

FastSSV: A Semantic Static validator for LLM-Driven Clinical Queries on the OMOP CDM

Shihao Shenzhang¹, Niko Möller-Grell¹, 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

The OHDSI community has developed a broad set of tools for OMOP-based research, including ACHILLES for database characterisation, DataQualityDashboard for data quality assessment, CohortDiagnostics for cohort and phenotype evaluation, PheValuator for phenotype validity assessment, and PatientLevelPrediction for building and validating predictive models. Within this ecosystem, analysts can work through custom SQL, reusable OHDSI R packages, or interactive tools such as ATLAS.¹

Recent advances in large language models (LLMs), Model Context Protocol (MCP)-based interfaces, and agentic frameworks are enabling researchers to query OMOP databases directly through natural language. This lowers technical barriers to generating real-world evidence, particularly for clinical researchers who may not routinely write SQL. At the same time, it changes the workflow's safety assumptions. Traditional OHDSI tools constrain how users interact with the CDM and, in doing so, provide a degree of protection against invalid analytical patterns. When SQL is generated by an LLM, these constraints may no longer apply.

The main concern is not syntax. An LLM-generated query can succeed while still being semantically incorrect. For example, it may use non-standard concepts where standard concepts are required, traverse concept hierarchies incorrectly, join tables through inappropriate paths, or define temporal windows in a way that leaks future information. These errors are difficult to identify because they often return plausible results rather than obvious failures.

Recent work has shown that LLM-generated clinical SQL remains highly dependent on domain-specific context. In pharmacovigilance tasks, reported accuracy of natural-language queries to SQL ranged from 8.3% when using schema information alone to 78.3% when additional contextual guidance was provided.² This highlights the importance of domain knowledge, but it also leaves a separate problem unresolved: existing OHDSI tools do not directly verify whether LLM-generated SQL conforms to OMOP-specific semantic constraints before it is executed.

Methods

We introduce the Fast Semantic Static Validator (FastSSV), an open-source Python framework that acts as a semantic guardrail layer for OMOP SQL. FastSSV validates queries before execution by checking them against rules derived from the analytical assumptions of the OMOP CDM v5.4 specification. The framework is AI-agnostic and can be applied to SQL written by researchers, generated by scripts, produced by LLMs, or obtained from other query-generation workflows.

FastSSV uses static analysis. Each SQL query is parsed with `sqlglot3` into a canonical abstract syntax tree (AST), which is then evaluated by a plugin-based rule engine without executing the query against a live database. Because validation occurs before execution, FastSSV does not require database credentials or access to patient-level data, making it suitable for privacy-sensitive multi-site settings.

The current release implements 154 rules across six categories: anti-patterns, concept standardisation, joins, temporal logic, data quality, and domain-specific constraints. These rules check for OMOP-specific semantic problems, such as inappropriate source-concept use, unsupported join paths, incorrect temporal anchoring, future information leakage, and domain-specific inconsistencies. Each rule is implemented as an independent plugin and returns zero or more structured violations. Each violation includes a rule identifier, severity level, a human-readable message, and, where available, the query location. FastSSV returns violations as structured JSON, supporting both manual review and automated feedback loops in which an LLM or agentic system can revise a query based on the detected semantic errors.

Results

We applied FastSSV to 11,654 OMOP SQL queries generated by an AI-assisted query-generation workflow across real-world analytical tasks. Overall, 47.5% of queries violated at least one CDM convention, resulting in 12,413 violations. Most violations (89.8%) were semantic rather than syntactic, most commonly involving temporal anchoring, concept standardisation, deprecated concepts, and hierarchy expansion. Vocabulary-related violations frequently reflected unsafe string-based filtering rather than concept-identifier-based lookups. These findings suggest that semantic violations are common in AI-generated OMOP SQL and may not be detected solely through execution-based validation.

Conclusion

As agentic AI systems are increasingly adopted for clinical data analysis, OMOP-based research workflows will require safeguards that can operate before generated SQL reaches the database. FastSSV addresses this need by providing a semantic guardrail layer for OMOP SQL that validates whether queries accurately encode OMOP analytical assumptions, regardless of how they are written. By treating analysis logic as a constraint-checking problem, FastSSV provides a plugin-based, AI-agnostic, and deterministic framework for identifying silent semantic violations before queries are executed or reused in downstream analyses. Its structured outputs support both human review and AI-assisted refinement, while its extensible design allows new rules to be added as study designs, governance requirements, and analytical patterns evolve. We encourage the OHDSI community to evaluate FastSSV in real study workflows, contribute additional rules, and help formalise the analytical conventions that currently remain implicit in OMOP-based research. FastSSV is available at <https://github.com/fastomop/FastSSV>.⁴

References

1. Informatics OHDS and. The Book of OHDSI [Internet]. ohdsi.github.io. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/>
2. Painter JL, Chalamalasetti VR, Kassekert R, Bate A. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. JAMIA Open. 2024 Dec 26;8(1).
3. SQLGlot • Leading SQL parser, transpiler & optimizer [Internet]. Tobikodata.com. 2026 [cited 2026 Apr 30]. Available from: <https://www.tobikodata.com/sqlglot>
4. fastomop. GitHub - fastomop/FastSSV: Fast Semantic Static Validator [Internet]. GitHub. 2025. Available from: <https://github.com/fastomop/FastSSV>