

# TAXIS: Building an OMOP-native clinical relationship layer to support reusable OHDSI analytics

J. Marc Overhage, MD, PhD<sup>1,2</sup>; Stephen H. Bandean, MD, JD<sup>3</sup>; Shaun Grannis, MD, MS<sup>1,2</sup>; Gowtham Rao, MD, PhD<sup>4</sup>  
<sup>1</sup>Regenstrief Institute; <sup>2</sup>Indiana University; <sup>3</sup>Johns Hopkins School of Medicine, Biomedical Informatics & Data Science; <sup>4</sup>CoReason

## Background

OHDSI gives observational research a shared foundation in the OMOP Common Data Model, standardized vocabularies, and open-source methods.<sup>1</sup> The vocabularies say what a concept is and where it sits in a hierarchy, but not how concepts relate clinically: which drug treats a disorder, which test monitors it, which finding argues against it. Analysts rebuild that logic by hand for each phenotype, covariate set, and outcome definition. Knowledge graphs can represent such relationships,<sup>2</sup> and co-occurrence in health records can propose them at scale, but co-occurrence alone cannot establish clinical meaning.

TAXIS builds an OMOP-native relationship layer by pairing the two: association mining proposes concept pairs, and an ensemble of large language models (LLMs) assigns typed, directed relationships, each graded by its evidence.<sup>3,4</sup> Our 2025 report covered diagnosis-to-diagnosis pairs with one model and eleven labels.<sup>5</sup> This report extends TAXIS to five concept classes, describes the classification method in reproducible detail, measures it against physicians and four external reference sets, and shows a first application: generating phenotype definitions from the graph.

## Methods

**Candidate generation.** Association-rule mining ran on an Indiana Network for Patient Care (INPC) OMOP CDM subset of 2,157,525 persons and 11,299,055 person-years.<sup>6</sup> Concepts were typed into five classes: Disorder, Finding, Intervention (drug, procedure, device), Test, and Result (a test plus an HL7-style value category such as low, normal, or high). We grouped drugs by ingredient and dose form. Each pair carries person-level lift and support. The August 2026 export holds 5.52 million pairs with support of at least 100 in 14 of the 15 pair types (Result–Result pairs were not mined). We drew a prespecified stratified sample of 37,023 pairs (14 pair types × 4 lift bands) for classification.

**Taxonomy.** A relationship taxonomy (v6.0) defines 112 relationship codes across the 15 pair types, in 32 semantic families that fall into five groups: causes, worsens or predisposes; sign, finding or test evidence; treats, manages or prevents; adverse effect or contraindication; and procedural or hierarchical. Each code has a definition, a fixed direction, and qualifiers, and every pair type offers “no direct relationship” and “too abstract to judge”. A relationship must hold between the two concepts without an unstated third concept. Hierarchical relationships are left to the OMOP vocabulary.

**LLM ensemble.** Classification has two stages. A relatedness screen asks four inexpensive models (gpt-4.1-mini, Gemma, gpt-oss-120b, Qwen) whether the two concepts are directly clinically related, yielding 0–4 screen votes. Four coders (Claude Sonnet 5, GPT-5, Gemma-4-31B, gpt-oss-120b) then receive the catalog for the pair type and assign every applicable relationship, with direction, qualifiers, and a rationale. Models are used as released, with instruction-only prompts generated from the versioned catalog: no fine-tuning and no examples drawn from any validation set. Temperature is 0 where a model accepts it; GPT-5 ran at minimal and Claude Sonnet 5 at high reasoning effort. In this sample, a concept screen first set aside container concepts such as “Disorder of respiratory system”. We store every vote with the model, prompt version, and run identifiers. We derive an evidence grade from the votes, screen first: Strong (screen majority and at least two coders), Moderate (screen majority, one coder), then Candidate, Weak, and Refuted. Strong and Moderate edges are published; Refuted pairs are kept as negative knowledge.

**Validation.** We compared the ensemble output with four reference sets that do not derive from INPC: ClinVec clinician ratings,<sup>7</sup> PACES expert-curated episode associations,<sup>8</sup> relationships implied by OHDSI Phenotype Library cohort definitions,<sup>9</sup> and SemMedDB literature predictions.<sup>10</sup> All coded pairs were also checked against SNOMED CT and UMLS relationships. Two physicians (internists; authors JMO and SHB) independently judged 291 INPC pairs, drawn across the screen-vote × coder-vote grid and covering all 14 pair types. Each decided whether a direct relationship exists and which code applies. We reweighted rates from the stratified draw to the coded population.

**Application.** A template-driven builder read the graph’s edges around a disease concept (subtypes, treatments, results, “confusable with” rule-outs), added OMOP vocabulary descendants, and wrote an executable Circe cohort definition. We compared five generated cohorts with their OHDSI Phenotype Library counterparts, as concept sets and as patients in three synthetic OMOP CDMs, using HADES CirceR and CohortGenerator.

## Results

**Classification at scale.** Of the 37,023 sampled pairs, 26,901 (72.7%) passed the concept screen and were coded (Figure 1). At least two coders asserted a typed relationship for 4,085 (15.2%). Causal relationships led (38%),

followed by signs, findings, and test evidence (22%); treatment, safety, and procedural relationships were 13–14% each. The share asserted rose with lift, from 8.9% below lift 1 to 27.0% at lift 20 or more, but varied more by pair type (44% for intervention–finding, 2% for test–test). Grading left 3,020 pairs Strong and 999 Moderate. Projected over all 5.52 million mined pairs, the graph would hold about 1.9 million published edges (range 1.4–2.1 million) among about 17,000 concepts; 46,650 published edges touching 5,644 concepts were loaded as of September 25 2026.

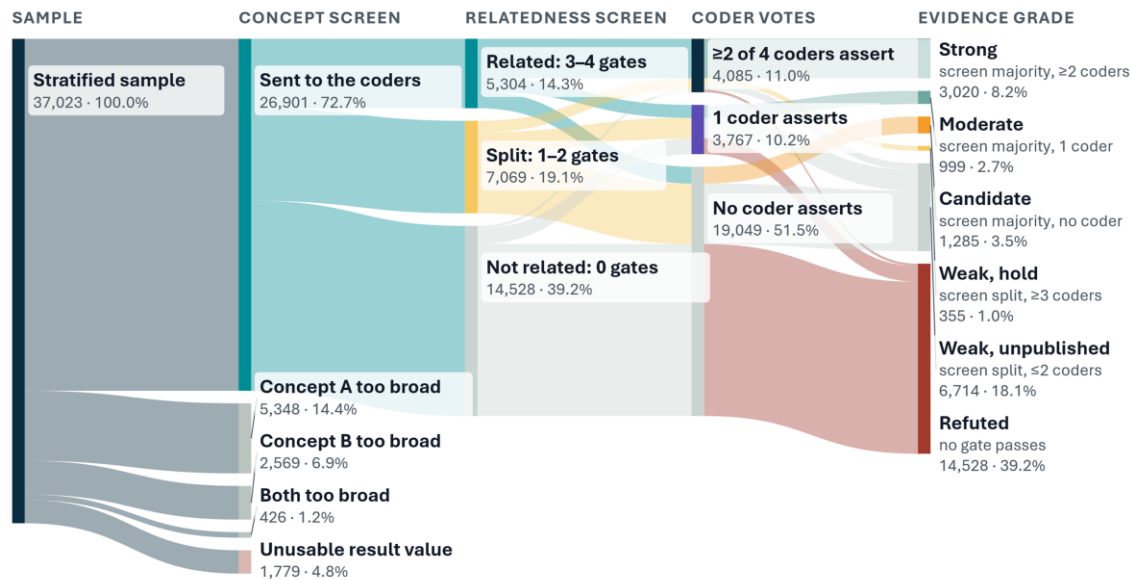


**Figure 1.** Flow of the 37,023 sampled INPC concept pairs through the concept screen, the relatedness screen (four screening models, labeled gates), and the four-coder vote to an evidence grade. The width of the band is proportional to the number of pairs. Strong and Moderate edges are published; Refuted pairs are retained as negative knowledge.

**Agreement with reference sets and physicians.** Table 1 summarizes the validation. The coder vote count separated related from unrelated ClinVec pairs with an area under the curve of 0.81. In our June analysis, person-level lift alone agreed only moderately with the same clinician ratings (Spearman  $\rho = 0.46$ ). Co-occurrence also reflects shared causes, shared care settings, and routine ordering; it is unstable at low counts; and the clinicians rated relatedness, not the existence of a direct relationship. Of the 193 confident disagreements we adjudicated, 45 were clear model errors. Because a public benchmark could have entered model training data, we also conducted physician review of unpublished INPC pairs.

**Table 1.** Validation against external reference sets, standard terminologies, and physician review.

| Reference (size)                                          | Question                            | Result                                                                                                                   |
|-----------------------------------------------------------|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| ClinVec clinician ratings <sup>7</sup> (1,976 pairs)      | Is the pair related?                | AUC 0.81 for the 0–4 coder vote count; at two or more coders, sensitivity 0.72 and specificity 0.77.                     |
| PACES expert associations <sup>8</sup> (2,850 assertions) | Are expert associations recovered?  | Recall 68% per assertion; 97% after blinded review removed PACES's own errors. Direction agrees in 99%.                  |
| OHDSI Phenotype Library <sup>9</sup> (1,629 cohort rules) | Is curated cohort logic recovered?  | Recall 56% per cohort rule; 73–92% for rule types the literature confirms.                                               |
| SemMedDB literature <sup>10</sup> (130.5 M predications)  | Does the literature agree?          | Supports 59% of PACES associations and contradicts none; 8 of 51 literature-only flags the ensemble declined were real.  |
| SNOMED CT and UMLS (26,901 coded pairs)                   | Do terminologies hold this already? | A relationship exists for 119 pairs (0.4%); where one is asserted exactly, the ensemble agrees in 99%.                   |
| Two physicians (291 INPC pairs)                           | Precision and recall on mined pairs | At two or more coders: precision 56% and 54%, recall 67% and 76%*, specificity 92%. Strong grade: precision 72% and 67%. |

\*Population-weighted, with one heavily weighted pair set aside.

For the first physician, 56% (95% CI 39–69) of pairs asserted by two or more coders and 72% (48–89) of Strong pairs held as direct relationships; for the second, 54% (37–65) and 67% (44–82). Requiring both physicians to agree lowers these to 38% and 49%; accepting either raises them to 71% and 89%. The physicians agreed with each other on 79% of pairs (Cohen's  $\kappa$  0.40 on the sample as drawn; 0.58 when population-weighted with one outlying pair set aside), and the screen majority agreed with each physician about as well (weighted  $\kappa$  0.68 and 0.60); no single model did. Where the first physician confirmed a relationship, the ensemble's top code matched in 58% of pairs and its relationship group in 88%. Most rejections involved result-status, encounter-context or normal-value concepts that can be screened out before coding.

**Application and reliance on results.** Table 2 compares five graph-generated cohort definitions with their curated counterparts. Diagnosis concept sets matched closely for four of five phenotypes, and for four of five, every generated patient was also in the curated cohort. The gaps were in entry logic and numeric thresholds, such as a body-mass-index cut-off for obesity, not the graph's relationships. Pair types that include a Result make up 55% of the mined frame and 48% of projected published edges. The other 52% need no result values, and 25% involve only disorders, findings, and interventions. On a claims-type CDM without laboratory values, the generated type 2 diabetes cohort overlapped the curated one at 99.5%.

**Table 2.** Phenotype definitions generated from TAXIS edges compared with OHDSI Phenotype Library cohorts.

| Phenotype (Library cohort)    | Diagnosis codes: recall / precision | Overlap, SynPUF | Overlap, Synthea | Overlap, Synthea + labs | Generated also in Library |
|-------------------------------|-------------------------------------|-----------------|------------------|-------------------------|---------------------------|
| Type 2 diabetes (1032)        | 0.94 / 0.99                         | 99.5%           | 100%             | 100%                    | 100%                      |
| COPD (1263)                   | 0.96 / 0.77                         | 56%             | 97%              | 97%                     | 68–100%                   |
| Chronic kidney disease (1191) | 0.18 / 0.78                         | 76%             | 99.8%            | 99.1%                   | 100%                      |
| Hyperkalemia (940)            | 1.00 / 1.00                         | 86%             | no cases         | no cases                | 100%                      |
| Obesity (1179)                | 0.96 / 1.00                         | 78%             | 0%               | 0%                      | 100%                      |

Overlap is the Jaccard index of generated and Library cohorts in CMS SynPUF 1k (claims), Synthea 117k and Synthea 115k with laboratory values; the last column is the share of generated patients also in the Library cohort. Cohort 1191 codes all kidney disease. COPD: a 10% exclusion-cost cap dropped the asthma rule-out. Obesity: Synthea records it as a body-mass-index finding, a route the graph does not yet express.

## Conclusions

TAXIS now types and grades clinical relationships across five concept classes, and its validation no longer rests on one condition-centered benchmark. The ensemble separates related from unrelated pairs, names the relationship group well and the exact code adequately, and assigns direction reliably. Precision is the binding constraint at scale: about half of the pairs asserted by two or more coders hold under a strict physician reading, and about seven in ten when the relatedness screen agrees, so we publish edges with evidence grades rather than as a binary graph. SNOMED CT and UMLS carry a relationship for only 0.4% of the same pairs, so the layer adds knowledge the vocabularies lack, and it can already draft phenotype definitions that agree closely with curated ones.

**Limitations.** Mining ran on one data source; the reference sets and synthetic CDMs test classification and application, not the portability of mining. The physician readers are authors and agreed only moderately, and the grade cut was chosen on the sample it is reported on. Edges that need result values are less useful where results are not captured.

**Availability and next steps.** The code is available in the OHDSI github repository along with an initial TAXIS Core KG

## References

1. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics: opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574-578.
2. Cui H, Lu J, Xu R, Wang S, Ma W, Yu Y, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *J Biomed Inform.* 2025;169:104861. doi:10.1016/j.jbi.2025.104861.
3. Bandeian S, Tompkins CP, Davison A. A future health care analytic system: Part 1 - What the destination looks like. In: *Healthcare Information Management Systems*. 5th ed. Springer; 2022.
4. Bandeian S, Tompkins CP, Davison A. A future health care analytic system: Part 2 - What is needed and getting it done. In: *Healthcare Information Management Systems*. 5th ed. Springer; 2022.
5. Bandeian SH, Overhage JM. TAXIS: enhancing OMOP CDM capabilities. *OHDSI Global Symposium*; 2025.
6. Dixon BE, Wen C, French T, Williams JL, Duke JD, Grannis SJ. Extending an open-source tool to measure data quality: case report on Observational Health Data Science and Informatics (OHDSI). *BMJ Health Care Inform.* 2020 Mar;27(1). doi:10.1136/bmjhci-2019-100054.
7. Johnson R, Gottlieb U, Shaham G, Eisen L, Waxman J, Devons-Sberro S, et al. ClinVec: unified embeddings of clinical codes enable knowledge-grounded AI in medicine. *medRxiv.* 2025 May 7. doi:10.1101/2024.12.03.24318322.
8. PACES Center for Value in Healthcare. Episode Definition Data (EDD), version 5.4.0. **complete citation: publisher details, year, URL**
9. OHDSI. Phenotype Library. <https://github.com/OHDSI/PhenotypeLibrary>.
10. Kilicoglu H, Shin D, Fiszman M, Rosembat G, Rindfleisch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics.* 2012;28(23):3158-3160.

## Working notes for co-authors (delete this page and everything after it before making the PDF)

*Draft of October 1 2026, revised from the brief report submitted on June 5 2026 (v8). It was rewritten rather than redlined because most of the results are new since June. Nothing below this line is part of the submission.*

**1. What OHDSI asked for.** The acceptance email of July 15 2026 (poster 127) says to “feel free to edit your brief report based upon the reviewers’ feedback”. The email of September 24 asks for the final brief report as a PDF, with the poster, through the Dropbox request link in that email by **Friday October 9 2026**, even if it is unchanged. File names carry the poster number, last name and title, for example 127-Overhage-TAXIS-BriefReport.pdf. Changes to the title, authors or author order on the printed handout are due Monday October 5; OHDSI confirmed the new author order on September 29. The emails give no page limit for the final version; this draft keeps the three pages of the June submission.

## 2. Reviewer comments and where this draft answers them.

| Reviewer comment                                                                                                                                                                      | How the draft responds                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Where                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|
| <b>Reviewer 1, point 1.</b> Done in one data set; needs validation in other data sets before the community can rely on it.                                                            | Stated as the first limitation. New since June: the classification is checked against four reference sets that do not come from INPC, plus SNOMED CT and UMLS, and graph-built phenotypes ran on three other OMOP CDMs. Mining itself has still run only on INPC, those three CDMs are synthetic, and the draft says so and invites data partners. If a second site is lined up, name it.                                                                                                                                                                        | Methods (Validation, Application); Limitations; last paragraph |
| <b>Reviewer 1, point 2.</b> Show the LLM work. Could ClinVec have been in the LLM training data?                                                                                      | Methods now name the eight model roles, the two-stage design, the taxonomy, the prompt style (instruction only, no fine-tuning, no examples from validation sets), temperature and reasoning settings, and stored provenance. On leakage: it cannot be excluded for a public benchmark, so the draft rests the case on evidence no model could have memorized, which is the blinded review of unpublished INPC pairs by two physicians. ClinVec’s clinicians also rated relatedness, a different question from the typed, direct relationship the coders assign. | Methods (LLM ensemble); Results, second paragraph and Table 1  |
| <b>Reviewer 1, point 3.</b> Spearman and rank-biserial correlations are only moderate. Why?                                                                                           | Explained in Results: lift measures co-occurrence, which also reflects shared causes, shared settings and routine ordering; it is unstable at low counts; and ClinVec scores relatedness, not a direct relationship. Lift is used only to order and weight pairs. The typed ensemble tracks clinicians better (AUC 0.81), and 45 of 193 adjudicated disagreements were true model errors.                                                                                                                                                                        | Methods (Candidate generation); Results, second paragraph      |
| <b>Reviewer 1, point 4 and Reviewer 2.</b> Give an example of use; expand real-world applications.                                                                                    | New Application section and Table 2: phenotype definitions generated from graph edges for five diseases, compared with Phenotype Library cohorts as concept sets and as patients. Other uses on the WIP deck (negative-control and confounder candidates, prescribing cascades, care-pathway skeletons) are not in the draft because they have not been evaluated. Add one sentence if you want them named as next uses.                                                                                                                                         | Methods (Application); Results, last paragraph; Table 2        |
| <b>Reviewer 1, point 5.</b> Many relationships rest on measurement and observation results, which many datasets lack.                                                                 | Now quantified: Result pair types are 55% of the mined frame and 48% of projected published edges; 52% of edges need no result values and 25% involve only disorders, findings and interventions. The type 2 diabetes cohort generated from the graph matched the curated one at 99.5% on a claims CDM with no laboratory values. Edges are knowledge derived where results exist and usable where they do not.                                                                                                                                                  | Results, last paragraph; Limitations                           |
| <b>AI review 1 and 2.</b> Methodological transparency: models and versions, prompts, hyper-parameters, few-shot provenance, public code and SQL.                                      | Models, prompt style, temperature and reasoning settings are in Methods. The availability sentence has a highlighted placeholder for the repository link. Decide what is public by October 9: the classification code and prompts, and the candidate-generation SQL.                                                                                                                                                                                                                                                                                             | Methods (LLM ensemble); last paragraph                         |
| <b>AI review 1 and 2.</b> Precision, recall and agreement against clinicians on a stratified human-annotated sample covering all pair types; no gold standard beyond condition pairs. | Two physicians judged 291 INPC pairs blind, stratified across the vote grid and all 14 pair types. The draft reports precision, recall, specificity and kappa, plus recall against PACES and the Phenotype Library, which reach drug, test and procedure pairs. F1 is not reported; it can be computed from Table 1 if wanted.                                                                                                                                                                                                                                   | Methods (Validation); Table 1; Results, third paragraph        |
| <b>AI review 1 and 2.</b> Governance: versioning, provenance, refresh, conflict resolution.                                                                                           | One sentence on stored provenance and the planned OMOP-vocabulary-format release with an evidence table. Coder disagreement is kept as votes and resolved by the grade, not discarded. Refresh cadence is not decided and is not mentioned.                                                                                                                                                                                                                                                                                                                      | Methods (LLM ensemble); last paragraph                         |
| <b>AI review 1.</b> Computational requirements and run times.                                                                                                                         | The draft gives the measured model cost of the sample (about US\$440). Not included: the screen’s cost (about US\$0.27 per 1,000 pairs), total validation spend (under US\$2,000), a full-frame estimate (the September 7 cost note projects a few thousand dollars and about a week of wall-clock time, before later pipeline changes) and IU’s mining run time. Add what you are comfortable stating.                                                                                                                                                          | Last paragraph                                                 |
| <b>AI review 1.</b> Real-world impact such as time saved, covariate balance or signal discovery.                                                                                      | Not measured, so not claimed. The phenotype comparison is the only application reported.                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Not addressed                                                  |

### 3. Decisions and open items.

1. **Authors and affiliations** follow the September 28 email to OHDSI and poster v5: Overhage (Regenstrief, Indiana University), Bandeian (Johns Hopkins), Grannis (Regenstrief, Indiana University), Rao (CoReason). The June submission listed The Overhage Group for Overhage and coreason.ai for Rao. Confirm.
2. **Cohort size.** 2,157,525 persons and 11,299,055 person-years are IU's June 2026 profile. The profile of the August run was still pending on September 25. Confirm that the June figures describe the August export, and how its 5.52 million pairs (support of at least 100) relate to the 14.2 million observed pairs and 95,968 concepts reported in June.
3. **Which precision to quote.** The first physician's 56% and 72% are the September 17 reweighting used on the poster (unsure answers excluded). The two-reader analysis of September 26, which counts unsure as no, gives 55% and 71% for the first physician; the second physician's and the both/either figures come from that analysis. After the first physician's reconciliation pass of September 29 the figures are 68% and 85% (both physicians: 54% and 67%); they are not used because that pass was not blind and the second physician's review of the 26 remaining disputes is pending.
4. **Recall figures** (67% and 76%) set aside pair #9, one sampled pair that stands for about 2,000 in the weighting. With it included the second physician's recall falls to 32%. The table footnote says a pair was set aside.
5. **Graph size.** 46,650 published edges touching 5,644 concepts is the September 25 figure on the poster. The batch 4 export of September 29 holds 54,237 published edges among 9,410 concepts but is not yet loaded. Keep the report and poster on the same number.
6. **Repository link.** The poster change log names [github.com/overhage/cab-llm](https://github.com/overhage/cab-llm) and notes its visibility was not checked. Fill the highlighted placeholder or reword the sentence.
7. **References.** Reference 8 (PACES Episode Definition Data v5.4.0) needs a complete citation. MED-RT and HADES are named in Methods without references to save space.
8. **Disclosures.** The brief report has no disclosure line. Consider one if any author has a relationship with PACES or CoReason that readers should know about, and note that the two physician readers are authors (the draft already says so).
9. **Terms.** The draft uses "LLM ensemble" and "coder" for the models, "screen" for the relatedness stage (Figure 1 labels the same models "gates"), and "association" for PACES records.
10. **Poster consistency.** Numbers match poster v5 of September 25. The second-reader results (26 to September 29) are in this draft but not on the poster, whose limitations still say a second reader is pending.

### 4. Where the numbers come from (all in the Knowledge Graph folder).

- Reviewer comments and deadlines: Gmail, OHDSI Events Manager, July 15, September 21 and September 24 2026.
- Sample flow, grades and Figure 1: `analysis/v60_evidence/v60_sankey_flows.json`; `cab_llm/docs/INPC_FIRST_RUN.md`.
- Reference-set results, Table 1 rows 1 to 5: `analysis/TAXIS_validation_synthesis_2026-09-08.md` (revised September 29), findings F1 to F4 and F10.
- Blinded review, reweighted: `analysis/v60_evidence/a19_cell_weights.py` (17 September); two readers: `v60_interrater_report.txt`; physician against model: `v60_human_vs_gate_report.txt`; reconciliation: `v60_reconciliation_report.txt`.
- Full-corpus projection and shares by pair type: `poster_build/v5/analysis_scripts/estimate/projection_summary.json` and `projection_by_type.csv`; `TAXIS_OHDSI_2026_poster_v5_changes.md`.
- Phenotype application, Table 2: `phenotype_builder/poster_panel/phenotype_panel.pptx` and `phenotype_builder/BUILDER.md` (Circe runs of 24 and 25 September).
- Models and settings: `cab_llm/models.yaml`, `cab_llm/src/cab_llm/prompts_v6_0.py` and `prompts_gate.py`; cost: `cab_llm/docs/COST_LEVERS.md`.