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

NeurIPS 2026 [†]Corresponding author: imran.razzak@mbzuai.ac.ae

1 Why causal compilation?

Medical foundation models answer with fluent text and citations; clinical decisions need quantified, auditable intervention effects.

- **No interventional estimand**
  Contrast, scale, horizon and population stay implicit, so claims cannot be recomputed.
- **No effect estimation**
  No scale conversion or uncertainty propagation: no answer to "how large?"
- **No real-world validation**
  Claims are never tested against patient data.

	Narrative AI	Causal compilation
Represents	Text + citations	**Estimand objects**
Outputs	"Studies suggest…"	**θ = 0.73 [0.60, 0.88]**
Effect size	Qualitative	**Estimate + CI**
Conflicts	Implicit	**Explicit, typed**
Validation	None	**Real-world data**
Updates	Retraining	**Incremental**

2 Formal core

Estimand intermediate representation

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

population · contrast · outcome · horizon · measure; claim $c = (\theta, CI)$

Canonicalization $N : (\varepsilon, c) \to (\varepsilon_{canon}, c_{canon}, \alpha)$

12-mo MACE	Reported	Canonical
Study A	HR 0.73 [0.58, 0.92]	**logHR −0.315**
Study B	logHR −0.315	**same bucket**
Study C	RR 0.76 [0.61, 0.94]	**same family**

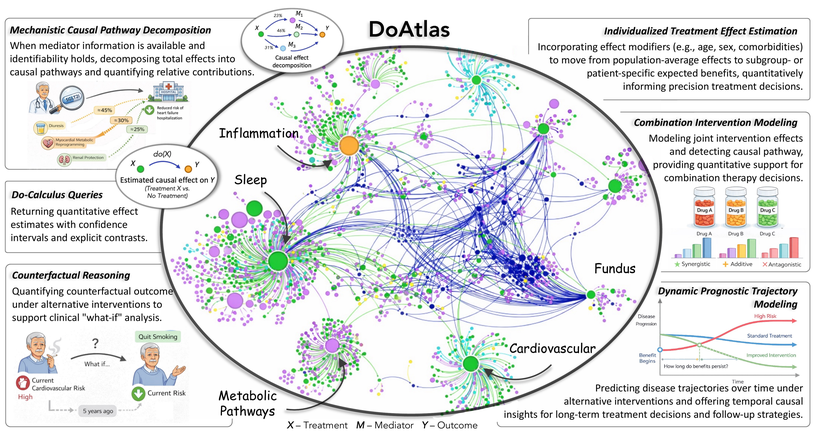
Theorem 1 Unique normal form
Canonicalization is deterministic, idempotent, semantics-preserving and invertible.

Theorem 2 Complete conflict detection
No false negatives for directional, interval or heterogeneity conflicts.

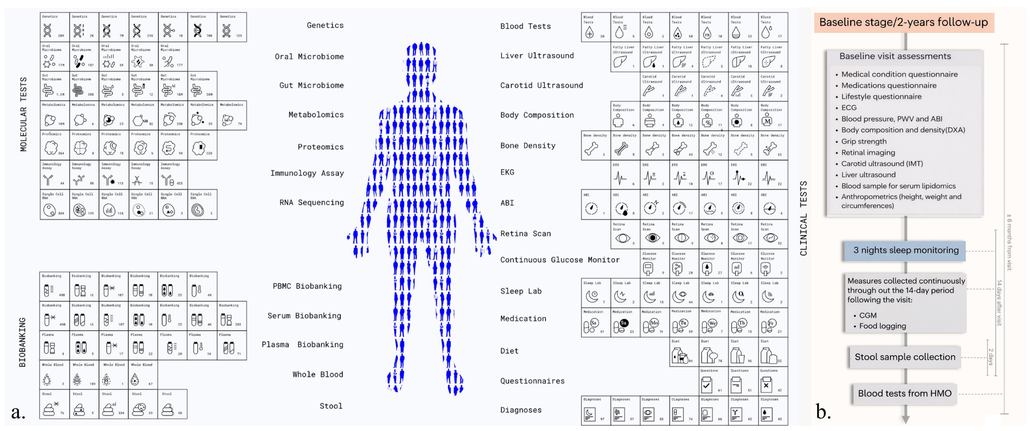
Proposition 1 Consistent evidence selection
A monotone, interpretable quality score picks the default effect.

Answer (ε, θ, CI, π, F(B), flags)
Every query returns estimand, estimate, CI, provenance, conflict annotation and status flags; missing conditions return a flag, never a made-up number.

3 DoAtlas-1: compile → organize → validate



Causal compilation. Effects from linear, logistic, Cox, MR and mediation models are standardized, pooled into comparable buckets with typed conflicts, and queried.



Human Phenotype Project. **10,000** participants · **17** physiological systems · multi-omics, imaging, DXA, CGM and sleep monitoring.

4 Executable queries

915 queries over six types; **80.3%** execute, the rest return an explicit flag.

Query	Asks for	Executable
Q_{do}	Interventional effect	200 / 250
Q_{cf}	Counterfactual	200 / 250
Q_{traj}	Trajectory over time	36 / 45
Q_{CATE}	Subgroup effect	163 / 200
Q_{med}	Mediation	29 / 36
Q_{joint}	Joint intervention	107 / 134

Example Q_{med}: BMI → CVD via KDM-BA
direct logOR **0.199** [0.113, 0.285], indirect **0.020** [0.010, 0.030]

Clinical decision support

Patient	do(·)	Outcome	Effect
68 y M, T2DM, HFrEF, eGFR 45	Dapagliflozin	HF worsening / CV death	**HR 0.73 (0.60–0.88)**
	Dapagliflozin	Kidney composite	**HR 0.61 (0.51–0.72)**
72 y F, HFpEF, dyspnea	Dapagliflozin	KCCQ at 12 wk	**+5.8 (2.3–9.2)**
64 y M, CKD 3, eGFR 38	Empagliflozin	Kidney progression / CV death	**HR 0.72 (0.64–0.82)**
70 y M, HFpEF, recent HF stay	Empagliflozin	HF hospitalization	**HR 0.71 (0.60–0.83)**

Each answer carries its CI, a verifiable DOI and an evidence grade.

5 Atlas & real-world check

1,113	1,053	362	705	46
nodes	edges	interventions	outcomes	mediators

Top intervention	Deg.	Top outcome	Deg.
TRE	22	**All-cause mortality**	41
Finerenone	17	**Stroke**	26
Dapagliflozin + CR	15	**Cardiovascular disease**	20
Dapagliflozin	15	**Myocardial infarction**	16
Early ECLS	14	**Heart failure**	13

External validation on HPP

Method	Direction agreement	False-positive rate
DoAtlas (ours)	**1.000**	**0.000**
No canonicalization	0.713	0.047
Text-only + cohort	0.712	0.630

Compiled claims mapped to HPP fields agree in direction with the cohort data, with no false positives.

Takeaway
Medical evidence becomes executable code: quantified, conflict-aware, auditable and checked on real-world data.

arXiv 2602.19158

6 Results

1,445
effect kernels from 754 studies

98.5%
semantic preservation (expert review)

0.982
mean F1 of estimand extraction

42
evidence conflicts found and typed

Compilation quality

Check	Score
Time horizon τ (extraction)	99.6%
Contrast ι (extraction)	98.3%
Population Π (extraction)	92.7%
Scale transformation (expert)	99.0%
Overall quality (expert, κ = 0.94)	97.0%

Ablation: canonicalization matters

Configuration	Bucket purity	Query acc.
Full DoAtlas	**96.8%**	**100%**
No scale normalization	68.3%	87.3%
No time alignment	83.7%	93.6%
Weak bucket key	78.9%	94.8%