

Reproducing gold-standard OHDSI cohorts with an LLM agent: an execution-based, patient-level evaluation of the FastOMOP ATLAS Agent

Niko Möller-Grell¹, Shihao Shenzhang¹, Ian Farr^{6,7}, Richard Dobson^{1,2,3,4,5}, Vishnu V Chandrabalan^{6,7}

1 Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK

2 Institute for Health Informatics, University College London, London, UK

3 NIHR Biomedical Research Centre, University College London Hospitals National Health Service Foundation Trust, London, UK

4 Health Data Research UK London, University College London, London, UK

5 NIHR Biomedical Research Centre, South London and Maudsley National Health Service Foundation Trust and King's College London, London, UK

6 Lancashire Teaching Hospitals Trust, Preston, UK

7 Lancaster University, Lancaster, UK

Background

Large-language-model (LLM) agents promise to translate natural-language phenotype descriptions into executable OHDSI/ATLAS CIRCE cohort definitions, yet rigorous evaluation is lacking. Most assessments compare generated text or structure; what ultimately matters is whether the generated definition selects the same patients as a trusted reference. We report an exact, execution-based evaluation of an LLM cohort-definition agent against an OHDSI phenotype gold standard.

Methods

We evaluated the FastOMOP ATLAS Agent [1], an interactive GPT-4.1 agent that converts a clinical description into an OMOP CIRCE cohort definition, resolving clinical terms to standard OMOP concepts via a vector concept-retrieval service. The FastOMOP ATLAS agent reproduces the full capability suite of ATLAS cohort creation towards a live OMOP CDM and allows for interactive cohort creation, size preview, cohort comparison and cohort profiling. Using a purpose-built evaluation harness [2], we drew a structurally stratified sample of 100 cohort definitions (50 SIMPLE / 30 MODERATE / 20 COMPLEX, graded by number of entry events, inclusion rules, correlated criteria and concept exclusions) from the OMOP Cohort Ground-Truth dataset [3], whose logic-bearing pairs derive chiefly from the peer-reviewed OHDSI Phenotype Library [4]. For each phenotype, both the agent-generated and the gold CIRCE were compiled by the canonical OHDSI circe-be engine (ATLAS Java backend) to SQL and executed against a MIMIC OMOP CDM in DuckDB. Agreement was the Jaccard index of the resulting patient sets, with exact-set match as a secondary endpoint. Using the identical engine and substrate for both arms makes the comparison apples-to-apples and ATLAS-faithful; phenotypes absent from the ICU-centric substrate were excluded.

Results

Across 100 cohorts the mean patient-set Jaccard was 0.64 (median 0.90); 44% of cohorts reproduced the gold patient set near-exactly (Jaccard ≥ 0.95), 68% reached ≥ 0.50 , and 25% were exact-set matches. Accuracy declined monotonically with structural complexity Table 1: mean Jaccard 0.76 (SIMPLE), 0.58

(MODERATE) and 0.45 (COMPLEX). The dominant error mode was concept-set granularity - generated cohorts too narrow (under-selection, 14%) or too broad (over-selection, 16%) relative to gold - rather than errors in cohort entry/inclusion logic (structural mismatch, 2%) (Figure 1).

The histogram in Figure 1 shows that the majority of cohorts were reproduced near-perfectly and the low mean was produced by the skewedness of the Jaccard-distribution, rules introduced by the agents in autonomous cohort creation of mostly complex cohorts led to catastrophic collapse of some and resulted in Jaccard overlaps of 0.

Stratum (n)	Mean Jaccard	Median	Exact-set match	Jaccard >=0.95
SIMPLE (50)	0.76	0.95	36%	52%
MODERATE (30)	0.58	0.67	7%	40%
COMPLEX (20)	0.45	0.37	25%	30%
Overall (100)	0.64	0.90	25%	44%

Table 1. Patient-set agreement between the FastOMOP ATLAS Agent and gold CIRCE, by structural-complexity stratum (n = 100).

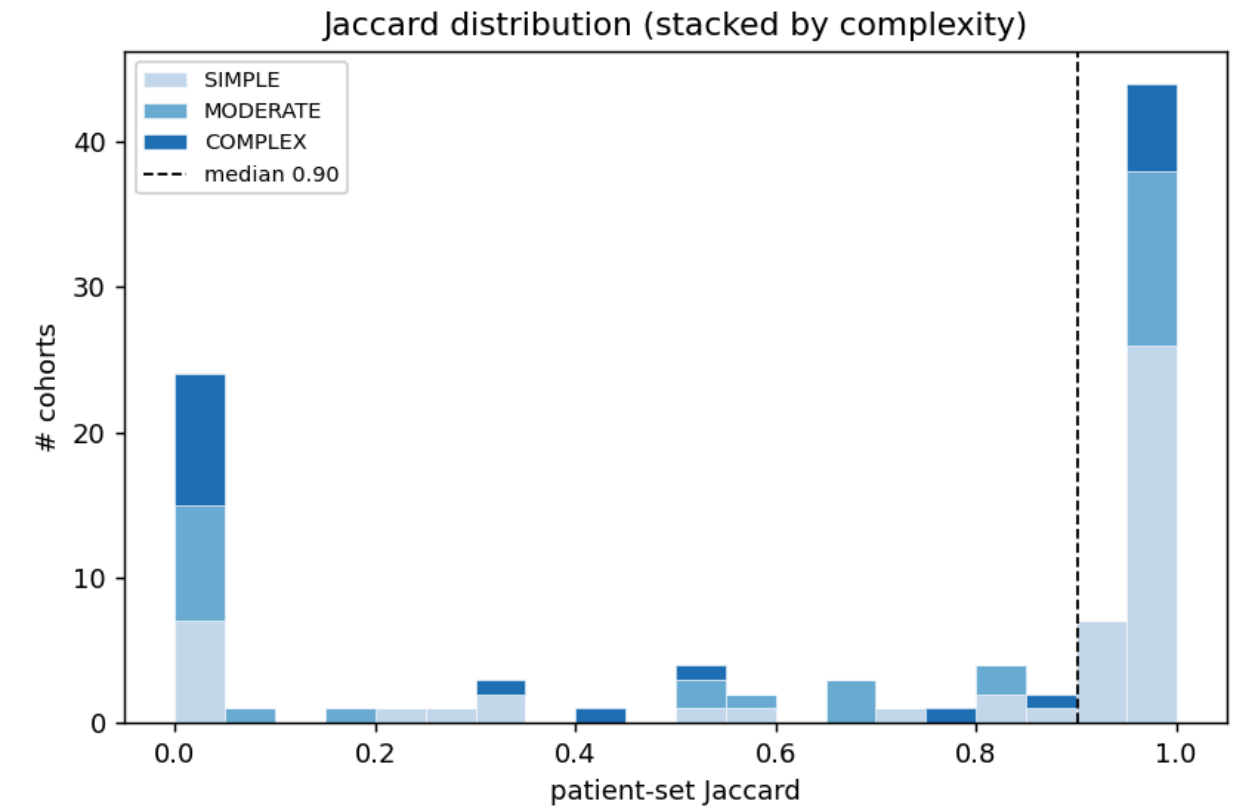


Figure 1. Patient-set Jaccard overlap across 100 stratified cohorts: distribution by complexity

Conclusions

Measured by exact execution on a real OMOP CDM, an LLM agent reproduces a substantial share of OHDSI gold-standard phenotypes; the high median (0.90) shows that most cohorts are reproduced

perfect or near-perfect, with residual error concentrated in concept-set breadth and in structurally complex definitions. This draws attention to the intended use of agentic pipelines in cohort generation. While agents are quite capable in producing semantically valid phenotypes and simpler cohorts, for RWE grade complexity, the agent should assist in the execution rather than tackle the task autonomously. Execution-based, patient-set evaluation provides a rigorous, ATLAS-faithful benchmark for agentic phenotyping and pinpoints the interactive cohort design as the priority for improvement. The gold standard, evaluation harness and agent are all contributed to the community.

References

1. FastOMOP ATLAS Agent. <https://github.com/fastomop/fastomop-atlas-agent>
2. Cohort evaluation harness (cohort-eval-lab). <https://github.com/fastomop/cohort-eval-lab>
3. OMOP Cohort Ground-Truth dataset. <https://github.com/fastomop/omop-cohort-gold-standard>
4. OHDSI Phenotype Library. <https://github.com/OHDSI/PhenotypeLibrary>