

Making the Most of What We Have: Regional Characterization of OHDSI Network Databases

Evelyn C. Goh¹; Jack Janetzki²; Nicole Pratt²; Feng Mengling³

¹ College of Design and Engineering, National University of Singapore, Singapore ² UniSA Clinical and Health Sciences, University of South Australia, Adelaide, Australia ³ Saw Swee Hock School of Public Health, National University of Singapore, Singapore

Background

Distributed network studies in the OHDSI community rely on three main diagnostic metrics: absolute standard difference of the mean (ASDM), empirical equipoise, and EASE^{1, 2}. These metrics form the basis for evaluating whether a given database contributes interpretable effect estimates. In a previously conducted study, non-North American (N. Am) databases, particularly from Asia-Pacific (APAC) and Europe, were seen to have disproportionately failed diagnostics in an OHDSI multi-region study⁴.

As most of the non-U.S. databases passed EASE but failed on one or both of ASDM or empirical equipoise, structural covariate sparsity became a leading hypothesis. Low prevalence of FeatureExtraction covariates in non-N. Am databases produces numerically unstable ASDM estimates post-matching, collapses preference score distributions, and inflates EASE. If the pre-existing diagnostic framework, developed based on North American/U.S. databases, are systematically excluding non-N. Am databases, an adapted framework may need to be developed to encourage additional participation in non-N. Am/regional databases. As such, this study explores the characterization of such databases to understand patterns that may be helpful in diagnosing reasons for systematic failure of objective diagnostics.

Methods

For this report, only covariates measured from previously executed network studies were evaluated. We analysed cohort characterization outputs from two independent OHDSI APAC studies: (1) the 2024 study evaluating the association between use of GLP-1 receptor agonists and acute liver injury⁴; and (2) an earlier study studying the association between use of fluoroquinolone-class antibiotics and abdominal aortic aneurysms (AAA)⁵.

In the 2024 GLP-1 study, fifteen databases across four regions contributed data: North America (n=9), Australia/New Zealand (n=1), Asia-Pacific (n=3), and Europe (n=2)⁴. Total analysed population was 9,118,492 individuals⁴. In the fluoroquinolone study, nine databases across two regions contributed data: North America (n=8) and Asia-Pacific (n=1)⁵. Total analysed population was 1,734,913 individuals⁵.

For the studies, FeatureExtraction² was used to generate covariate sets per cohort across short-term (30-day) and long-term (365-day) pre-index windows¹. Covariates falling below the assigned subject count threshold were recorded as -1; those absent from a database were recorded as missing.

For each covariate, a regional sparsity ratio was computed as the proportion of databases within each region returning a -1 or null value. Each covariate was thus represented as a four-dimensional vector [ratio_{N.Am}, ratio_{ANZ}, ratio_{APAC}, ratio_{EU}]. K-means clustering (k=5, scikit-learn⁶, random_state=42) was applied independently to each cohort's covariate set. An elbow analysis confirmed k=5 as appropriate. Cluster profiles were characterized by mean regional sparsity ratio and dominant covariate domain via regex-based category mapping. Cross-study pattern concordance was assessed by comparing cluster profiles across both studies.

Results

For the GLP-1 study, each cohort yielded approximately 20,000 covariates after FeatureExtraction. K-means clustering identified five clusters per cohort. Three patterns were consistent across all cohorts in both studies (Table 1).

- (1) Pattern 1 demonstrated universal sparsity. A large fraction of covariates (dominant cluster by size in all cohorts) showed high -1 ratios across all four regions. These were predominantly observation-period and rare condition-era covariates.
- (2) Pattern 2 demonstrated North America preferential availability. Up to 20% of covariates were available in N.Am databases but recorded as -1 or absent elsewhere. Drug era group and condition era group covariates dominated this cluster, consistent with the breadth of outpatient prescription and ambulatory claims data available in N.Am administrative datasets but not in hospital-centric or regional registry systems.
- (3) Pattern 3 showed APAC-enriched availability. A distinct cluster showed near-complete missingness in N.Am, ANZ, and EU but non-trivial availability in APAC databases, predominantly JMDC (Japan). These covariates were dominated by measurement and procedure concepts, consistent with structured hospital billing data in Japanese administrative claims⁷.

Cohort-specific patterns also emerged: the ALI cohort showed an additional N.Am-preferential cluster distinct from Pattern 2; the GLP-1 cohort had a Europe-specific cluster (~4.9% of covariates available predominantly in French and German databases); and the DPP-4 cohort showed an Australia-New Zealand (ANZ)-specific cluster (~4.8%).

The three cross-study patterns identified in the GLP1 analysis are broadly replicated in the FQ/AAA results with one structural caveat.

- (1) Pattern 1 (universal sparsity across all databases) is clearly present. The dominant cluster in every FQ/AAA cohort contains covariates with high NaN ratios across all nine databases, consistent with covariates that are globally rare or systematically uncaptured in OMOP-mapped data.
- (2) Pattern 2 (available in N.Am but not non-N.Am) is directionally supported. Cluster profiles showed subsets of condition and drug era covariates with low NaN rates across N.Am databases but complete missingness in JMDC, mirroring the N.Am-preferential pattern seen in GLP1.
- (3) Pattern 3 (available in APAC but sparse elsewhere) is also detectable. Measurement covariates such as LDH, CRP, and blood pressure readings appeared in Japan's JMDC with meaningful prevalence but were absent in several N.Am databases, consistent with JMDC's structured inpatient billing data capture⁷.

The key limitation is that the FQ/AAA dataset includes only one non-N.Am database (JMDC), with no ANZ or European sites, so the GLP1 study's study-specific EU and ANZ niche clusters cannot be assessed. Pattern 3 cannot be distinguished from general JMDC structural differences rather than a broader APAC phenomenon.

Table 1. Summary of cross-study covariate sparsity patterns identified by k-means clustering

Pattern	Prevalence	Dominant domain	Consistent across 2 studies?
1 - Universal sparsity	Largest cluster (30–40% of covariates)	Observation period, rare conditions	Yes
2 - N.Am preferential	5–20% of covariates	Drug era group, condition era group	Yes
3 - APAC enriched	Moderate cluster (10–15%)	Measurements, procedures (JMDC)	Yes
Study-specific (EU/ANZ/N.Am niches)	4.8–15%	Varies by study drug / region	No - cohort-dependent

Conclusion

Covariate sparsity in OHDSI network studies follows reproducible structural patterns that are consistent across independently conducted studies with different study drugs, outcomes, and cohort definitions. Three patterns (universal sparsity, N.Am preferential availability, and APAC-enriched coverage) were identified across all cohorts in both the GLP-1/DILI and FQ/AAA analyses. The consistent emergence of study-specific regional niches (EU, ANZ) suggests that individual study populations further modulate regional availability beyond the structural baseline. These findings have possible direct implications for OHDSI study design.

Future work will extend this characterization to further studies, as well as conduct formal analyses of all aspects of the diagnostic framework, including for EASE and the availability of data for standard negative control outcomes. These results are intended to support all regions to contribute meaningfully to meta-analyses of distributed network studies.

References

1. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans A Math Phys Eng Sci.* 2018;376(2128):20170356. doi:10.1098/rsta.2017.0356
2. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul.* 2013;23(1):10. doi:10.1145/2414416.2414421
3. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574–8. PMID:26262116
4. Goh E, You SC, Janetzki JL, Asao K, Dorajoo S, Pratt N, Ryan PB, Schuemie M, Suchard MA, Van Zandt M, Mengling F. Use of GLP-1 Receptor Agonists and Risk of Acute Liver Injury: A Cohort Analysis in the OMOP CDM (GLP1-DILI) [poster]. OHDSI Global Symposium; 2025; [place unknown]. National University of Singapore.
5. Janetzki JL, Kim JH, Minty E, Lee JA, Morales DR, Khera R, Kim C, Alshammari TM, DuVall SL, Matheny ME, Falconer T, Kim S, Phan TP, Nguyen PA, Hsu MH, Hsu JC, Park RW, Man KKC, Seager S, Van Zandt M, Gilbert JP, Ryan PB, Schuemie MJ, Suchard MA, Hripcsak G, Pratt N, Chan You S. Risk of aortic

aneurysm or dissection following use of fluoroquinolones: a retrospective multinational network cohort study. *eClinicalMedicine*. 2025 Feb 1;81:103096. doi: 10.1016/j.eclinm.2025.103096.

6. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
7. Kimura S, Sato T, Ikeda S, Noda M, Nakayama T. Development of a database of health insurance claims: standardization of disease classifications and anonymous record linkage. *J Epidemiol*. 2010;20(5):413–9. doi:10.2188/jea.JE20090066
8. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol*. 2018;47(6):2005–14. doi:10.1093/ije/dyy120
9. Pasternak B, Inghammar M, Svanstrom H. Fluoroquinolone use and risk of aortic aneurysm and dissection: nationwide cohort study. *BMJ*. 2018;360:k678. doi:10.1136/bmj.k678