

Automated Semantic Data Quality Assessment for Federated OMOP Networks: Flagging and Documenting Meaning Differences Before Aggregation

OHDSI 2026 Global Symposium · Collaborator Showcase · Software Demonstration #150

Sandra Estremera-Zink | Founder, Vercori | Sausalito, CA | sandy@vercori.com

Background

Federated OMOP research relies on standardized vocabularies so that data from many institutions can be analyzed together [1]. Standardization supports interoperability, but it does not by itself establish that a concept is represented comparably at every site for a specific study [2–4]. Differences in clinical capture practices, EHR configurations, vocabulary versions, and ETL implementations can produce **semantic divergence**: two or more sites map source data to the same standardized concept identifier while the underlying clinical meaning differs in ways that matter for the study.

A concrete example: heart failure with reduced ejection fraction (HFrEF, OMOP concept 45766164) may be recorded at one institution only when a documented echocardiogram confirms an ejection fraction of 40% or below, while another applies the same term based on clinical judgment alone [5]. Both map to the same OMOP concept. Both can pass the OHDSI Data Quality Dashboard (DQD), because each site's data is internally valid: conformant, complete, and plausible [6]. The divergence is invisible to any check that looks at one site at a time. It becomes visible only through cross-site comparison.

Recent OHDSI work provides reusable study-specific data quality checks, including single- and multi-site analysis [7], and privacy-preserving federated data quality assessment [8]. Vercori addresses a distinct question: whether sites represent the same standardized concept comparably for a given study, assessed at the concept level before aggregation, together with a tamper-evident record of what was found and what the study team decided.

Methods

Vercori is a working prototype that assesses, flags, and documents semantic divergence across distributed OMOP CDM sites. By design, no patient-level data is transmitted between institutions: computation runs locally inside each site's security perimeter, and only aggregate statistical summaries are shared. The workflow has four stages:

1. **Local profiling.** Each site builds a privacy-preserving profile of how it represents each standardized concept across six signals: source codes, diagnostic confirmation, co-occurring conditions, demographics, guideline treatment rate, and referring specialty mix. A key design goal is separating true population differences (for example, one site treats sicker patients) from differences in what a concept means at each site; the multi-signal profile is designed to support that separation.
2. **Cross-site comparison.** Profiles are compared across sites to produce an overall mismatch score for each concept, along with an indication of which signals are driving any mismatch. The comparison is automated and scales with concept set size.
3. **Flag and structured review.** Each concept is either cleared or flagged against prototype thresholds, which are pending validation with network sites. Vercori flags differences; it does not make the study decision. For a flagged concept, a study reviewer chooses one of four actions (Proceed, Change, Exclude, or Consult) and records a rationale. The flagged concept is held from aggregation until a decision is recorded.
4. **Audit trail and documentation.** Every assessment and decision is recorded in a tamper-evident log: what was found, what was decided, when, why, and by whom. Vercori generates a Study Readiness Check report from this log. The report is designed to help sponsors document the data relevance and reliability assessments emphasized in FDA's real-world evidence guidance for medical devices [9], and

to support documentation expectations under the EMA Data Quality Framework and the European Health Data Space.

Results

The software demonstration runs on two simulated sites with synthetic OMOP CDM data. The full pipeline executes in under two minutes for a representative concept set on commodity hardware. Two concepts are assessed in the demonstration (Table 1).

Type 2 diabetes shows a 1% composite mismatch and is cleared. **HFrEF**, the same OMOP concept at both sites, shows a 59% composite mismatch and is flagged for review. The dominant driver is diagnostic confirmation: 95% of HFrEF patients at Site A have a recorded ejection-fraction result, compared with 5% at Site B. The same concept could therefore identify different patients at each site. In the demonstration, the reviewer chooses Change, tightens the concept definition so it is applied consistently at both sites, and records the rationale. The audit trail then shows what was found, what was decided, when, why, and by whom.

Table 1. Synthetic two-site demonstration results (prototype thresholds).

Concept	Composite mismatch	Assessment	Status	Outcome
Type 2 diabetes	1%	Profiles aligned across sites	Cleared	Proceeds to aggregation
HFrEF	59%	Material divergence; diagnostic confirmation dominant	Flagged	Reviewer chose Change; decision and rationale recorded in audit trail

Controlled validation. To assess detection against known ground truth, five distinct divergence patterns were injected into the Eunomia GiBleed reference dataset on OMOP CDM 5.3, including a population-only difference that should not be flagged. Vercori correctly identified all five patterns by flag status, classification, and attribution of the driving signal. This shows the detection logic behaves as designed against injected divergence with known ground truth. It does not, by itself, establish performance on unlabeled real-world data; real-world validation is the subject of planned pilot work.

Conclusions

Vercori is designed as an upstream complement to DQD, Achilles, and federated analytics frameworks. DQD checks each site on its own; Vercori compares how sites represent the same concept before data are combined. Because phenotypes are built from concepts, a concept that diverges across sites can cause the same phenotype to find different patients at each site. Vercori addresses two gaps: concept-level assessment of cross-site semantic comparability, and a tamper-evident record of what divergence was found, how it was classified, and what the reviewer decided and why.

Because it operates in a federated architecture in which, by design, no patient-level data leaves a site, Vercori is compatible with existing OHDSI network governance structures and institutional data use agreements. Vercori is designed to complement existing open-source OHDSI tools, not replace them. The central open question is how study-specific divergence thresholds should be established and validated, which requires validation across multiple real OMOP implementations and is the primary objective of planned pilot partnerships.

Intellectual Property and Collaboration

The core detection method is patent pending. This report describes Vercori at a functional level. Implementation details, including the statistical methods, scoring weights, threshold parameters, and audit architecture, are available to prospective pilot sites and methodological collaborators under a mutual non-disclosure agreement.

The demonstration uses synthetic data, and we are seeking OHDSI network sites and methodological collaborators to validate Vercori on real data. Live demonstration is available at the symposium. To discuss collaboration, contact sandy@vercori.com.

References

- [1] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574-8. PMID: 26262116.
- [2] Dymshyts D, DeFalco F, Ostroplets A, Rao G, Shoaibi A, Blacketer C. Evaluating the impact of different vocabulary versions on cohort definitions and CDM. Collaborator Showcase #23. OHDSI Global Symposium; 2024 Oct 22-24; New Brunswick, NJ.
- [3] Lee H, Gupta S, Blacketer C, Cook M, Naka S, Fan R, et al. Comparative analysis of OMOP CDM database profiles across institutions and future research implications. Collaborator Showcase #49. OHDSI Global Symposium; 2024 Oct 22-24; New Brunswick, NJ.
- [4] Dickinson K, Wieand K, Bailey C, Razzaghi H. A systematic and sustainable solution for assessing network data quality. Collaborator Showcase #42. OHDSI Global Symposium; 2024 Oct 22-24; New Brunswick, NJ.
- [5] Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J Am Coll Cardiol*. 2022;79(17):e263-e421. doi:10.1016/j.jacc.2021.12.012.
- [6] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4(1):1244. doi:10.13063/2327-9214.1244.
- [7] Wieand K, Razzaghi H, Dickinson K, Kahn M, Roy J, Bailey C. Executing a reusable framework for study-specific data quality analysis. Collaborator Showcase #128. OHDSI Global Symposium; 2024 Oct 22-24; New Brunswick, NJ.
- [8] Tomášik R, Kussel T, Dudová Z, Kacová R, Hrstka R, Lablans M, et al. Privacy-preserving data quality assessment for federated health data networks. *BMC Med Inform Decis Mak*. 2026;26:49. doi:10.1186/s12911-025-03328-6.
- [9] U.S. Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and Food and Drug Administration staff. Silver Spring (MD): FDA; 2025 Dec 18. Docket No. FDA-2023-D-4395.