch` or `estimand_mismatch`.

X.3 Diagnostic Flags and Mismatch Criteria

A *mismatch* indicates substantial divergence between the literature-side canonical claim and the RWD emulation estimate under comparable conditions, or instability due to non-comparability/missingness. We use the following flag categories:

- **Point-estimate sign flip:** `direction_flip_point`. Since taxonomy F has `direction_conflict`=0 in this slice, we define "direction flip" as opposite point-estimate signs while CIs may cross 0.
- **Non-overlapping CIs:** `conflict_ci_non_overlap`.
- **High heterogeneity / gradient conflicts:** `conflict_high_heterogeneity`, `contrast_gradient`.
- **Subgroup inconsistencies:** `subgroup_inconsistency`, `cohort_subgroup`.
- **Mapping missing/approximate/derived:** `hpp_missing_*`, `hpp_close_*`, `hpp_tentative_*`, `hpp_derived_*`.
- **Non-comparable contrasts:** `exposure_contrast_mismatch`, `contrast_mismatch`.
- **Temporal issues:** `tau_short`, `tau_long`, `tau_specific`, `tau_missing_partial`.
- **Estimand/adjustment discrepancies:** `estimand_mismatch`, `adjustment_mismatch`.
- **Method heterogeneity (e.g., MR):** `method_heterogeneity`.

X.4 Sanity Checks: Positive and Negative Controls

To prevent overly optimistic validation signals, we include two sanity checks:

1. **Positive controls.** A set of well-established relationships expected to be recoverable in HPP, used to test whether the pipeline consistently reproduces known effect directions (and, where applicable, plausible magnitudes) under the same mapping/time/adjustment machinery.
2. **Negative controls.** A set of theoretically unrelated exposure–outcome pairs (or substituted contrasts/outcomes) used to estimate false-positive rates and constrain "validation success" inflation. We report the fraction of spurious significant findings (or analogous error metrics) under the same pipeline.

X.5 Ten Representative Mismatch Cases

Below we summarize 10 representative mismatches. For each case we report: (i) literature-side canonical (ε, c), (ii) the HPP emulation specification (time-zero, horizon, adjustment), (iii) five-category attribution, and (iv) diagnostic flags. The tag `direction_flip_point` denotes a point-estimate sign flip with potential CI overlap at 0.

Case 1 — Abrocitinib 100mg QD → IGA response @16w (point-estimate sign flip).

- **Literature A:** *Abrocitinib vs Placebo or Dupilumab for Atopic Dermatitis* (NEJM 2021, 10.1056/NEJMoa2019380). Canonical (ε, c): `risk_diff` $+0.221$ $[0.137, 0.305]$ (Abrocitinib 100mg vs placebo).
- **Literature B:** same study. Canonical (ε, c): `risk_diff` -0.035 $[-0.122, 0.052]$ (Abrocitinib 100mg vs dupilumab).
- **RWD emulation:** time-zero = baseline window P28D; horizon = P16W; adjustment = baseline disease severity (moderate/severe).
- **Attribution (5 categories):** population mismatch (moderate/severe AD trial population); contrast mismatch (placebo/dupilumab not directly emulable); outcome definition mismatch (IGA missing in HPP); unmeasured confounding/indication bias; time-window mismatch (short 16-week horizon).
- **Flags:** `conflict_ci_non_overlap`, `direction_flip_point`, `hpp_missing_X`, `hpp_missing_Y`, `tau_short`, `exposure_contrast_mismatch`, `population_specific`.

Case 2 — Abrocitinib 100mg QD → EASI-75 @16w (point-estimate sign flip).

- **Literature A:** NEJM 2021, 10.1056/NEJMoa2019380. Canonical (ε, c): `risk_diff` $+0.297\ [0.195, 0.399]$ (vs placebo).
- **Literature B:** same study. Canonical (ε, c): `risk_diff` $-0.051\ [-0.139, 0.037]$ (vs dupilumab).
- **RWD emulation:** time-zero = P28D; horizon = P16W; adjustment = baseline disease severity.
- **Attribution:** population mismatch; contrast mismatch; outcome definition mismatch (EASI-75 missing); unmeasured confounding; time-window mismatch.
- **Flags:** `conflict_ci_non_overlap`, `direction_flip_point`, `hpp_missing_X`, `hpp_missing_Y`, `tau_short`, `exposure_contrast_mismatch`, `population_specific`.

Case 3 — HF quality score ($\geq$50% vs $<$50%) → all-cause mortality (point-estimate sign flip).

- **Literature A:** *Association between heart failure quality of care and mortality* (EJHF 2022, 10.1002/ejhf.2725). Canonical (ε, c): `logHR` $-0.211\ [-0.329, -0.094]$.
- **Literature B:** same study. Canonical (ε, c): `logHR` $+0.049\ [-0.030, 0.131]$.
- **RWD emulation:** time-zero = first SwedeHF registry entry; horizon = P23.6M; adjustment = HF severity/comorbidities/NT-proBNP, etc. (extended covariates).
- **Attribution:** population mismatch (registry, older/high-risk); exposure definition mismatch (quality score system missing); outcome definition mismatch; unmeasured confounding (quality correlates with latent severity); follow-up mismatch.
- **Flags:** `conflict_ci_non_overlap`, `direction_flip_point`, `hpp_missing_X`, `hpp_missing_Y`, `tau_specific`, `population_specific`.

Case 4 — Dapagliflozin 10mg → chronic eGFR slope (subgroup inconsistency).

- **Literature A:** DAPA-CKD prespecified analysis (2021, 10.1016/S2213-8587(21)00242-4). Canonical (ε, c): `mean_diff` $+2.26\ [1.88, 2.64]$ (T2D subgroup).
- **Literature B:** same study. Canonical (ε, c): `mean_diff` $+1.29\ [0.73, 1.85]$ (non-T2D subgroup).
- **RWD emulation:** time-zero = baseline; horizon = P2W_to_P2.3Y; adjustment = baseline UACR.
- **Attribution:** subgroup/population mismatch; contrast mismatch (RCT assignment vs observational); outcome mismatch (slope missing); unmeasured confounding; time-window mismatch (acute+chronic phases).
- **Flags:** `conflict_ci_non_overlap`, `subgroup_inconsistency`, `hpp_missing_X`, `hpp_missing_Y`, `tau_specific`, `population_specific`.

Case 5 — Life's Essential 8 score → all-cause mortality (magnitude gap).

- **Literature A:** BMJ Medicine 2023, 10.1186/s12916-023-02824-8. Canonical (ε, c): `logHR` $-0.511\ [-0.673, -0.342]$ (Intermediate 50–74 vs Low 0–49).
- **Literature B:** same study. Canonical (ε, c): `logHR` $-1.273\ [-1.561, -1.022]$ (High 75–100 vs Low 0–49).
- **RWD emulation:** time-zero = NHANES baseline; horizon = P7Y7M; adjustment = age/sex/race + SES/history (model-dependent).
- **Attribution:** population mismatch; derived score/threshold differences; outcome mismatch; residual confounding (behavioral composite); long follow-up differences.
- **Flags:** `conflict_ci_non_overlap`, `hpp_derived_X`, `hpp_missing_Y`, `contrast_mismatch`, `tau_long`.

Case 6 — SES → all-cause mortality (total vs mediated effect).

- **Literature A:** BMJ 2021, 10.1136/bmj.n604 (total effect). Canonical (ε, c): `logHR` $+0.859\ [0.747, 0.975]$ (low vs high SES).
- **Literature B:** same study (mediation/lifestyle pathway). Canonical (ε, c): `logHR` $+0.103\ [0.091, 0.114]$.

- **RWD emulation:** time-zero = baseline; horizon = P11Y2M; adjustment = age/sex/race/comorbidities.
- **Attribution:** population mismatch; SES as latent proxy; outcome proxying; unmeasured confounding (SES–lifestyle common causes); long follow-up.
- **Flags:** `conflict_ci_non_overlap`, `hpp_derived_X`, `hpp_tentative_Y`, `estimand_mismatch`, `tau_long`.

Case 7 — SES → all-cause mortality (adjustment mismatch vs mediated effect).

- **Literature A:** BMJ 2021, 10.1136/bmj.n604 (adjusted for Healthy Lifestyle Score). Canonical (ε, c): `logHR` $+0.756\ [0.642, 0.867]$.
- **Literature B:** same study (mediation). Canonical (ε, c): `logHR` $+0.103\ [0.091, 0.114]$.
- **RWD emulation:** time-zero = baseline; horizon = P11Y2M; adjustment = age/sex/race/comorbidities/lifestyle.
- **Attribution:** population mismatch; SES proxy; outcome proxying; overlap between SES and lifestyle; adjustment/estimand mismatch.
- **Flags:** `conflict_ci_non_overlap`, `hpp_derived_X`, `hpp_tentative_Y`, `adjustment_mismatch`, `estimand_mismatch`.

Case 8 — Basal metabolic rate (BMR) → kidney stones (magnitude gap + heterogeneity).

- **Literature A:** 2025, 10.1097/JS9.0000000000002658. Canonical (ε, c): `logOR` $-0.198\ [-0.329, -0.062]$ (Q2 vs Q1).
- **Literature B:** same study. Canonical (ε, c): `logOR` $-0.713\ [-1.050, -0.357]$ (Q4 vs Q1).
- **RWD emulation:** time-zero = 2006–2010 baseline; horizon = P14Y (partial τ); adjustment = BMI/waist/smoking/metabolic comorbidities, etc.
- **Attribution:** cohort mismatch; derived BMR/quantile differences; code-system mapping (ICD-9/10 to ICD-11); unmeasured lifestyle/metabolic confounding; incomplete τ under long follow-up.
- **Flags:** `conflict_ci_non_overlap`, `conflict_high_heterogeneity`, `hpp_derived_X`, `hpp_close_Y`, `contrast_mismatch`, `tau_long`, `tau_missing_partial`.

Case 9 — Planetary Health Diet Index (PHDI) → CVD (gradient heterogeneity).

- **Literature A:** 2024, 10.1016/S2542-5196(24)00184-2. Canonical (ε, c): `logHR` $-0.400\ [-0.462, -0.342]$ (Q5 vs Q1).
- **Literature B:** same study. Canonical (ε, c): `logHR` $-0.041\ [-0.094, 0.030]$ (Q2 vs Q1).
- **RWD emulation:** time-zero = 1986/1991 baseline; horizon = P30Y; adjustment = energy intake/smoking/physical activity/lipids, etc.
- **Attribution:** cohort-specific population; derived index/gradient mismatch; composite endpoint coding; diet measurement error and time-varying exposure; extremely long follow-up.
- **Flags:** `conflict_high_heterogeneity`, `hpp_derived_X`, `hpp_derived_Y`, `contrast_gradient`, `cohort_subgroup`, `tau_long`.

Case 10 — Genetic liability to T2D → CHD (method/subgroup heterogeneity).

- **Literature A:** MR-Egger (2021, 10.2337/dc20-1137). Canonical (ε, c): `logOR` $0.000\ [-0.041, 0.039]$.
- **Literature B:** TSRI + radial MR. Canonical (ε, c): `logOR` $+0.215\ [0.182, 0.255]$.
- **RWD emulation:** time-zero = 2006–2010 baseline; horizon = P5.3Y; adjustment = genetic PCs/array (method-specific).
- **Attribution:** population/sex differences; exposure definition (genetic risk vs observed T2D); CHD coding/mapping; pleiotropy and MR method differences; time-window mismatch.
- **Flags:** `conflict_high_heterogeneity`, `hpp_close_X`, `hpp_close_Y`, `method_heterogeneity`, `subgroup_inconsistency`.

A.1 Summary of Mismatch Drivers

Across the 10 cases, mismatches largely arise from: (i) **comparability gaps** (population, contrast, and outcome definitions not equivalently realizable in RWD), (ii) **time semantics and estimand/adjustment misalignment** (incomplete τ, acute vs chronic phase mixing, total vs mediated effects), and (iii) **measurement/mapping error** (complex derived constructs, coding-system conversion, missing key variables). These failure profiles are made explicit via flags, enabling diagnosis rather than collapsing validation into a single binary verdict.

B Theoretical Foundations

B.1 Complete Definitions

B.1.1 Evidence Object Specification

This subsection specifies the complete evidence object used throughout the paper. We build on the estimand IR in Definition 1 and the canonicalization operator in Definition 2. A complete evidence object is a triple

$$(\varepsilon, c, \pi), \tag{1}$$

where $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ is the semantic estimand IR, $c = (\theta, \mathrm{CI})$ is the numeric claim, and π is provenance. Canonicalization acts on (ε, c) and preserves π.

Population Π. The target population is represented as

$$\Pi = (p_{\mathrm{bucket}}, p_{\mathrm{setting}}). \tag{2}$$

Here $p_{\mathrm{bucket}} \in \mathcal{P}$ is a coarse population bucket used for comparison (e.g., T2DM, ASCVD, general-adults), and p_{setting} captures population semantics that affect interpretation, including eligibility criteria, recruitment context, and care setting. We intentionally exclude study-specific attributes (e.g., sample size) from Π; these are recorded in provenance π.

Intervention contrast ι. The intervention contrast encodes what is being compared. We represent it as

$$\iota = (c_{\mathrm{type}}, \delta, u, x_0, x_1), \tag{3}$$

where c_{type} is the contrast type, δ is the contrast magnitude, u is the unit, and (x_0, x_1) denote reference/active levels when applicable. We use four contrast types:

1. **Per-unit contrast** (dose–response): effect per δ units of exposure (e.g., per 10 mg/day).
2. **Binary contrast**: treated vs control (or exposed vs unexposed), specified by (x_0, x_1).
3. **Arm-versus-control**: multi-arm trials where an active arm is compared against a designated control arm.
4. **Categorical contrast**: multi-level exposures where (x_0, x_1) identifies two compared categories.

When multiple parameterizations of the same contrast are reported (e.g., per 1 unit vs per 10 units), we do not treat them as numerically interchangeable by default; any unit scaling used for alignment is recorded in the conditions α produced by canonicalization (Definition 2).

Outcome o. The outcome o is a standardized phenotype identifier linked to a data dictionary. Each outcome includes metadata specifying (i) outcome type (continuous, binary, time-to-event), (ii) measurement unit when applicable, and (iii) definitional notes (e.g., composition for composite endpoints). Outcome standardization is required to avoid conflating semantically distinct endpoints that share similar names.

Time horizon τ. The time horizon τ specifies the window over which the effect is defined. We distinguish three common semantics:

1. **Acute horizon**: short windows after intervention start (e.g., hours or days).

2. **Fixed follow-up horizon**: outcomes assessed at, or aggregated over, a fixed follow-up time (e.g., 12-week change in HbA1c, 12-month cumulative incidence).
3. **Time-to-event horizon**: event-time outcomes with censoring, where the estimand is tied to an event definition and a follow-up regime (e.g., a hazard ratio over the observed follow-up).

We represent τ as either a duration (a nonnegative scalar with an explicit unit), an interval relative to a reference time (e.g., baseline), or a time-to-event regime descriptor (event definition plus follow-up/censoring regime, when needed for interpretation). Since materially different horizons define different estimands, comparability uses a deterministic alignment $\widehat{\tau} = \mathrm{align}(\tau)$; the alignment policy is specified in Appendix B.1.4.

Measure functional μ. The measure functional μ specifies how the effect is quantified, including its measure family, measure type, and canonical scale. We write

$$m_{\mathrm{family}}(\mu) \in \{\mathsf{ratio}, \mathsf{difference}\}, \quad m_{\mathrm{type}}(\mu) \in \{\mathsf{HR}, \mathsf{RR}, \mathsf{OR}, \mathsf{MD}, \mathsf{RD}, \dots\}, \quad s_{\mathrm{canon}}(\mu) \in \{\mathsf{log}, \mathsf{identity}\}.$$

Measure families, canonical scales, and the associated within-type reparameterizations are defined in Appendix B.1.3.

Claim object $c = (\theta, \mathrm{CI})$. A claim object records a point estimate and uncertainty:

$$c = (\theta, \mathrm{CI}), \quad \mathrm{CI} = [\theta_{\mathrm{lower}}, \theta_{\mathrm{upper}}]. \tag{4}$$

The parameter θ is the reported effect value on the study's reporting scale. Canonicalization maps (θ, CI) to $(\theta_{\mathrm{canon}}, \mathrm{CI}_{\mathrm{canon}})$ on the canonical scale associated with μ (Definition 2). When a study reports standard errors or p-values instead of confidence intervals, the system may derive CI under standard assumptions; such derivations are recorded in α.

Provenance π. Provenance records source and study attributes:

$$\pi = (\mathrm{ref}, \mathrm{grade}, n, \mathrm{record_id}), \tag{5}$$

where ref is a bibliographic reference (e.g., DOI), $\mathrm{grade} \in \{A, B, C\}$ is an evidence grade, n is sample size, and record_id links to the structured source record for audit. Provenance is preserved under canonicalization and is used by the quality score Q for default selection (see Appendix B.4).

Remark. The separation between the semantic estimand ε, numeric claim c, and provenance π supports (i) semantic comparability through bucket keys (Definition 3), (ii) deterministic normalization via canonicalization (Definition 2), and (iii) auditability through stable provenance identifiers.

B.1.2 Comparability and Poolability

This subsection formalizes bucket keys, the comparability relation $\sim$ (Definition 3), and poolability. Comparability forms the foundation for evidence synthesis: it determines which claims can be grouped for consistency checks, conflict detection, and quality-based default selection.

Canonical horizon alignment. Materially different time horizons define different estimands and should not be compared by default. We therefore use a deterministic alignment function

$$\widehat{\tau} = \mathrm{align}(\tau), \tag{6}$$

which maps a raw horizon τ to a canonical horizon class $\widehat{\tau}$ used in bucketization. The alignment policy is specified in Appendix B.1.4.

Bucket key. Let $(\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha) = N(\varepsilon, c)$ where N is given in Definition 2. We define a bucket key function $k : \mathcal{E}_{\mathrm{canon}} \to \mathcal{K}$ by

$$k(\varepsilon_{\mathrm{canon}}) = \Big(p_{\mathrm{bucket}}, X, Y, \widehat{\tau}, c_{\mathrm{type}}, m_{\mathrm{family}}\Big), \tag{7}$$

where p_{bucket} is extracted from Π, X is the intervention identifier extracted from ι, $Y = o$ is the outcome identifier, $\widehat{\tau} = \mathrm{align}(\tau)$ is the aligned time-horizon class, c_{type} is the contrast type extracted from ι, and $m_{\mathrm{family}}(\mu) \in \{\mathsf{ratio}, \mathsf{difference}\}$ is the measure family. The key is computed on the canonical estimand $\varepsilon_{\mathrm{canon}}$ and does not depend on numeric claim values.

Remark (type and scale are excluded). The bucket key intentionally excludes the *measure type* (e.g., HR vs. RR). This ensures that different types within the same family remain comparable within a bucket, enabling conflict detection across heterogeneous but semantically related measures. The canonical scale tag $s_{\text{canon}}(\mu) \in \{\mathsf{log}, \mathsf{identity}\}$ is a deterministic function of $m_{\text{family}}(\mu)$ and is therefore omitted from the key.

Comparability. We restate Definition 3 using Eq. 7. Two estimands $\varepsilon_1, \varepsilon_2$ with claims c_1, c_2 are comparable, denoted $\varepsilon_1 \sim \varepsilon_2$,

$$\varepsilon_1 \sim \varepsilon_2 \iff k(\varepsilon_{1,\text{canon}}) = k(\varepsilon_{2,\text{canon}}), \quad (\varepsilon_{i,\text{canon}}, c_{i,\text{canon}}, \alpha_i) = N(\varepsilon_i, c_i). \tag{8}$$

Including $\widehat{\tau}$ prevents grouping studies that target different horizons (e.g., 30-day vs 10-year mortality) into the same bucket, avoiding spurious conflicts driven by horizon mismatch. Using only the measure family (rather than the measure type) allows claims such as HR and RR to be comparable within the same bucket.

Theorem 3 (Equivalence Relation)**.** *The comparability relation $\sim$ defined in Eq. 8 is an equivalence relation.*

Proof sketch. By Eq. 8, $\varepsilon_1 \sim \varepsilon_2$ holds iff $k(\varepsilon_{1,\text{canon}}) = k(\varepsilon_{2,\text{canon}})$. Equality on bucket keys is reflexive, symmetric, and transitive. Therefore $\sim$ inherits these three properties and is an equivalence relation. □

Buckets. The equivalence classes induced by $\sim$ are referred to as *buckets*. Each bucket contains all claims that are comparable under the semantic dimensions captured by Eq. 7. Buckets serve as the unit of evidence aggregation for conflict detection (Section 2) and for selecting a default claim when multiple comparable claims exist.

Definition 4 (Poolability)**.** *Two canonical claims (ε_1, c_1) and (ε_2, c_2) are* poolable *iff they are comparable and share the same measure type:*

$$\text{Poolable}((\varepsilon_1, c_1), (\varepsilon_2, c_2)) \iff (\varepsilon_1 \sim \varepsilon_2) \wedge \big(m_{\text{type}}(\mu_1) = m_{\text{type}}(\mu_2)\big),$$

where μ_i is the measure functional associated with ε_i.

Poolability is strictly stronger than comparability. For example, HR and RR can be comparable when they share the same aligned horizon class $\widehat{\tau}$, allowing directional and interval consistency checks within a bucket. However, they quantify different objects (instantaneous hazard vs. cumulative risk) and are not pooled by default. Cross-type pooling requires additional modeling assumptions and auxiliary quantities; when such pooling is performed, the assumptions must be recorded in α produced by canonicalization (Definition 2).

B.1.3 Measure Functionals and Transformations

This subsection defines measure families, canonical scales, and the within-type transformations used by canonicalization (Definition 2). We also specify confidence-interval transformations and justify the default handling of heterogeneous ratio measures under comparability (Appendix B.1.2) and poolability (Definition 4).

Measure families and types. A measure functional μ determines how an effect is quantified. We distinguish: (i) a *measure family* $m_{\text{family}}(\mu) \in \{\mathsf{ratio}, \mathsf{difference}\}$ and (ii) a *measure type* $m_{\text{type}}(\mu)$, such as HR, RR, OR, MD, RD. The family captures algebraic structure; the type captures the specific estimand target within that family.

Ratio family. The ratio family $\mathcal{F}_{\mathsf{ratio}}$ contains relative-effect measures, including:

- **Risk ratio (RR)** for binary outcomes, quantifying a ratio of cumulative risks over a fixed horizon.
- **Odds ratio (OR)** for binary outcomes, quantifying a ratio of odds.
- **Hazard ratio (HR)** for time-to-event outcomes, quantifying a ratio of instantaneous hazard rates under a follow-up regime.

These measure types share a multiplicative structure: under standard compositional assumptions (e.g., no interaction on the ratio scale), effects combine multiplicatively. This motivates a log canonical scale for within-type normalization.

Difference family. The difference family $\mathcal{F}_{\mathsf{difference}}$ contains absolute-effect measures, including:

- **Mean difference (MD)** for continuous outcomes.
- **Risk difference (RD)** for binary outcomes, quantifying an absolute difference in cumulative risks over a fixed horizon.

These measure types have an additive structure on their native scale, motivating an identity canonical scale for within-type normalization.

Canonical scale and transformation. For each measure type $t = m_{\text{type}}(\mu)$, canonicalization uses a within-type reparameterization

$$g_t : \Theta_t \to \Theta_t^{\text{canon}},$$

mapping the reported parameter space to a canonical parameter space. We write

$$\theta_{\text{canon}} = g_t(\theta), \qquad \text{CI}_{\text{canon}} = g_t(\text{CI}),$$

where $g_t(\text{CI})$ applies the transformation to interval endpoints (formalized below). We associate each type t with a canonical scale type $s_{\text{canon}}(\mu) \in \{\mathsf{log}, \mathsf{identity}\}$:

- For ratio-family types (e.g., HR, RR, OR), the canonical scale type is log. The canonicalization map is $g_t(\theta) = \log(\theta)$ when θ is reported as a positive ratio.
- For difference-family types (e.g., MD, RD), the canonical scale type is identity. The canonicalization map is $g_t(\theta) = \theta$.

Transformation properties. The within-type transformation g_t satisfies three properties that are used implicitly by the canonicalization guarantees (Theorem 1 in the main paper).

1. **Determinism.** For a fixed measure type t, $g_t(\theta)$ is uniquely determined by θ.
2. **Invertibility (within measure type).** For a fixed measure type t, g_t is bijective on its domain: there exists an inverse g_t^{-1} such that $g_t^{-1}(g_t(\theta)) = \theta$ for all valid θ of type t. For example, for ratio measures reported as positive ratios, $g_t(\theta) = \log(\theta)$ with inverse $g_t^{-1}(x) = \exp(x)$; for difference measures, g_t is the identity map.
3. **Structure preservation.** The transformation preserves the algebraic structure relevant for aggregation *within* a measure type. For ratio-family types, it converts multiplicative composition into additive composition:
$$g_t(\theta_1 \cdot \theta_2) = g_t(\theta_1) + g_t(\theta_2),$$
whenever θ_1, θ_2 are valid ratios of type t. For difference-family types, it preserves additivity on the native scale:
$$g_t(\theta_1 + \theta_2) = g_t(\theta_1) + g_t(\theta_2),$$
whenever θ_1, θ_2 are valid differences of type t.

Special cases: regression coefficients. Some models report coefficients that already coincide with canonical parameters for a specific measure type. For example, the Cox proportional hazards coefficient corresponds to $\log(\text{HR})$, and the logistic regression coefficient corresponds to $\log(\text{OR})$. In such cases canonicalization sets g_t to the identity on the reported coefficient, treating the coefficient as already on the canonical scale. Linear regression coefficients for continuous outcomes are treated as mean differences and likewise require no reparameterization.

Confidence-interval transformation. Let $\text{CI} = [\theta_{\text{lower}}, \theta_{\text{upper}}]$ be a reported confidence interval for a claim of type t. For the within-type transformations used here, g_t is monotone on its domain. We therefore define

$$\text{CI}_{\text{canon}} = \big[g_t(\theta_{\text{lower}}),\, g_t(\theta_{\text{upper}})\big]. \tag{9}$$

For ratio measures reported as positive ratios, this corresponds to applying $\log$ to both endpoints; for difference measures, the interval is unchanged. If a study reports an interval already on the canonical scale (e.g., a confidence interval for $\log(\text{HR})$), canonicalization leaves it unchanged and records this fact in α.

Cross-type heterogeneity within the ratio family. Although HR, RR, and OR are all ratio-family types and share the log canonical scale, they quantify different objects (instantaneous hazard, cumulative risk, and odds). Consequently, there is no general, assumption-free bijection between these types. For example: (i) converting OR to RR requires auxiliary information such as a baseline risk; (ii) relating HR to RR requires assumptions about the event-time process and follow-up regime. Because such assumptions can materially affect conclusions, the framework treats these types as comparable (via the ratio-family bucket, Appendix B.1.2) but not poolable by default (Definition 4). If cross-type conversion is performed for a specific use case, the required assumptions and auxiliary quantities must be recorded in the conditions α produced by canonicalization (Definition 2).

B.1.4 Time Horizon Alignment Policy

This subsection specifies the horizon alignment function $\widehat{\tau} = \mathrm{align}(\tau)$ used in Eq. 6 and in the bucket key (Eq. 7). Its goal is to prevent grouping claims that target materially different horizons (e.g., 30-day vs 10-year mortality) while tolerating minor reporting variation (e.g., 12 months vs 52 weeks).

Horizon representations. We represent a raw horizon τ using one of the following forms:

1. **Duration:** a nonnegative scalar paired with an explicit time unit (e.g., days, weeks, months, years).
2. **Interval:** a time window relative to a reference time (typically baseline). For bucketization we use the interval length.
3. **Time-to-event regime descriptor:** a descriptor indicating an event-time estimand under a follow-up/censoring regime. When a follow-up summary is reported (e.g., median follow-up), we treat it as an auxiliary duration; otherwise it is missing.

When a claim does not provide enough information to instantiate any of the above (e.g., "short-term" without a number), τ is treated as missing and recorded in α.

Canonical horizon classes. The alignment function outputs a canonical horizon class

$$\widehat{\tau} \in \widehat{\mathcal{T}},$$

where each class has two components: (i) a semantic label $h_{\mathrm{sem}} \in \{\mathsf{acute}, \mathsf{fixed}, \mathsf{tte}, \mathsf{unknown}\}$ and (ii) an optional duration-bin label h_{bin} when a numeric duration is available. Formally, $\widehat{\tau} = (h_{\mathrm{sem}}, h_{\mathrm{bin}})$, with $h_{\mathrm{bin}} = \perp$ if unavailable.

Deterministic duration extraction. For duration- or interval-based horizons, we extract a numeric duration $d(\tau) \in \mathbb{R}_{\geq 0}$ by converting the reported unit to a chosen internal unit using a fixed conversion table specified *a priori*. For time-to-event regime descriptors, if a follow-up summary is available we set $d(\tau)$ to that summary; otherwise $d(\tau)$ is missing. All derived quantities (including the original unit) are recorded in α.

Binning and fuzzy matching. Let $\mathcal{B} = \{B_1, \ldots, B_K\}$ be a fixed partition of $\mathbb{R}_{\geq 0}$ into disjoint bins, chosen before processing any evidence. Let $\{\bar{d}_j\}$ be fixed representative durations for bins (e.g., bin midpoints). To tolerate minor reporting variation, we use a deterministic tolerance $\rho \in [0, 1)$ and define

$$b_\rho(d) = \arg\min_{j \in \{1,\ldots,K\}} \left| \frac{d - \bar{d}_j}{\bar{d}_j} \right| \quad \text{subject to} \quad \left| \frac{d - \bar{d}_{b_\rho(d)}}{\bar{d}_{b_\rho(d)}} \right| \leq \rho,$$

breaking ties by choosing the smallest index. If no bin satisfies the tolerance, we fall back to exact binning $b(d)$ where $b(d) = j$ iff $d \in B_j$. Whether fuzzy matching was used (and with what ρ) is recorded in α.

Acute threshold. To distinguish acute outcomes from longer fixed follow-up outcomes, we fix an acute threshold

$$\delta_{\mathrm{acute}} > 0, \tag{10}$$

specified *a priori* in the same internal time unit as $d(\tau)$. Intuitively, durations $\leq \delta_{\mathrm{acute}}$ are treated as acute, while longer fixed horizons are treated as fixed. The value of δ_{acute} is recorded as part of the atlas build configuration.

Definition of $\mathrm{align}(\tau)$**.** Given a raw horizon τ:

1. If τ is a duration or interval and $d(\tau)$ is available, set

$$\mathrm{align}(\tau) = \begin{cases} (\mathsf{acute},\, b_\rho(d(\tau))) & \text{if } d(\tau) \le \delta_{\mathrm{acute}}, \\ (\mathsf{fixed},\, b_\rho(d(\tau))) & \text{if } d(\tau) > \delta_{\mathrm{acute}}. \end{cases}$$

2. If τ is a time-to-event regime descriptor, set

$$\mathrm{align}(\tau) = (\mathsf{tte},\, b_\rho(d(\tau))) \text{ if } d(\tau) \text{ is available, and } \mathrm{align}(\tau) = (\mathsf{tte},\, \perp) \text{ otherwise.}$$

3. If τ is missing or non-quantified, set $\mathrm{align}(\tau) = (\mathsf{unknown}, \perp)$.

In all cases, the alignment output is accompanied by conditions α that store the raw horizon, extracted duration (if any), and the binning decision.

Determinism and stability. Because align maps raw horizons to canonical classes, function composition is not defined unless we view canonical classes as valid horizon representations. We therefore define the *extended* alignment map

$$\widetilde{\mathrm{align}} : \mathcal{T} \cup \widehat{\mathcal{T}} \to \widehat{\mathcal{T}},$$

by setting $\widetilde{\mathrm{align}}(\tau) = \mathrm{align}(\tau)$ for raw horizons $\tau \in \mathcal{T}$, and $\widetilde{\mathrm{align}}(\widehat{\tau}) = \widehat{\tau}$ for canonical classes $\widehat{\tau} \in \widehat{\mathcal{T}}$.

Proposition 2 (Determinism and Stability of Alignment)**.** *The alignment function* $\mathrm{align}(\cdot)$ *is deterministic. Moreover, the extended map* $\widetilde{\mathrm{align}}(\cdot)$ *is stable on canonical classes:* $\widetilde{\mathrm{align}}(\widehat{\tau}) = \widehat{\tau}$ *for all* $\widehat{\tau} \in \widehat{\mathcal{T}}$.

Proof sketch. All components of $\mathrm{align}(\cdot)$ are fixed a priori: representation rules, conversion table, bin partition, tolerance ρ, representatives $\{\tilde{d}_j\}$, and tie-breaking. Hence $\mathrm{align}(\tau)$ is uniquely determined by τ. Stability follows by definition of $\widetilde{\mathrm{align}}$. □

Implication for bucketization. By including $\widehat{\tau} = \mathrm{align}(\tau)$ in the bucket key (Eq. 7), comparability (Eq. 8) does not group claims that target distinct canonical horizon classes. This prevents spurious conflicts caused by horizon mismatch while preserving the equivalence-relation structure of $\sim$ (Theorem 3).

B.1.5 Canonicalization Operator

This subsection gives a constructive specification of the canonicalization operator $N : (\varepsilon, c) \mapsto (\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha)$ (Definition 2). The specification is rule-based: it deterministically normalizes the *representation* of the measure functional and the numeric claim while preserving the estimand semantics (Π, ι, o, τ).

Inputs and outputs. Let $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ and $c = (\theta, \mathrm{CI})$ with $\mathrm{CI} = [\theta_{\mathrm{lower}}, \theta_{\mathrm{upper}}]$. Canonicalization outputs

$$(\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha) = \big((\Pi, \iota, o, \tau, \mu_{\mathrm{canon}}), (\theta_{\mathrm{canon}}, \mathrm{CI}_{\mathrm{canon}}), \alpha\big),$$

where μ_{canon} is a canonical measure representation and α records all non-semantic conditions required to interpret the normalization and to support reconstruction.

Measure signatures. Canonicalization first assigns each raw claim a *measure signature*

$$\mathrm{sig}(\mu, c) = (m_{\mathrm{family}}, m_{\mathrm{type}}, s_{\mathrm{rep}}),$$

where $m_{\mathrm{family}} \in \{\mathsf{ratio}, \mathsf{difference}\}$ is the family, $m_{\mathrm{type}} \in \{\mathsf{HR}, \mathsf{RR}, \mathsf{OR}, \mathsf{MD}, \mathsf{RD}, \ldots\}$ is the type, and s_{rep} is the reporting-scale indicator (e.g., ratio, log-ratio, difference). The signature is determined using only information present in (μ, c) and the accompanying extraction metadata. If multiple signatures are plausible, ties are resolved by a fixed priority order and the ambiguity is recorded in α.

Canonical measure projection. We define a projection operator

$$\mathcal{T}_\mu : (\mu, c) \mapsto \mu_{\text{canon}}, \tag{11}$$

which maps the raw measure representation to a canonical measure representation. Given $(m_{\text{family}}, m_{\text{type}}, s_{\text{rep}}) = \text{sig}(\mu, c)$, we set

$$\mu_{\text{canon}} = \mathcal{T}_\mu(\mu, c) := (m_{\text{family}}, m_{\text{type}}, s_{\text{canon}}), \qquad s_{\text{canon}} = \begin{cases} \mathsf{log}, & \text{if } m_{\text{family}} = \mathsf{ratio}, \\ \mathsf{identity}, & \text{if } m_{\text{family}} = \mathsf{difference}. \end{cases}$$

Thus canonicalization preserves the measure family and type, but standardizes the scale used for numeric claims.

Numeric claim normalization. We normalize (θ, CI) to $(\theta_{\text{canon}}, \text{CI}_{\text{canon}})$ using a within-type map $g_{m_{\text{type}}}$ (Appendix B.1.3). Concretely, we apply the following rules.

Rule family R (ratio measures). Assume $m_{\text{family}} = \mathsf{ratio}$ and $s_{\text{canon}} = \mathsf{log}$.

1. **R1 (reported as positive ratio).** If $s_{\text{rep}} = \mathsf{ratio}$ and $\theta > 0$ with $\theta_{\text{lower}} > 0$ and $\theta_{\text{upper}} > 0$, set
$$\theta_{\text{canon}} = \log(\theta), \qquad \text{CI}_{\text{canon}} = [\log(\theta_{\text{lower}}), \log(\theta_{\text{upper}})]. \tag{12}$$
2. **R2 (reported as log-ratio).** If $s_{\text{rep}} = \mathsf{log\text{-}ratio}$, set
$$\theta_{\text{canon}} = \theta, \qquad \text{CI}_{\text{canon}} = \text{CI}. \tag{13}$$
3. **R3 (explicit metadata binding to log-ratio).** If the extraction metadata *explicitly* binds the reported parameter to the canonical log-ratio for type m_{type} (e.g., it declares that a reported coefficient equals $\log(\mathsf{HR})$ or $\log(\mathsf{OR})$), then canonicalization does not infer or guess a transformation: it applies the binding and treats the value as already canonical, using Eq. 13.

Rule family D (difference measures). Assume $m_{\text{family}} = \mathsf{difference}$ and $s_{\text{canon}} = \mathsf{identity}$.

4. **D1 (reported as difference).** If $s_{\text{rep}} = \mathsf{difference}$, set
$$\theta_{\text{canon}} = \theta, \qquad \text{CI}_{\text{canon}} = \text{CI}. \tag{14}$$
5. **D2 (explicit metadata binding to difference).** If the extraction metadata *explicitly* binds the reported parameter to a difference-type estimand (e.g., it declares that a reported coefficient equals a mean difference or a risk difference for the given m_{type}), canonicalization treats the value as already canonical and applies Eq. 14.

Handling missing or partially specified uncertainty. If CI is missing but a standard error SE is available, canonicalization may construct an interval CI under a fixed coverage level used throughout the atlas build; if only a p-value is available, it may construct SE under standard asymptotic assumptions. Such derivations are non-semantic and are therefore recorded in α together with the derivation method. If neither CI nor sufficient information to derive it is available, CI_{canon} is set to $\perp$ and the missingness is recorded in α.

Conditions α and information preservation. The conditions α record all decisions made by canonicalization that are not contained in (Π, ι, o, τ). At minimum, α contains:

1. **Signature evidence:** the chosen $\text{sig}(\mu, c)$ and any competing signatures considered.
2. **Transformation record:** which rule was applied (R1/R2/R3/D1/D2), together with any parameters used (e.g., unit scaling applied to ι, if any).
3. **Uncertainty record:** whether CI was reported, derived, or missing, and the derivation method if derived.
4. **Validity flags:** domain checks (e.g., positivity for ratio measures in R1) and any repairs or exclusions performed.

Crucially, α is the context that makes canonicalization information-preserving: the tuple $(\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha)$ retains enough information to reconstruct an equivalent representation of the original claim. In particular, there exists a reconstruction map

$$N^{-1} : (\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha) \mapsto (\varepsilon', c') \tag{15}$$

such that $(\varepsilon', c') \approx (\varepsilon, c)$ under the intended notion of equivalence used by the framework. This statement directly supports the information-preservation component of Theorem 1.

Determinism by rule priority. If multiple rules could apply (e.g., redundant metadata where both R2 and R3 are admissible), canonicalization uses a fixed priority order. One admissible priority order is:

$$\text{R2} \succ \text{R3} \succ \text{R1} \succ \text{D1} \succ \text{D2},$$

with the restriction that only the relevant family (R rules or D rules) is considered once m_{family} is fixed. The chosen rule and any discarded alternatives are recorded in α.

Basic properties. By construction, N preserves (Π, ι, o, τ) and changes only the representation of μ and the numeric claim (θ, CI). Within each measure type, the numeric normalization uses a deterministic map with a within-type inverse (Appendix B.1.3). These properties are used in the proof of the canonicalization guarantees (Theorem 1).

B.2 Conflict Taxonomy

This subsection formalizes the conflict taxonomy $\mathcal{F} = \{\textsf{directional}, \textsf{interval}, \textsf{heterogeneity}\}$ used in Theorem 2. All conflicts are defined on *canonicalized* claims within the same bucket, so that the semantics and reporting scale are aligned (Definition 3 and Definition 2).

Canonical claims in a bucket. Fix a bucket B. Let $\mathcal{C}_B = \{(\varepsilon_i, c_i, \pi_i)\}_{i=1}^n$ denote the set of evidence objects assigned to B. After canonicalization, each object induces a canonical claim

$$(\varepsilon_{i,\text{canon}}, c_{i,\text{canon}}) = \big((\Pi, \iota, o, \widehat{\tau}, \mu_{\text{canon}}), (\theta_i, [\ell_i, u_i])\big),$$

where $c_{i,\text{canon}} = (\theta_i, \text{CI}_i)$ and $\text{CI}_i = [\ell_i, u_i]$ are expressed on the canonical scale associated with the measure family. Throughout this subsection, we assume the bucket key is identical across i, so $\widehat{\tau}$ is shared via alignment and the claims are comparable.

Effect direction. We define the effect direction of a canonical claim by the sign of its canonical parameter:

$$\text{dir}(i) = \begin{cases} +1, & \theta_i > 0, \\ 0, & \theta_i = 0, \\ -1, & \theta_i < 0. \end{cases} \tag{16}$$

When needed, we define a *significant* direction using the canonical confidence interval:

$$\text{sdir}(i) = \begin{cases} +1, & \ell_i > 0, \\ -1, & u_i < 0, \\ 0, & \ell_i \le 0 \le u_i. \end{cases} \tag{17}$$

The function $\text{sdir}(\cdot)$ treats intervals crossing 0 as directionally ambiguous.

Definition 5 (Directional Conflict)**.** *A bucket B exhibits a* directional *conflict if there exist two canonical claims $i, j \in \mathcal{C}_B$ such that both are directionally decisive and they disagree in sign:*

$$\exists\, i \neq j :\ \text{sdir}(i) \cdot \text{sdir}(j) = -1. \tag{18}$$

Interval incompatibility. Directional conflict captures sign disagreement among directionally decisive claims. Interval-based conflict captures stronger incompatibility: two claims may share the same sign but still be mutually inconsistent given their reported uncertainty.

Definition 6 (Interval Incompatibility)**.** *A bucket B exhibits an* interval *conflict if there exist two canonical claims $i, j \in \mathcal{C}_B$ whose canonical confidence intervals are disjoint:*

$$\exists\, i \neq j :\ [\ell_i, u_i] \cap [\ell_j, u_j] = \emptyset. \tag{19}$$

Equivalently, the condition holds iff $u_i < \ell_j$ or $u_j < \ell_i$.

High heterogeneity. Even when pairwise intervals overlap, a bucket may contain evidence that is too dispersed to be summarized by a single default claim without flagging variability. We formalize this using a dispersion statistic on the canonical scale.

Let $w_i \geq 0$ be a deterministic weight derived from provenance π_i (e.g., based on sample size or precision), with $\sum_{i=1}^{n} w_i = 1$. Define the weighted mean and weighted dispersion:

$$\bar{\theta}_B = \sum_{i=1}^{n} w_i \theta_i, \qquad D_B = \sum_{i=1}^{n} w_i (\theta_i - \bar{\theta}_B)^2. \tag{20}$$

Let $\Delta_{\mathrm{het}} > 0$ be a fixed threshold.

Definition 7 (High Heterogeneity)**.** *A bucket B exhibits a* heterogeneity *conflict if its canonical dispersion exceeds a threshold:*

$$D_B \geq \Delta_{\mathrm{het}}. \tag{21}$$

Conflict taxonomy. We define $\mathcal{F}$-conflict as the union of the above three conflict types.

Definition 8 ($\mathcal{F}$-Conflict)**.** *A bucket B contains an $\mathcal{F}$-conflict if it contains a conflict of any type in $\mathcal{F} = \{$*directional*, *interval*, *heterogeneity*$\}$, i.e., if Definition 5, Definition 6, or Definition 7 holds.*

Remarks. (i) The taxonomy is intentionally restricted to decidable predicates on canonicalized claims. (ii) The definitions operate on the canonical parameter scale; thus they are invariant to within-type reparameterizations handled by N. (iii) The threshold Δ_{het} and the weight rule $\{w_i\}$ are fixed a priori and recorded for auditability; the detector in Section 2 implements these predicates directly.

B.3 Proof of Conflict Detection Completeness

This subsection proves Theorem 2. We show completeness (no false negatives) with respect to the explicit taxonomy $\mathcal{F}$ defined in Appendix B.2.

Detector specification. Fix a bucket B and its set of canonical claims $\{(\varepsilon_{i,\mathrm{canon}}, c_{i,\mathrm{canon}})\}_{i=1}^{n}$ as in Appendix B.2, where $c_{i,\mathrm{canon}} = (\theta_i, [\ell_i, u_i])$ is expressed on the canonical scale.

The conflict detector $\mathrm{Detect}_{\mathcal{F}}(B)$ implemented in Section 2 evaluates the following predicates:

1. **Directional predicate** $\mathrm{Dir}(B)$: compute $\mathrm{sdir}(i)$ by Eq. 17 and return true iff there exist $i \neq j$ such that $\mathrm{sdir}(i) \cdot \mathrm{sdir}(j) = -1$.
2. **Interval predicate** $\mathrm{Int}(B)$: return true iff there exist $i \neq j$ such that $[\ell_i, u_i] \cap [\ell_j, u_j] = \emptyset$ (Eq. 19).
3. **Heterogeneity predicate** $\mathrm{Het}(B)$: compute D_B by Eq. 20 using the fixed weight rule $\{w_i\}$ and return true iff $D_B \geq \Delta_{\mathrm{het}}$ (Eq. 21).

Finally, the detector reports an $\mathcal{F}$-conflict iff

$$\mathrm{Detect}_{\mathcal{F}}(B) \;=\; \mathrm{Dir}(B) \;\vee\; \mathrm{Int}(B) \;\vee\; \mathrm{Het}(B). \tag{22}$$

Theorem 4 (Conflict Detection Completeness (w.r.t. $\mathcal{F}$))**.** *For any bucket B, if an $\mathcal{F}$-conflict exists in B (Definition 8), then $\mathrm{Detect}_{\mathcal{F}}(B)$ returns true.*

Proof. Assume B contains an $\mathcal{F}$-conflict. By Definition 8, at least one of the following holds.

Case 1: Directional conflict. By Definition 5, there exist $i \neq j$ such that $\mathrm{sdir}(i) \cdot \mathrm{sdir}(j) = -1$. The detector computes $\mathrm{sdir}(\cdot)$ for every claim and checks the existence of such a pair in $\mathrm{Dir}(B)$. Therefore $\mathrm{Dir}(B) = \mathsf{true}$, and hence $\mathrm{Detect}_{\mathcal{F}}(B) = \mathsf{true}$.

Case 2: Interval incompatibility. By Definition 6, there exist $i \neq j$ such that $[\ell_i, u_i] \cap [\ell_j, u_j] = \emptyset$. The detector explicitly checks this predicate over all pairs in $\mathrm{Int}(B)$. Therefore $\mathrm{Int}(B) = \mathsf{true}$, and hence $\mathrm{Detect}_{\mathcal{F}}(B) = \mathsf{true}$.

Case 3: High heterogeneity. By Definition 7, $D_B \geq \Delta_{\mathrm{het}}$, where D_B is the dispersion statistic defined in Eq. 20 under the fixed weight rule $\{w_i\}$. The detector computes the same D_B and compares it to the same threshold in $\mathrm{Het}(B)$. Therefore $\mathrm{Het}(B) = \mathsf{true}$, and hence $\mathrm{Detect}_{\mathcal{F}}(B) = \mathsf{true}$.

In all cases, if an $\mathcal{F}$-conflict exists, the detector returns true. This proves completeness with respect to $\mathcal{F}$. □

Remark. The theorem establishes a no-false-negative guarantee only for the explicitly defined, decidable taxonomy $\mathcal{F}$. It does not claim completeness for informal notions of disagreement outside Appendix B.2.

B.4 Quality Score Construction

This subsection defines the quality score $Q(\varepsilon, c, \pi)$ used for default evidence selection within a bucket. The design goals are: (i) *determinism* (fixed output for fixed inputs), (ii) *monotonicity* with respect to each quality dimension, and (iii) *auditability* (explicit, decomposable contributions).

Inputs. For each evidence object assigned to a bucket, we assume provenance metadata

$$\pi_i = (\mathrm{ref}_i,\ \mathrm{grade}_i,\ n_i,\ \mathrm{adj}_i,\ \mathrm{meta}_i),$$

where ref_i is a bibliographic identifier, grade_i is an ordinal evidence grade, n_i is the study sample size, adj_i encodes adjustment quality, and meta_i denotes any additional non-semantic metadata used only for tie-breaking (e.g., extraction provenance). The canonical claim is $c_{i,\mathrm{canon}} = (\theta_i, \mathrm{CI}_i)$ with $\mathrm{CI}_i = [\ell_i, u_i]$ on the canonical scale.

Four quality dimensions. We map each dimension to a normalized score in $[0, 1]$.

(1) Evidence grade. Let the grade set be totally ordered (e.g., $\mathsf{A} \succ \mathsf{B} \succ \mathsf{C}$). We fix a deterministic embedding $\phi_g : \{\mathsf{A}, \mathsf{B}, \mathsf{C}\} \to [0, 1]$ such that

$$\phi_g(\mathsf{A}) > \phi_g(\mathsf{B}) > \phi_g(\mathsf{C}). \tag{23}$$

In experiments we instantiate $\phi_g(\mathsf{A}) = 1$, $\phi_g(\mathsf{B}) = \frac{2}{3}$, $\phi_g(\mathsf{C}) = \frac{1}{3}$, but any fixed strictly order-preserving map satisfies the theory.

(2) Sample size. We use a monotone saturation transform $\phi_n : \mathbb{N} \to [0, 1]$:

$$\phi_n(n) = \min\Big\{1,\ \frac{\log(1+n)}{\log(1+n_{\max})}\Big\}, \tag{24}$$

where $n_{\max}$ is a fixed constant chosen a priori (e.g., the maximum sample size observed in the curated corpus). This mapping is monotone in n and compresses extreme values.

(3) Precision. Let the canonical interval width be $w_i = u_i - \ell_i$ (defined only when CI_i is available). We map width to a precision score via a monotone decreasing transform $\phi_p : \mathbb{R}_{>0} \to [0, 1]$:

$$\phi_p(w) = \min\Big\{1,\ \frac{w_{\mathrm{ref}}}{w}\Big\}, \tag{25}$$

where $w_{\mathrm{ref}} > 0$ is a fixed reference width (e.g., a corpus-level median width on the canonical scale). If CI_i is missing, we set $\phi_p(w_i) = 0$ and record this missingness in α during canonicalization.

(4) Adjustment quality. We assume adj_i belongs to a finite totally ordered set (e.g., $\mathsf{none} \prec \mathsf{basic} \prec \mathsf{rich}$), and define a deterministic order-preserving embedding ϕ_a:

$$\phi_a(\mathrm{adj}_1) \geq \phi_a(\mathrm{adj}_2) \quad \text{whenever} \quad \mathrm{adj}_1 \succeq \mathrm{adj}_2. \tag{26}$$

This dimension captures how strongly the study design/analysis addresses confounding; its categories are defined by the evidence schema.

Score aggregation. Let $G_i = \phi_g(\mathrm{grade}_i)$, $N_i = \phi_n(n_i)$, $P_i = \phi_p(u_i - \ell_i)$, and $A_i = \phi_a(\mathrm{adj}_i)$. We define the quality score as a fixed positive-weight aggregation:

$$Q_i \ = \ Q(\varepsilon_i, c_i, \pi_i) \ = \ w_g G_i + w_n N_i + w_p P_i + w_a A_i, \tag{27}$$

where $w_g, w_n, w_p, w_a > 0$ and $w_g + w_n + w_p + w_a = 1$ are fixed constants. Eq. 27 is monotone in each dimension by construction.

Deterministic tie-breaking. If multiple candidates attain the same maximal Q_i, we select by a fixed lexicographic rule on an audit tuple:

$$\mathrm{tie}(i) = \Big(G_i,\ A_i,\ P_i,\ N_i,\ \mathrm{ref}_i\Big), \tag{28}$$

choosing the candidate with the maximal tuple under lexicographic order. If ref_i is not unique, we extend Eq. 28 by an additional deterministic field from meta_i (e.g., extraction identifier). This ensures a unique default choice for fixed inputs.

Auditability. Eq. 27 decomposes Q_i into explicit contributions from the four dimensions. Together with the tie-breaking tuple (Eq. 28), the system can attribute the default selection to a small set of observable metadata fields.

Remark. The same provenance-derived weights $\{w_i\}$ used in the heterogeneity statistic (Eq. 20) can be chosen consistently with ϕ_n and/or ϕ_p (e.g., proportional to sample size or inverse interval width), but the completeness result in Appendix B.3 only requires that the weight rule is fixed and deterministic.

B.5 Derivation of Evidence Selection Consistency

This subsection proves Proposition 1. We first formalize the notions of determinism, monotonicity under weak dominance, and auditability, then show they follow from the construction of Q in Appendix B.4.

Setup. Fix a bucket B containing candidates indexed by $i \in \{1, \ldots, n\}$. Each candidate has canonical claim $c_{i,\text{canon}} = (\theta_i, [\ell_i, u_i])$ and provenance π_i, inducing four normalized quality components

$$G_i \in [0,1], \quad N_i \in [0,1], \quad P_i \in [0,1], \quad A_i \in [0,1],$$

as defined in Appendix B.4. The quality score is

$$Q_i = w_g G_i + w_n N_i + w_p P_i + w_a A_i,$$

with fixed weights $w_g, w_n, w_p, w_a > 0$ summing to 1, and deterministic tie-breaking by the audit tuple $\text{tie}(i)$ in Eq. 28.

Definition 9 (Default Selection)**.** *The default selection in bucket B is the unique index*

$$i^* \in \arg\max_{i \in \{1,\ldots,n\}} \big(Q_i, \text{tie}(i)\big) \tag{29}$$

under lexicographic order, and the default claim is $(\varepsilon^*, c^*) = (\varepsilon_{i^*}, c_{i^*})$.

Weak dominance. We formalize the statement "candidate a is no worse than b in all quality dimensions".

Definition 10 (Weak Dominance)**.** *For two candidates a and b in the same bucket, we say a weakly dominates b, written $a \succeq b$, if*

$$G_a \geq G_b, \quad A_a \geq A_b, \quad P_a \geq P_b, \quad N_a \geq N_b. \tag{30}$$

Lemma 1 (Determinism)**.** *For fixed bucket contents and fixed provenance fields, the default selection in Definition 9 is uniquely determined.*

Proof. Each Q_i is a deterministic function of $(\varepsilon_i, c_i, \pi_i)$ by construction of G_i, N_i, P_i, A_i and fixed weights. If multiple indices maximize Q_i, Eq. 29 selects the unique maximal audit tuple under a fixed lexicographic order. If needed, the final tie-breaking field is a deterministic unique identifier (from ref_i and/or meta_i), hence the maximizer is unique. □

Lemma 2 (Monotonicity under Weak Dominance)**.** *If $a \succeq b$ in the sense of Definition 10, then $Q_a \geq Q_b$.*

Proof. By Eq. 27 and positivity of weights,

$$Q_a - Q_b = w_g(G_a - G_b) + w_a(A_a - A_b) + w_p(P_a - P_b) + w_n(N_a - N_b) \geq 0,$$

since each parenthesized term is nonnegative under $a \succeq b$. □

Lemma 3 (Auditability)**.** *The default choice admits an audit trail decomposable into the four quality dimensions and deterministic tie-breaking fields.*

Proof. Eq. 27 expresses Q_i as an explicit sum of four terms with fixed weights, and Eq. 28 provides a fixed ordered list of secondary criteria. Therefore, for the chosen i^* the system can report $(G_{i^*}, A_{i^*}, P_{i^*}, N_{i^*})$ and the tie-breaking comparisons that separate i^* from other candidates, yielding a finite, deterministic explanation trace. □

Theorem 5 (Evidence Selection Consistency)**.** *The default selection rule in Definition 9 is deterministic (Lemma 1), monotone under weak dominance (Lemma 2), and auditable (Lemma 3).*

Proof. Immediate by Lemma 1, Lemma 2, and Lemma 3. □

Remark. The result does not claim global optimality under an unrestricted utility class. Instead, it establishes consistency with respect to component-wise dominance and a deterministic, auditable aggregation of explicitly defined quality dimensions.

B.6 Proof of Canonicalization Guarantees

This subsection proves Theorem 1. We formalize (i) determinism/uniqueness, (ii) semantic preservation, and (iii) information preservation via an explicit reconstruction map.

Domain and notation. Let $\mathcal{D}$ denote the space of input pairs (ε, c) where $\varepsilon = (\Pi, \iota, o, \tau, \mu)$ and $c = (\theta, \mathrm{CI})$. Let $\widehat{\mathcal{D}}$ denote the space of outputs $(\varepsilon_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha)$.

We write $\varepsilon_{\mathrm{canon}} = (\Pi, \iota, o, \widehat{\tau}, \mu_{\mathrm{canon}})$, where $\widehat{\tau} = \mathrm{align}(\tau)$ is the aligned horizon representation (Definition B.1.4). Let $\mathcal{T}_\mu$ denote the canonical reparameterization operator for the measure component:

$$\mathcal{T}_\mu : (\mu, c) \mapsto (\mu_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha_\mu),$$

where α_μ records the transformation context needed for reconstruction (Appendix B.1.5). The full canonicalization is

$$N(\varepsilon, c) = \Big((\Pi, \iota, o, \mathrm{align}(\tau), \mu_{\mathrm{canon}}),\ c_{\mathrm{canon}},\ \alpha \Big), \tag{31}$$

with $\alpha = (\alpha_\mu, \alpha_\tau, \alpha_{\mathrm{meta}})$, where α_τ records horizon-alignment context and α_{meta} records any additional deterministic metadata bindings used by canonicalization.

Equivalence up to representation. Information preservation is stated "up to equivalence" because multiple raw representations can encode the same semantic claim. Define a relation $\approx$ on $\mathcal{D}$ by

$$(\varepsilon, c) \approx (\varepsilon', c') \iff (\Pi, \iota, o, \tau) = (\Pi', \iota', o', \tau') \wedge (\mu, c) \text{ and } (\mu', c') \text{ agree up to within-type reparameterization.} \tag{32}$$

The second conjunct means that there exists a within-type bijection g (Appendix B.1.3) such that $\theta' = g(\theta)$ and $\mathrm{CI}' = g(\mathrm{CI})$, and μ' denotes the same measure type as μ but possibly expressed under an equivalent parameterization.

Rule system with total priority. Canonicalization is defined by a finite set of rules $\mathcal{R}$ operating only on (μ, c), each rule of the form

$$r :\ (\mu, c) \mapsto (\mu', c', \alpha_r),$$

together with a total priority order $\prec$ over $\mathcal{R}$. For any input (μ, c), let $\mathcal{R}(\mu, c) \subseteq \mathcal{R}$ be the subset of applicable rules. The chosen rule is

$$r^*(\mu, c) = \min_{\prec}\ \mathcal{R}(\mu, c), \tag{33}$$

and $\mathcal{T}_\mu$ applies r^* once to produce $(\mu_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha_\mu)$. Any missingness or binding decisions used for applicability (e.g., whether a reported coefficient is already on the canonical scale) must be supported by explicit metadata fields and recorded in α_{meta}.

B.6.1 Uniqueness (Determinism and Stability)

Lemma 4 (Determinism of $\mathcal{T}_\mu$)**.** *For any input (μ, c), the output $(\mu_{\mathrm{canon}}, c_{\mathrm{canon}}, \alpha_\mu) = \mathcal{T}_\mu(\mu, c)$ is uniquely determined.*

Proof. The set of applicable rules $\mathcal{R}(\mu, c)$ is determined by (μ, c) and the fixed rule predicates. Since $\prec$ is a total order, Eq. 33 selects a unique $r^*(\mu, c)$ whenever $\mathcal{R}(\mu, c) \neq \emptyset$. Applying r^* yields a unique output by functional definition of rules. If $\mathcal{R}(\mu, c) = \emptyset$, $\mathcal{T}_\mu$ returns the identity output with a fixed tag in α_μ. □

Lemma 5 (Stability on canonical forms)**.** *Let $\mathcal{D}_{\text{canon}} \subseteq \mathcal{D}$ denote the subset of inputs whose (μ, c) are already in canonical form and whose horizon representation is aligned, i.e., $\tau = \text{align}(\tau)$ and $(\mu, c) = (\mu_{\text{canon}}, c_{\text{canon}})$. Then for any $(\varepsilon, c) \in \mathcal{D}_{\text{canon}}$,*

$$N(\varepsilon, c) = \big(\varepsilon, c, \alpha_{\text{id}}\big),$$

where α_{id} is a fixed identity-context record.

Proof. For canonical (μ, c), either (i) no nontrivial rule is applicable, so $\mathcal{T}_\mu$ returns the identity output by definition, or (ii) an explicit identity rule is applicable and dominates by priority, yielding the same result. Since τ is already aligned, $\text{align}(\tau) = \tau$ and α_τ takes a fixed "already-aligned" value. Thus Eq. 31 returns the input unchanged up to recording α_{id}. □

Lemma 6 (Uniqueness of the normal form)**.** *For any $(\varepsilon, c) \in \mathcal{D}$, the canonical form $(\varepsilon_{\text{canon}}, c_{\text{canon}})$ produced by N is unique.*

Proof. The components (Π, ι, o) are copied directly. The aligned horizon $\widehat{\tau} = \text{align}(\tau)$ is uniquely determined by the alignment operator (Proposition 2, Appendix B.1.4). By Lemma 4, $(\mu_{\text{canon}}, c_{\text{canon}})$ is uniquely determined. Therefore $(\varepsilon_{\text{canon}}, c_{\text{canon}})$ is unique. □

B.6.2 Semantic Preservation

Lemma 7 (Semantic core is invariant)**.** *For any input (ε, c) with $\varepsilon = (\Pi, \iota, o, \tau, \mu)$, the output $N(\varepsilon, c)$ satisfies*

$$(\Pi_{\text{out}}, \iota_{\text{out}}, o_{\text{out}}) = (\Pi, \iota, o), \qquad \widehat{\tau}_{\text{out}} = \text{align}(\tau),$$

and N does not modify (Π, ι, o).

Proof. By construction (Eq. 31), N applies align only to τ and applies $\mathcal{T}_\mu$ only to (μ, c). Neither operator touches (Π, ι, o). Hence the semantic core is preserved and only the horizon is aligned. □

B.6.3 Information Preservation and Reconstruction

Context record α. The context α is defined to be sufficient for reconstruction. Concretely, α includes: (i) the selected rule identifier r^* (and its parameters, if any), (ii) the original measure type and parameterization tag, (iii) any explicit metadata bindings used to interpret the raw report, and (iv) the horizon-alignment class and binning outcome. All fields are deterministic functions of the input and are stored as part of $(\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha)$.

Lemma 8 (Existence of a reconstruction map)**.** *There exists a map $N^{-1} : \widehat{\mathcal{D}} \to \mathcal{D}$ such that for all $(\varepsilon, c) \in \mathcal{D}$,*

$$N^{-1}\big(N(\varepsilon, c)\big) \;\approx\; (\varepsilon, c). \tag{34}$$

Proof. Given an output triple $(\varepsilon_{\text{canon}}, c_{\text{canon}}, \alpha)$, define N^{-1} as follows.

Step 1 (invert the measure transform). From α_μ, recover the selected rule identifier r^* and the original measure type/parameterization tag. Each rule $r \in \mathcal{R}$ is either an identity step or a within-type bijective reparameterization whose inverse is known (e.g., log/exp for ratio-type canonicalization; Appendix B.1.3). Apply the corresponding inverse transform to $(\mu_{\text{canon}}, c_{\text{canon}})$ to obtain some (μ', c') in the original representation class recorded by α_μ.

Step 2 (recover the horizon representation). From α_τ, choose a representative τ' consistent with the aligned output $\widehat{\tau}$ and the recorded alignment class. This step may not be unique; any deterministic choice yields an equivalent representation under $\approx$ because $\widehat{\tau}$ is the aligned semantic horizon used for comparability.

Step 3 (assemble). Return (ε', c') with $\varepsilon' = (\Pi, \iota, o, \tau', \mu')$, using (Π, ι, o) copied from $\varepsilon_{\text{canon}}$.

By construction, (Π, ι, o) matches the original input, and (μ', c') is in the same within-type reparameterization class as the original (μ, c), because α_μ pins down the representation class and the inverse reparameterization. Therefore $(\varepsilon', c') \approx (\varepsilon, c)$, establishing Eq. 34. □

B.6.4 Proof of Theorem 1

Proof of Theorem 1. Uniqueness follows from Lemma 6 and stability on canonical forms from Lemma 5. Semantic preservation is Lemma 7. Information preservation follows from Lemma 8. □

C Six Queries: Executability, Status Flags, and Examples

C.1 Flags and the Unified Answer Object

The unified answer object is

$$\mathsf{Ans} = \big(\varepsilon, \hat{\theta}, \mathrm{CI}, \pi, \mathcal{F}(\mathcal{B}), \mathrm{flags}\big), \qquad \mathrm{flags} \subseteq \mathcal{F}_{\mathrm{flag}}.$$

Group	Flags
Executability	`executable`
Missingness	`missing_edge`, `missing_path`, `missing_field`
Incompatibility	`mixed_mtype`
Disagreement	`heterogeneity`, `conflict`
Premise	`assumption_required`
Coverage	`no_subgroup_evidence`, `insufficient_time_coverage`

C.2 Q_{do}: Interventional Effect

Executability. There exists a key s such that $\mathcal{B}_s \neq \varnothing$ and s matches the queried (X, Y, ε) semantic dimensions (see the definition of `matches` in Table 2).

Return. When executable, let $(\varepsilon_s^*, c_s^*, \pi_s^*)$ be the default kernel for s and let $\mathcal{B}_s$ be the associated bucket. The returned object is

$$\mathsf{Ans} = (\varepsilon_s^*, \hat{\theta}_s^*, \mathrm{CI}_s^*, \pi_s^*, \mathcal{F}(\mathcal{B}_s), \mathrm{flags}).$$

C.3 Q_{med}: Mediation Decomposition

Executability. There exist keys s_1, s_2 such that $\mathcal{B}_{s_1} \neq \varnothing$ and s_1 matches (X, M, ε), and $\mathcal{B}_{s_2} \neq \varnothing$ and s_2 matches (M, Y, ε).

Return. When executable, the query returns the triple

$$(\mathsf{Ans}_{\mathrm{TE}}, \mathsf{Ans}_{\mathrm{NDE}}, \mathsf{Ans}_{\mathrm{NIE}}),$$

where each component is an answer object of the form Ans defined on the corresponding required evidence pattern, with status flags recording executability conditions and certified conflicts.

C.4 Q_{joint}: Joint Intervention

Executability. There exist keys s_1, s_2 such that $\mathcal{B}_{s_1} \neq \varnothing$ and s_1 matches (X_1, Y, ε), and $\mathcal{B}_{s_2} \neq \varnothing$ and s_2 matches (X_2, Y, ε).

Return. When executable, the query returns an answer object Ans associated with the required joint evidence pattern.

C.5 Q_{cf}: Counterfactual / Individual-Context Query

Executability. There exists a key s such that $\mathcal{B}_s \neq \varnothing$ and s matches (X, Y, ε), together with a well-formed individual context identifier z for counterfactual querying.

Return. When executable, the query returns an answer object Ans whose estimand semantics are conditioned on the individual context z.

C.6 Q_{CATE}: Subgroup Effect

Executability. There exists a key s such that $\mathcal{B}_s \neq \varnothing$, s matches (X, Y, ε), and the population category encoded in s equals the subgroup identifier z (i.e., $\mathrm{pbucket}(\Pi) = z$).

Return. When executable, the query returns an answer object Ans corresponding to the subgroup constraint z.

C.7 Q_{traj}: Time Trajectory

Executability. Given a set of aligned time classes $\mathcal{T}$, for every $\hat{\tau} \in \mathcal{T}$ there exists a key $s(\hat{\tau})$ such that $\mathcal{B}_{s(\hat{\tau})} \neq \varnothing$ and $s(\hat{\tau})$ matches $(X, Y, \varepsilon, \hat{\tau})$.

Return. When executable, the query returns the sequence $\{\mathsf{Ans}(\hat{\tau})\}_{\hat{\tau} \in \mathcal{T}}$.

C.8 Examples

Example 1: Missing mediation path. Given (X, M, Y), if a non-empty bucket matches (X, M) but no non-empty bucket matches (M, Y), then the executability condition of Q_{med} fails and flags includes `missing_path`.

Example 2: Conflict annotation in trajectories. For two aligned time classes $\hat{\tau}_1, \hat{\tau}_2$ of the same (X, Y, ε), if the default kernels in the corresponding buckets have opposite directions on the canonical scale and are both significant, then $\mathcal{F}(\mathcal{B}_{s(\hat{\tau}_1)})$ and $\mathcal{F}(\mathcal{B}_{s(\hat{\tau}_2)})$ include `dir`; the corresponding answers record `conflict` and return the triggering witnesses for auditing.

D Extended Related Work

This appendix provides an extended discussion of prior work along the technical threads most closely related to DoAtlas: (i) medical AI and medical foundation models for clinical assistance, (ii) literature-grounded generation and evidence-centric NLP pipelines, (iii) evidence-based medicine and systematic evidence synthesis, (iv) causal inference and machine learning for clinical decision-making, and (v) multi-source evidence fusion, conflict governance, external validity, and auditability.

D.1 Medical AI for clinical assistance: a historical perspective

Rule-based expert systems (1970s–1990s). Early medical artificial intelligence systems were dominated by rule-based expert systems, which encoded clinical heuristics through manually curated rules and decision trees. Representative systems such as MYCIN demonstrated how expert knowledge could be formalized for tasks including infectious disease diagnosis and antibiotic selection, with explicit rule traces and uncertainty handling mechanisms [36]. Similarly, large-scale diagnostic systems such as INTERNIST-1 and its successor QMR focused on hypothesis generation and differential diagnosis using structured disease finding associations [30]. Later systems such as DXplain further operationalized paradigms for clinical use by mapping patient findings to ranked diagnostic hypotheses with explanatory support [6]. While transparent and interpretable, from a modern perspective, these systems were brittle, difficult to maintain, and exhibited limited generalization beyond predefined clinical scenarios, particularly in complex clinical environments.

Statistical learning and predictive modeling (2000s–early 2010s). With the digitization of healthcare data in the early 2000s, medical AI increasingly shifted toward statistical and machine learning models trained on structured clinical variables. Regression-based clinical risk scores became widely adopted for diagnosis, risk stratification, and prognosis, exemplified by the Framingham risk functions for cardiovascular disease [9]. In parallel, early electronic health record (EHR)–driven predictive models emerged to estimate disease risk and clinical outcomes at scale, with evaluation focused primarily on discrimination and calibration rather than explicit decision semantics or intervention definitions [33, 35]. This generation of models supported population-level risk assessment but typically exposed outputs as probabilities or scores rather than executable representations of clinical actions.

Deep learning for high-dimensional clinical data (mid-2010s). Around the mid-2010s, deep learning substantially expanded the scope of medical AI by enabling end-to-end modeling of high-dimensional clinical inputs, including medical images, physiological waveforms, and free-text clinical notes. Systems achieved near-expert or expert-level performance on multiple diagnostic tasks, particularly in medical imaging, as demonstrated by large-scale convolutional models for skin cancer classification [10]. At the same time, deep neural architectures were increasingly applied to longitudinal EHR data for outcome prediction and patient risk modeling [35]. Despite these advances, most deep learning systems continued to operate within a predictive paradigm, leaving clinical decision-making and intervention interpretation to downstream human judgment, without explicit causal or decision-theoretic representations.

Foundation models and natural-language clinical interfaces (2020s–present). Since approximately 2020, foundation models and large language models have been explored as general-purpose interfaces for clinical assistance. General-purpose LLMs demonstrated strong performance on medical question answering and professional examination benchmarks, highlighting the feasibility of natural-language clinical interaction at scale [31]. Domain-adapted medical LLMs, such as Med-PaLM, further advanced performance through targeted training and evaluation on medical reasoning tasks [38]. In parallel, industry-developed medical LLMs including HuatuoGPT, DoctorGLM, and Baichuan-M1 emphasized medical dialogue, domain knowledge understanding, and clinician-facing assistance [46, 43, 41]. Across this generation, retrieval-augmented generation techniques have been widely adopted to ground responses in biomedical literature and clinical guidelines [27]. While these systems prioritize interaction and synthesis, clinical decision support is still predominantly operationalized through narrative outputs rather than explicit, recomputable representations of interventional effects. Beyond the historical evolution of medical AI models, several parallel research threads have sought to improve reliability, evidence usage, and decision support. We briefly summarize these directions below, focusing on their relationship to executable causal decision-making.

D.2 Literature-grounded generation and evidence-centric NLP

Recent work has explored grounding language model outputs in external knowledge sources to improve factual reliability and reduce hallucination. Retrieval-augmented generation (RAG) combines parametric language models with non-parametric document retrieval, enabling responses to be conditioned on retrieved corpora rather than relying solely on internal parameters [27]. In clinical settings, RAG-style systems have been instantiated for guideline- and literature-backed assistance (e.g., retrieving medical references for treatment recommendations) [45] and systematically benchmarked in medical QA, where toolkits such as MedRAG highlight both accuracy gains and failure modes (e.g., retrieval sensitivity and "lost-in-the-middle") [43].

Beyond retrieval, a growing line of work aims to make grounding explicit through citation-aware generation and fine-grained attribution. "Attributed QA" formalizes the requirement that generated claims be supported by specific evidence spans [7], while datasets and evaluation protocols (e.g., expert-judged claim–evidence pairs) operationalize supportiveness and citeworthiness as measurable targets [28]. Complementary methods improve citation behavior through prompt- and reasoning-based control (e.g., generating with citations via structured reasoning) [23], or by designing end-to-end medical citation generation and evaluation pipelines [42]. Post-hoc attribution and verification further retrofit grounding onto black-box generators by searching for supporting sources and revising unsupported statements [12], and recent work has proposed automated pipelines to assess whether cited sources truly support model claims in health-related queries.

While these approaches improve transparency and verifiability at the text level—often reducing hallucinations or enabling claim-level evidence checking [29]—they typically treat evidence as narrative justification rather than as structured, executable objects. Retrieved studies are surfaced as references or rationales, but their estimand definitions (contrast, population, time horizon), effect scales/units, and uncertainty semantics are rarely normalized across sources, limiting downstream use for accountable decision-making and interventional planning.

D.3 Evidence-based medicine and systematic evidence synthesis

Evidence-based medicine (EBM) provides a principled framework for evaluating clinical knowledge through hierarchies of evidence, randomized controlled trials, and systematic reviews [11, 34].

Meta-analysis and clinical guideline development aim to aggregate results across studies to support evidence-based recommendations, with increasing emphasis on transparency, reproducibility, and methodological rigor [19]. However, traditional evidence synthesis pipelines remain largely manual and slow to update, and typically focus on population-average conclusions rather than individualized or context-specific effects [14]. Effect heterogeneity, conflicting findings, and limitations in external validity across populations are often summarized qualitatively, rather than represented in machine-readable or executable forms, posing challenges for integration with computational clinical decision support systems, particularly for real-time or adaptive decision-making [25, 22].

D.4 Causal inference and causal machine learning for clinical decision-making

Causal inference provides formal tools for estimating interventional effects, including structural causal models, potential outcomes, propensity-based adjustment, instrumental variables, mediation analysis, and heterogeneous treatment effect estimation [32, 17, 40]. These frameworks support principled answers to counterfactual and interventional queries central to clinical decision-making, and have been widely applied in epidemiology and health services research—for example, using propensity score methods to estimate treatment effects from observational EHR data [4], instrumental variables to address unmeasured confounding in pharmacoepidemiology [15], and mediation analysis to decompose mechanisms of treatment effects [40]. More recently, causal machine learning methods have improved scalability and flexibility, enabling estimation of heterogeneous treatment effects and non-linear causal relationships in high-dimensional settings [3, 8]. Despite these advances, causal analyses in medicine typically operate within individual datasets or single-study contexts, with estimands, populations, and effect scales defined ad hoc. As a result, causal estimates are rarely standardized across studies or compiled into reusable representations, limiting their integration with broader evidence ecosystems and decision-support infrastructures [17]. This limitation becomes critical in large-scale evidence clinical deployment.

D.5 Multi-source evidence fusion, conflict governance, and auditability

Combining evidence across heterogeneous sources raises challenges in effect standardization, conflict resolution, and uncertainty propagation [22, 21]. Prior work has explored ensemble learning, Bayesian model averaging, and sensitivity analysis to address model uncertainty and robustness, but systematic mechanisms for governing evidential conflicts across studies remain limited [20, 16]. In clinical settings, auditability and external validity are critical for safe deployment [39, 2]. Existing medical AI systems rarely provide explicit provenance, executable effect definitions, or validation signals against real-world data, making it difficult to assess reliability, adjudicate conflicting claims, or prevent unsafe recommendations at scale [24].

This part details a systematic methodological framework for constructing a structured, evidence-based medical dataset and establishing computable mappings from abstract literature variables to specific fields within the local Human Phenotype Project (HPP) database, ultimately achieving automated empirical verification of medical mechanisms and pathways. The dataset construction encompasses a closed-loop pipeline involving literature metadata cleaning, structured extraction of causal logic, semantic alignment with local data dictionaries, and empirical testing of statistical models. The primary objective is to verify that medical pathways identified in the literature exhibit consistent directionality, non-zero effects, and plausible magnitudes within the local HPP dataset.

D.6 Construction of Structured Evidence Dataset

D.6.1 Literature discovery and classification system and corresponding overview

This study established a systematic pipeline for literature discovery and hierarchical processing. This pipeline aims to extract high-quality research with clear methodological characteristics from massive amounts of unstructured biomedical text and map them to structured data templates using a rigorous classification framework.

To ensure the scientific rigor and timeliness of the source data, a strict retrieval strategy and hard guardrails were implemented during the construction phase of this dataset. In terms of time dimension, research published between January 1, 2020 and December 31, 2025 were included to ensure that the extracted evidence reflects the latest developments in the relevant field. For model utilization, Gemini

3 Pro and Grok 4.1 Thinking were employed for preliminary literature screening and extraction, Claude Sonnet 4.5 Thinking was utilized to read papers and extract structured evidence, and to map literature nodes to the local HPP dataset. Regarding publication source quality, candidate literature was restricted to journals with an Impact Factor ≥ 10, serving as a proxy for rigorous study design and result credibility. To mitigate the risk of hallucinated citations associated with generative models, we implemented a strict metadata validation mechanism. All candidate papers were required to possess a resolvable DOI with authoritative metadata verified via Crossref. Furthermore, valid landing pages for these DOIs must be identifiable on PubMed, PubMed Central (PMC), or domains within a trusted publisher whitelist. Entries lacking a DOI, failing to yield consistent Crossref metadata, or lacking association with a whitelisted domain were automatically excluded. Complementing this automated validation, a rigorous manual review process was executed, in which all final included data underwent human verification to ensure completeness, authenticity, and accuracy.

Based on distinct research foci, this dataset establishes a mutually exclusive and exhaustive taxonomy comprising four evidence categories: Interventional, Causal, Mechanistic, and Associational.

Interventional The primary objective of interventional studies is to confirm the effectiveness of interventions, evaluating the net causal effect of a specific exposure or treatment on clinical outcomes under strictly controlled experimental conditions. These studies constitute the primary benchmark within the overall framework. The classification criteria for this category are exclusive, relying primarily on PubMed publication types explicitly tagged as "Randomized Controlled Trial" or "Clinical Trial," as well as methodological keywords such as "randomized," "double-blind," "placebo," or the presence of clinical trial registration numbers (e.g., NCT identifiers). In the localization validation of the HPP dataset, interventional studies require the reproduction of statistical significance, with a specific emphasis on verifying the directional consistency of the Average Treatment Effect (ATE) between the intervention and control groups.

Causal Causal studies utilize specific statistical methods to control for confounding within observational data to estimate the causal effect between exposure and outcome. These studies mitigate the limitations of randomized controlled trials due to sample scarcity and high costs. The criteria for systematically identifying such studies are whether they use explicit causal identification strategies, including Mendelian randomization, instrumental variables, difference-in-differences, propensity score matching/IPTW, and target trial emulation. This classification distinguishes studies that employ causal identification strategies from those that only perform conventional regression adjustments.

Mechanistic Mechanism studies focus on analyzing how exposure factors influence outcomes through intermediate variables, delving into the internal transmission pathways and biological mechanisms of the study. Within the overall framework, the objective of this category is to validate the existence of biological pathways. The core criteria for evaluating such studies are the presence of explicit mediation analysis or indirect effect calculations, and whether a complete $X \rightarrow M \rightarrow Y$ chain is constructed in the research hypotheses. Key metrics include the statistical significance of path coefficients and the proportion of the effect explained by the mediator. In HPP data validation, the emphasis is on verifying the existence and validity of specific biological pathways in the HPP population.

Associational Associational studies characterize statistical correlations between variables, typically used to capture phenomenological covariation to generate preliminary scientific hypotheses. This category is defined via exclusion, encompassing cross-sectional and cohort studies that report adjusted Hazard Ratios (HR), Odds Ratios (OR), or regression coefficients, but do not employ randomization or the aforementioned specific causal inference methods. Distinct from the validation of causal effects, the validation process for this category primarily assesses the direction and magnitude of observed correlations.

D.7 HPP Field Mapping

Building upon the structured evidence dataset constructed in the preceding subsection, we establish a bridge between unstructured natural language descriptions from the literature and the specific local Human Phenotype Project (HPP) cohort data. This integration is achieved through deep alignment across semantic consistency, measurement methodology, and unit standardization.

D.7.1 Mapping rule

This study establishes a precise mapping rule grounded in a metadata dictionary. The core objective of this rule is to anchor abstract variable nodes extracted from the literature, including exposures (X), outcomes (Y), mediators (M), and instrumental variables (IV) to specific physical fields within the HPP database. The entire mapping system comprises a two-layer architecture featuring unique identifier locking and confidence grading. This framework first establishes variable node directionality via a strict data dictionary benchmark, and subsequently quantifies the match quality of each mapping relationship through a standardized status classification system.

The sole Ground Truth for this mapping process is restricted to the official Data Dictionary of the HPP platform Pheno AI Knowledgebase, which serves as the definitive metadata source for all subdatasets. Specifically, mappings must point strictly to the `tabular_field_name` within the database, which functions as the unique, persistent identifier bridging literature concepts and local data. The extraction protocol mandates that the mapping results must be spelled exactly as in the data dictionary, including the preservation of case sensitivity and underscore formatting. The use of non-standardized generic names or fabricated created field names is strictly prohibited to ensure that subsequent automated validation scripts can accurately call the corresponding column data. Acknowledging the inevitable heterogeneity between variable definitions in medical literature and the actual measurement methods of the local HPP cohort, this study introduces state classification to quantify the matching accuracy and confidence of each mapping relationship. Based on the degree of consistency in the definition, the system divides the mapping results into the following four mutually exclusive categories: Exact, Close, Derived, and Tentative. The criteria and operating procedures for determining each mapping state are as follows:

Exact Indicates that the variable from the literature is identical to the HPP field in definition, unit, and measurement methodology. For example, "Systolic BP (mmHg)" in the literature maps directly to `sbp_mmHg` in HPP. This category offers the highest validation validity.

Close Indicates high conceptual consistency between the literature variable and the HPP field, despite non-essential differences in specific device models, questionnaire versions, or unit magnitudes. These differences are judged not to fundamentally alter the biological significance of the variable. For instance, "AHI measured by PSG" in the literature versus "pAHI measured by portable devices" in HPP. In such cases, discrepancies must be explicitly documented in the `notes` field to facilitate bias assessment during validation.

Derived Indicates that the literature variable does not correspond to a single raw field in HPP but can be obtained via mathematical calculation, logical aggregation, or thresholding of existing fields. Common scenarios include unit conversion (e.g., mg/dL to mmol/L), multi-field synthesis (e.g., calculating BMI from height and weight), or discretization of continuous variables (e.g., generating a categorical variable by applying a "<6 hours" threshold to a continuous sleep duration field). Such mappings require the explicit recording of the calculation formula or derivation logic in the `notes` field.

Tentative Indicates that no clear corresponding field exists in the HPP data dictionary, or that ambiguous candidates exist without sufficient information to confirm consistency. These mappings are retained solely as potential leads for validation and are typically excluded from the automated validation of core mechanisms. The source of uncertainty and potential candidate fields must be noted in the `notes` field.

D.7.2 Variable conversion and alignment

Following basic field anchoring, we implemented a standardized variable transformation and alignment protocol for variables classified as Close or Derived to eliminate systematic bias arising from discrepancies in units, granularity, or definition thresholds. For cases where the units reported in the literature are inconsistent with the native units of the local HPP database, the system employs a conversion mechanism based on the UCUM (Unified Code for Units of Measure) standard, mandating the recording of precise conversion formulas in the remarks field (e.g., the literature uses the conventional unit mg/dL to report blood glucose, while HPP uses the international unit mmol/L), ensuring that the values are in the same dimensional space before entering the statistical validation model. For composite variables not representable by a single physical field, the protocol defines multi-field aggregation rules. Multiple source fields are linked via pipe separators, with synthesis

algorithms—such as summation, averaging, or logical derivation—explicitly recorded to construct physiological metrics consistent with the literature (e.g., calculating Total Energy Intake by summing specific nutrients, or BMI from height and weight). For complex derived variables, such as "energy balance" inferred from daily weight changes and Dual-energy X-ray Absorptiometry (DXA) data, the system requires derivation formulas and intermediate variables are required to ensure logical alignment with the source text. Given that medical literature often uses specific cutoff values to convert continuous indicators into categorical variables, the alignment protocol mandates the explicit definition of classification thresholds and reference groups. These are marked as Derived states, with discretization rules specified to facilitate subsequent data binning. During validation, the system dynamically applies these rules to bin or binarize raw HPP data, reconstructing exposure and control groups homogeneous to those in the literature to ensure the comparability of categorical effect sizes. Finally, the system enforces strict temporal anchoring, aligning measurement time points from the literature with specific local data collection waves (e.g., mapping "Visit 1" to the corresponding HPP wave). Where literature employs means over specific time windows (e.g., "average sleep duration over the past 7 nights"), the protocol requires aggregating repeated measures from the corresponding local timeframe. This mechanism effectively prevents interference from time-varying confounding factors, ensuring that the temporal logic of exposure and outcome conforms to the basic assumptions of causal inference.

D.8 Structural equation construction

D.8.1 Verification environment definition

To enable the automated validation and causal inference of medical mechanism pathways on the local HPP dataset, we established a standardized validation environment and variable domain. Given the complexity of causal inference and the sparsity of longitudinal data, the current validation framework adopts a "Single Time-point Slice" model. This means that all variables exposure X, outcome Y, mediator M, and covariates Z, are anchored within a specific temporal window surrounding the participant's enrollment baseline (default: baseline $\pm$ 90 days). This setting treats each pathway to be validated, including the involved exposure X, outcome Y, potential mediator M, and adjustment set Z, as static observations within the same time slice, focusing on validating the causal structure of the mechanism pathway in the cross-sectional or short-term context. The construction of the validation cohort strictly follows the inclusion and exclusion criteria defined in the `population.eligibility_signature` field of the extracted record. The automated system dynamically queries the HPP database to identify sub-cohorts matching specified age ranges, gender distributions, and disease histories, subsequently applying standardized preprocessing pipelines to key variables. For missing data, the system employs imputation strategies based on multiple imputation or full case analysis; for outliers in continuous variables, it uses interquartile range-based truncation. These definitions transform heterogeneous literature evidence into homogeneous, data-complete, static statistical tasks on the local dataset.

D.8.2 Unified Structural Equations (EL-GSE) Specification

To achieve a standardized transformation from unstructured documentary evidence to computable causal objects, this study defines and establishes the Edge-Local Generalized Structural Equation (EL-GSE) specification. EL-GSE is a general mathematical abstraction template designed to encapsulate each causal path extracted from the literature into an atomic unit with independent computational capabilities. For any candidate causal edge $X \to Y$, its EL-GSE formal definition is as follows:

$$g(\mathbb{E}[Y \mid X, Z]) = \alpha + \beta \cdot T_X(X) + \gamma^\top T_Z(Z)$$

In this equation, $g(\cdot)$ represents the link function, the form of which is strictly determined by the data distribution type of the outcome variable Y; $T_X(\cdot)$ is a contrastive coding function for the exposure variable X (including standardization, quantile contrast, or treatment/control coding) used to mathematically align the effect definitions reported in the literature; $T_Z(\cdot)$ is the vectorized coded representation of the adjustment set Z; β is the core causal parameter to be validated, corresponding to the effect size reported in the source study; γ is the regression coefficient vector of the adjustment set; and α is the intercept term.

Based on the distribution characteristics of the outcome variables, the EL-GSE specification further derives two standard sub-forms to adapt to actual data: For continuous outcomes, the equation adopts

the identity link function and the Gaussian noise assumption, degenerating into a linear structure equation $Y := \alpha + \beta T_X(X) + \gamma^\top T_Z(Z) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For binary outcomes, the equation adopts the Logit link function and the Bernoulli distribution assumption, transforming into a logistic structure equation $Y := \text{Bernoulli}(\pi)$, where $\text{logit}(\pi) = \alpha + \beta T_X(X) + \gamma^\top T_Z(Z)$.

The core advantage of the EL-GSE specification lies not only in its support for statistical reproducibility at the parameter level (such as verifying the direction and significance of β), but more importantly, in its inherent support for causal computation (Do-calculus). Under the premise of a single time point and identifiable backdoor path, this equation can be directly used as the execution carrier of the `do()` operator, utilizing the G-computation formula $\mathbb{E}[Y \mid do(X = x)] = \mathbb{E}_Z[g^{-1}(\alpha + \beta T_X(x) + \gamma^\top T_Z(Z))]$ to calculate the average intervention effect at the population level, or to calculate personalized counterfactual prognoses based on the covariate $Z = z$ for a specific individual. Finally, these equations and their parameters are serialized into a standard JSON object containing `form`, `parameters`, and `do_support`, completing the transition from static literature knowledge to a dynamic, executable causal model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the scope and contributions of the paper. The paper proposes causal compilation, instantiates it as DoAtlas-1, and supports the claimed executability, validation, and auditability through the formalization and experiments in Sections 4–6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses important scope and executability limitations through explicit query conditions, diagnostic flags for non-executable queries, conflict annotations, and validation constraints. These limitations clarify when DoAtlas can return a numeric causal answer and when it should instead surface missing-evidence or conflict status.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states theoretical results for canonicalization, conflict detection completeness, and evidence selection consistency in Section 4.2. The assumptions, constructions, and proofs are provided or referenced in the appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the core information needed to reproduce the main experimental results, including the estimand representation, canonicalization procedure, conflict taxonomy, query semantics, validation protocol, and ablation settings. The deterministic design and full provenance tracking further support reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not primarily contribute an open-access benchmark or open-source training pipeline requiring public release of data and code. The validation relies on existing clinical evidence and HPP-derived resources, whose access is governed by the corresponding data-use policies.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental setting, including the number of compiled kernels, source studies, buckets, query-suite sizes, negative-sample ratio, clinical-review protocol, ablation configurations, and external validation comparisons. Since DoAtlas is a deterministic causal compilation system rather than a trained neural model, optimizer and training hyperparameters are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports causal effect estimates together with confidence intervals for executable queries, and uses interval-based and heterogeneity-based criteria in conflict detection and validation. The main experimental conclusions are further supported by quantitative validation metrics, including canonicalization accuracy, conflict-detection performance, query executability, and external-validation agreement.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The reported experiments consist of deterministic evidence compilation, canonicalization, conflict detection, graph construction, query execution, and external validation rather than large-scale neural-network training. The paper reports the corpus size, graph size, query-suite size, and validation setting, which characterize the computational scale; these procedures can be reproduced with standard CPU-based compute and do not require specialized GPU training resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics by focusing on auditable, verifiable, and provenance-tracked clinical decision-support infrastructure. The system is presented as a research tool for causal evidence compilation rather than as an autonomous clinical decision-maker.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive societal impacts through improved auditability, verifiability, conflict detection, and evidence-grounded clinical decision support. It also addresses risks associated with incomplete evidence, non-executable queries, evidential conflicts, and deployment outside validated settings through explicit flags, provenance, and validation signals.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk generative model, pre-trained language model, image generator, or scraped dataset. Therefore, safeguards for responsible release of high-risk data or models are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing clinical studies, datasets, and related models used as evidence sources or baselines. HPP is explicitly identified as the external validation resource, and the paper uses these assets under their corresponding access and citation requirements.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release a new public dataset, model checkpoint, or standalone software asset as part of the submission. DoAtlas-1 is described as a compiled causal atlas and system instantiation rather than a separately released asset requiring additional documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not conduct new crowdsourcing experiments or recruit new human subjects. HPP is used as an existing validation resource, so participant instructions, screenshots, and compensation details are not applicable to this submission.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not report newly conducted human-subject experiments by the authors. The HPP data are used as an existing governed validation resource, so new IRB approval for participant recruitment is not applicable to this work.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method does not rely on LLMs as an important, original, or non-standard component. Any use of LLMs limited to writing, editing, or formatting would not affect the scientific methodology, rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.