

PRECISION Evidence Workflow: A Protocol-driven Semantic Infrastructure for Federated and AI-ready Observational Studies in OMOP CDM

Julio Cesar Barbour Oliveira¹, Tulio Sarmiento Rocha Sarmiento¹, Daniela Mayumi Usuda Prado Rocha¹

¹ Precision Data, São Paulo-SP, Brazil

Background

Observational studies using real-world data often rely on extensive and highly complex analytical protocols, in which cohort definitions, indicators, denominators, stratifications, filters, and aggregation rules must be operationalized as reproducible computational queries. Even when employing standardized data structures such as the OMOP Common Data Model, the practical implementation of these analyses can vary substantially across research teams, institutions, and studies, thereby generating challenges for traceability, interoperability, and reproducibility. Methodological initiatives such as STaRT-RWE, LEGEND, and the OHDSI Phenotype Library underscore the importance of structured protocol development, with transparent reporting, pre-specification of analyses, reusable cohort definitions, and scalable network-based evidence generation. In addition, the emergence of artificial intelligence (AI) agents capable of supporting evidence generation increases the requirement to represent analytical metadata in a structured, machine-interpretable, and sufficiently clear to guide the construction of consistent queries. Therefore, the objective of this abstract is to introduce the PRECISION method — **Protocol-driven Reproducible Evidence through Computationally Interoperable Semantic Infrastructure for Observational Networks** — a workflow designed to persist protocol-defined cohorts, metrics, dimensions, filters, and aggregation rules as reusable, machine-actionable semantic metadata, using OMOP CDM as its foundation.^{1, 2, 3, 4}

Methods

The PRECISION workflow uses an analytical metadata layer to explicitly and versionably define the computation rules for indicators of interest, their stratifications by relevant variables, and the cohorts associated with the study. This semantic layer operates as an analytical contract between the study protocol and its computational implementation. Each metric, dimension, cohort, and filter is described with detailed metadata, including its conceptual definition, calculation logic, referenced OMOP CDM fields, and any applicable usage constraints.

The semantic layer is encoded as a YAML file, which is used as input to create and deploy a REST API for access to aggregated data. Analytical use, Precision Data developed the Epiquery library, which allows users, within Python or R environments, to specify the metrics, dimensions, cohorts, and filters of interest. The request is transmitted to the REST API, which interprets the requested components, applies the corresponding rules defined in the semantic layer, and composes a single SQL query that executes the analysis directly on the OMOP CDM database. The implementation was built using the open-source Cube.js framework as the REST API layer. For the REST API deployment considered in this workflow, the minimum infrastructure comprises a single Cube API instance running from the Cube Docker image, access to the semantic data model files, environment-based configuration, authentication through API secrets and JWT, and read-only connectivity to the OMOP CDM database. Cube's architecture supports one or multiple API instances to handle incoming REST API requests, and these instances can be horizontally scaled as needed.⁵

Results

The result is an analytical architecture in which study definitions are no longer dispersed across protocol documents, SQL scripts, computational notebooks, and dashboards, but are instead maintained within a structured semantic layer. This approach increases methodological consistency across analyses derived from the same protocol, reduces reliance on manual SQL rewriting, and allows multiple downstream consumers to access shared analytical concepts through a standardized interface.

The underlying structure was additionally engineered to be agentic-ready. Because the components of the semantic layer include rich metadata and detailed descriptors, an AI agent can be provisioned with this contextual information and, from it, construct Epiqueries that are compatible with the study protocol. Instead of freely generating SQL, the agent operates within a controlled analytical space composed of predefined metrics, dimensions, cohorts, and filters. This promotes greater consistency, governance, and reproducibility of analyses, while also mitigate the risk of divergent interpretations of the study's original conceptual and operational definitions.

The system architecture is also federated by design. The REST API can be deployed within the data owner's own environment, thereby ensuring individual-level data under local control. In this model, the Epiquery library transmits only the desired analytical specification — metrics, dimensions, cohorts, and filters — to the API hosted by the data owner. The REST API then generates and executes the corresponding SQL queries locally on the OMOP CDM database and returns solely aggregated outputs. Consequently, several institutions can provide access to standardized analyses without exposing individual-level data, thereby establishing an infrastructure that is compatible with federated models of evidence generation (figure 1).

Currently, a total of 907 dimensions, 1,126 metrics, and 21 cohorts were defined and organized within the semantic layer structure. These analytical objects span multiple OMOP CDM domains, with the largest concentration of metrics in visit-related and cross-domain objects (table 1). This implementation demonstrates the feasibility of encoding study-relevant analytical definitions as reusable semantic metadata, supporting consistent consumption through Epiquery and the REST API across different observational research applications.

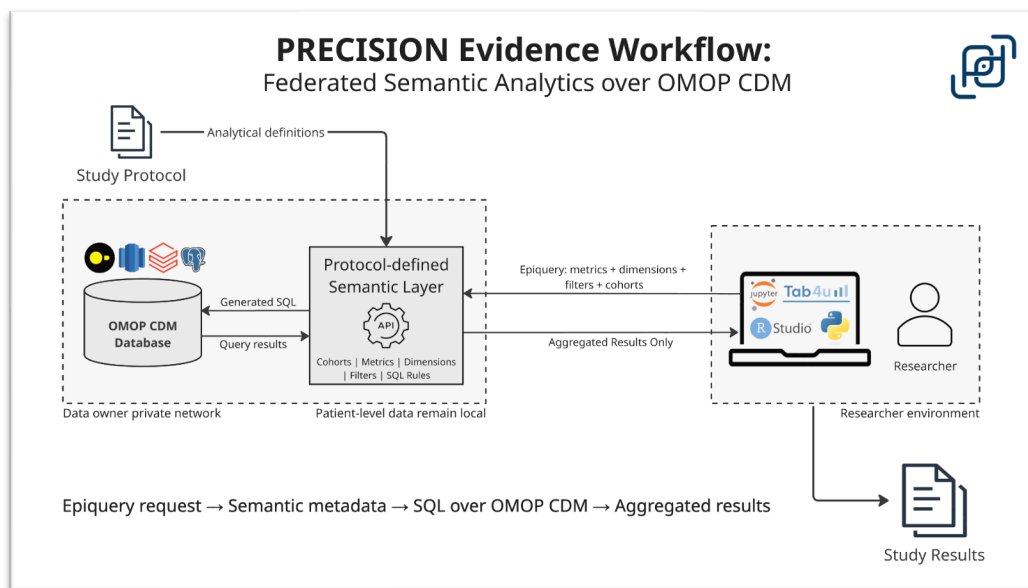


Figure 1. Federated REST API architecture for protocol-driven observational analytics over OMOP CDM: The study protocol is translated into a semantic layer that defines cohorts, metrics, dimensions, filters, and SQL generation rules over an OMOP CDM database. The REST API is deployed within the data owner's private network and receives Epiquery requests from an external Python or R session. Queries are executed locally against the OMOP CDM database, and only aggregated results are returned to the researcher environment, supporting federated, reproducible, and privacy-preserving observational analyses.

Domain	Dimensions	Metrics
Condition	47	8
Device	13	0
Drug	38	8
Measurement	37	0
Procedure	51	25
Observation	65	134
Visit	173	314
Others	483	637
TOTAL	907	1,126

Table 1. Semantic-layer analytical objects created by domain in the PRECISION Evidence Workflow: The table summarizes the number of dimensions and metrics implemented in the protocol-defined semantic layer across OMOP CDM domains. A total of 907 dimensions and 1,126 metrics were created, covering clinical domains such as condition, drug, procedure, observation, visit, device, and measurement, as well as additional cross-domain or derived analytical objects grouped as “Others.”

Conclusion

The PRECISION Evidence Workflow aligns with FAIR principles by making analytical objects more findable through explicit and searchable metadata; more interoperable by structuring definitions on top of the OMOP CDM and making them available through a standardized REST API; and more reusable by enabling metrics, dimensions, and cohorts to be consistently reused across different analyses, studies, and federated environments. Replicability is supported by the clear separation between analytical definition, computational execution, and consumption of aggregated results. By constraining AI agents to predefined metrics, dimensions, cohorts, and filters, the semantic layer provides a controlled and governed environment for generating answers and supporting reproducible data exploration across diverse observational analytics use cases.

A current limitation is that the semantic layer is mainly suited for aggregated, descriptive, and stratified analyses. Complex statistical modeling, advanced adjustment, causal inference, or dedicated algorithms should be implemented upstream, with outputs exposed as flags, derived variables, or pre-computed metrics.

In summary, the PRECISION Evidence Workflow introduces a semantic-layer approach to standardize, operationalize, and federate observational analyses using OMOP CDM. By connecting study protocols, analytical metadata, REST APIs, Epiquery, and AI agents, the methodology establishes an infrastructure oriented toward reproducible, interoperable evidence that can be consumed by both humans and machines. This approach can support collaborative observational research networks by allowing aggregated analyses to be executed consistently across multiple data owners, while preserving local governance and increasing transparency in the analytical process.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

2. Wang S V, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin J A et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies BMJ 2021; 372 :m4856 doi:10.1136/bmj.m4856.
3. Watson, H.J. A Statistical Analysis Plan Template for Observational Studies: Promoting Quality and Rigor in Research. *J Stat Theory Pract* **19**, 91 (2025). <https://doi.org/10.1007/s42519-025-00504-9>.
4. Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, Marc A Suchard, Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND), *Journal of the American Medical Informatics Association*, Volume 27, Issue 8, August 2020, Pages 1331–1337, <https://doi.org/10.1093/jamia/ocaa103>
5. Cube Dev, Inc. Cube Core: open-source semantic layer for AI, BI and embedded analytics [software]. GitHub. Available from: <https://github.com/cube-js/cube>