

Patient-level evaluation of LLM-assisted OMOP phenotype curation: a benchmark against expert gold standards on MIMIC-IV

Hamid Tavakoli, MD, MSc (ORCID 0000-0003-0574-7739), Alireza Moayyeri, MD, PhD

Meta Health Informatics; UCB

hamid.tavakoli@metahealthinfo.com

Background

The OHDSI community has invested heavily in standardized phenotype curation through the Phenotype Library and Common Data Model. In practice, building a high-quality concept set is slow and resource-intensive; an expert informatician can spend days on a single cohort in ATLAS and ATHENA. The stakes are high too: errors at this layer propagate silently into every cohort, estimate, and downstream regulatory or HTA submission, and are rarely revisited once fixed. LLMs offer a way to automate this work, and recent studies have begun benchmarking LLM-generated concept sets against expert references.

Every published benchmark we are aware of evaluates performance through concept-set overlap metrics: precision, recall, F1, and Jaccard. These measure naming agreement but miss what actually matters: whether the concept set identifies the right patients in a real OMOP CDM database. Patient-level overlap can mislead, too. In coded EHR data, a single anchor concept plus descendant expansion can capture most affected patients regardless of what else gets enumerated. We therefore benchmark at two levels: concept-level agreement against OHDSI Phenotype Library references, and patient-level performance against both those references and a trivial-seed baseline on a 5,000-patient MIMIC-IV subsample. The trivial baseline is the key addition: it isolates how much the LLM actually contributes beyond what one obvious anchor concept already achieves through descendant expansion.

Methods

We built an R-based pipeline with six phases targeting the OHDSI inclusion concept-set layer, the foundational building block of cohort definitions. Temporal and algorithmic cohort rules (e.g., 'new-onset' AF, '>= 2 failed treatments' for TRD) are deliberately excluded; this study evaluates concept-set membership only.

Phase 1: LLM concept-name generation. A physician-authored plain-language brief is submitted to an LLM (Claude Opus 4.6 or Gemini 2.5 Pro) via a constrained system prompt requiring JSON output of concept names only. Concept_ids are explicitly prohibited since LLMs frequently hallucinate them, and a validation layer enforces this.

Phase 2: SQL concept resolution. A five-tier PostgreSQL cascade (exact name, synonym, and fuzzy matching with domain filtering) maps each concept name to a verified standard concept_id using the OHDSI standard vocabulary.

Phase 3: Ancestor expansion. Concepts flagged include_descendants = true are expanded via concept_ancestor.

Phase 4: Concept-level evaluation. Precision, recall, F1, Jaccard, and ancestor-aware F1 are computed against the gold-standard concept set.

Phase 5: Patient-level evaluation. The generated concept set is run against OMOP CDM clinical tables to identify a predicted cohort, compared against the OHDSI reference concept set on the same data (concept-set-derived reference, not clinician adjudication).

Phase 6: Trivial-seed baseline. One anchor concept name per phenotype (e.g., 'Type 2 diabetes mellitus') is resolved in vocab.concept, expanded via concept_ancestor, and run through Phase 5 to establish a saturation floor measuring the LLM's marginal contribution.

Dataset. 5,000 randomly sampled patients from MIMIC-IV v3.1, transformed to OMOP CDM v5.3 via the OHDSI/MIMIC ETL framework, evaluated against the May 2026 OHDSI standard vocabulary.

Phenotypes. Five from the OHDSI Phenotype Library: type 2 diabetes mellitus (288), cardiac valve surgery with new-onset atrial fibrillation (1103), acute liver injury excluding viral or alcoholic etiologies (736), drug-induced acute pancreatitis (253), and treatment-resistant depression (1009).

Design. Two LLMs x five phenotypes x three trials at temperature 0.3; 30 runs across 10 phenotype-model cells. Two design rules applied: lab tests were excluded from patient identification (routinely ordered in ICUs, unreliable disease indicators); and when OHDSI references include look-alike conditions, the brief named them too, applied to acute liver injury and drug-induced pancreatitis. Code, briefs, and results: <https://github.com/tavakohr/omop-phenotype-pipeline> (Apache 2.0).

Results

Across the 10 phenotype-model cells (30 runs), mean LLM patient F1 was 0.90 versus a trivial-seed baseline mean of 0.66, a +0.25 absolute lift (Table 1, Figure 1). Mean concept-level F1 (ancestor-aware) was 0.43; patient-level evaluation yielded a more optimistic view of LLM performance than concept-overlap metrics alone. The aggregate numbers, though, obscure what the per-phenotype breakdown actually shows.

When the OHDSI inclusion set requires enumerating drug classes, differentials, or treatment-specific ingredients, the LLM contribution is substantial: drug-induced pancreatitis +0.60 (trivial 0.28 to LLM 0.89), acute liver injury +0.46 (trivial 0.46 to LLM 0.92), and treatment-resistant depression +0.27 (trivial 0.58 to LLM 0.85). When one or two dominant anchor concepts suffice, the trivial baseline saturates and the LLM adds little: T2DM +0.02 (trivial 0.95 to LLM 0.97) and cardiac valve + new-onset AF -0.11 (trivial 1.00 to LLM 0.89).

Claude Opus 4.6 scored a mean patient F1 of 0.93 across the five phenotypes; Gemini 2.5 Pro 0.88. The paired difference (median +0.01, mean +0.05) is not statistically distinguishable at $n = 5$; both models track each other on the same strongest and weakest phenotypes.

Conclusions

LLM-assisted curation delivers real patient-level F1 gains for phenotypes requiring enumeration of drug classes, differential diagnoses, or treatment-specific ingredients (mean lift +0.44). Where one or two

anchor concepts capture the phenotype, descendant expansion saturates and the LLM adds nothing. The practical rule: use LLM curation for combinatorially complex inclusion sets; elsewhere, anchor-plus-descendants is sufficient.

Concept-overlap metrics alone are misleading: they can miss real LLM gains or inflate apparent ones. Reporting patient-level F1 alongside a trivial-seed baseline could help clarify how much the LLM actually contributed beyond simple ancestor expansion.

Beyond methodology, the trivial-seed comparison functions as a governance instrument: in regulated evidence contexts such as EU Joint Clinical Assessment and FDA real-world-evidence submissions, demonstrating an LLM's marginal contribution above a transparent baseline is exactly the auditability these processes increasingly require, and a concept set that cannot show its value over that baseline is a liability rather than an asset.

This study evaluates the inclusion concept-set layer, not full OHDSI cohort logic. The patient-level reference is concept-set-derived rather than clinician-adjudicated; so the pipeline is benchmarked on agreement with the OHDSI reference, not on correct identification of true patients, and an LLM that faithfully reproduces the reference also inherits whatever coding and ontology biases that reference encodes. Physician-authored briefs reflect standard operational protocol rather than model fine-tuning. The evaluation is single-center (MIMIC-IV ICU); replication on broader EHR populations is needed.

Figure 1. Patient-level F1 per phenotype on MIMIC-IV (5,000 patients): trivial-seed baseline (grey), Claude Opus 4.6 (orange), Gemini 2.5 Pro (blue). LLM bars: mean of 3 trials, ± 1 SD. Ordered by LLM lift over trivial.

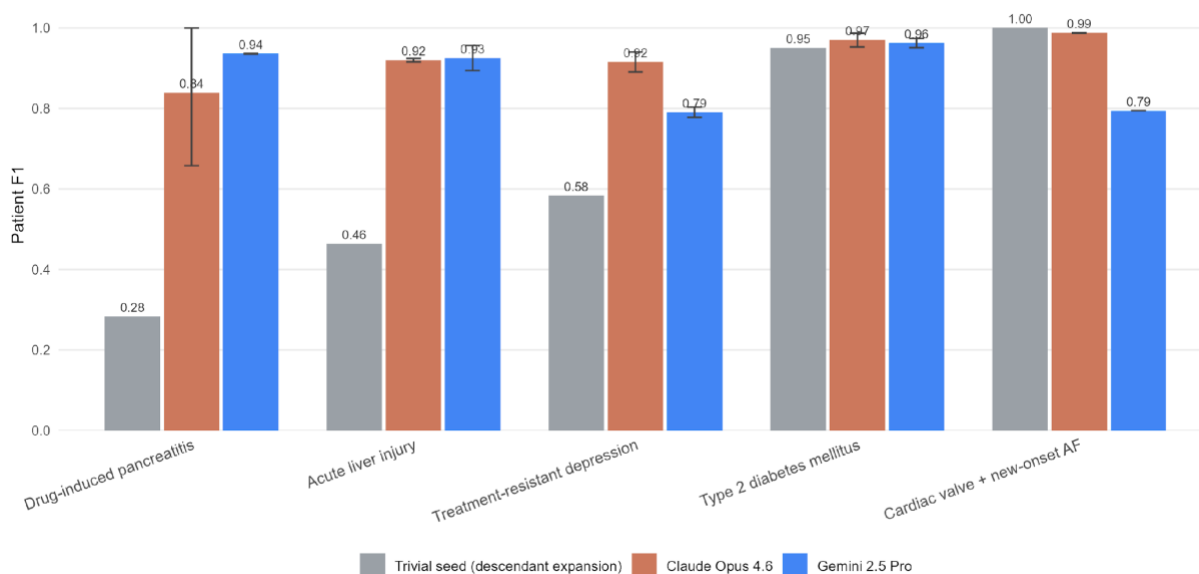


Table 1. Per-phenotype patient-level evaluation against MIMIC-IV (5,000 patients). Trivial-seed baseline (one anchor concept + concept_ancestor expansion) vs LLM-generated sets (Claude Opus 4.6, Gemini 2.5 Pro; 3 trials per cell, temperature 0.3). Phenotypes ordered by mean LLM patient F1.

Phenotype	Source	Patient F1 (mean $\pm$ SD)	Patient recall	Patient precision	Concept F1 (ancestor-aware)	Gold cohort size (n)
Type 2 diabetes mellitus	Trivial seed (descendant expansion)	0.95	0.90	1.00	—	662
Type 2 diabetes mellitus	Claude Opus 4.6	0.97 $\pm$ 0.02	0.94	1.00	0.10	662
Type 2 diabetes mellitus	Gemini 2.5 Pro	0.96 $\pm$ 0.01	0.93	1.00	0.19	662
Acute liver injury	Trivial seed (descendant expansion)	0.46	0.33	0.80	—	851
Acute liver injury	Claude Opus 4.6	0.92 $\pm$ 0.00	1.00	0.85	0.44	851
Acute liver injury	Gemini 2.5 Pro	0.93 $\pm$ 0.03	1.00	0.86	0.31	851
Cardiac valve + new-onset AF	Trivial seed (descendant expansion)	1.00	1.00	1.00	—	481
Cardiac valve + new-onset AF	Claude Opus 4.6	0.99 $\pm$ 0.00	1.00	0.98	0.08	481
Cardiac valve + new-onset AF	Gemini 2.5 Pro	0.79 $\pm$ 0.00	1.00	0.66	0.06	481
Drug-induced pancreatitis	Trivial seed (descendant expansion)	0.28	0.17	1.00	—	490
Drug-induced pancreatitis	Claude Opus 4.6	0.84 $\pm$ 0.18	0.96	0.78	0.75	490
Drug-induced pancreatitis	Gemini 2.5 Pro	0.94 $\pm$ 0.00	0.95	0.92	0.78	490
Treatment-resistant depression	Trivial seed (descendant expansion)	0.58	0.41	1.00	—	833
Treatment-resistant depression	Claude Opus 4.6	0.92 $\pm$ 0.02	0.85	1.00	0.77	833
Treatment-resistant depression	Gemini 2.5 Pro	0.79 $\pm$ 0.01	0.65	1.00	0.84	833
Mean (Trivial, n = 5)		0.66	0.56	0.96	—	
Mean (LLM only, n = 10 cells)		0.90	0.93	0.90	0.43	

Trivial seed: a single anchor concept name per phenotype, looked up in vocab.concept then expanded via concept_ancestor (same Phase 5 lab-domain filter as the LLM pipeline). T2DM: 'Type 2 diabetes mellitus'. Cardiac: 'Atrial fibrillation' + 'Heart valve replacement'. Acute liver injury: 'Disease of liver'. Drug-induced pancreatitis: 'Acute pancreatitis'. TRD: 'Major depressive disorder'.

Patient F1: harmonic mean of patient-level precision and recall against the cohort flagged by the OHDSI inclusion concept set after concept_ancestor expansion (concept-set-derived reference, not clinician-adjudicated).

Concept F1 (ancestor-aware): F1 crediting predicted concepts that are ancestor- or descendant-related to gold concepts via concept_ancestor; not applicable to the trivial seed (which by construction is one descendant subtree).

Gold cohort size: distinct MIMIC-IV patients flagged by the OHDSI inclusion concept set after concept_ancestor expansion. Cohort temporal/algorithmic logic (e.g. 'new-onset' AF for cardiac, '>= 2 failed treatments' for TRD) is NOT reproduced; this evaluates concept-set membership only.

References

1. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574-578.
2. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc.* 2015;22(3):553-564.
3. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc.* 2018;25(8):969-975.
4. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1.