Comparative Analysis of OMOP CDM Database Profiles Across Institutions and Future Research Implications

Haeun Lee[1], Snehil Gupta[2], Clair Blacketer[3], Michael Cook[1], Shinji Naka[2], Ruochong Fan[2], Benjamin Martin[1], Khyzer Aziz[1], Linying Zhang[2], Paul Nagy[1]

[1] Department of Biomedical Informatics and Data Science, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, Maryland
[2] Institute for Informatics, Data Science and Biostatistics, Washington University in St. Louis, St. Louis, Missouri
[3] Janssen Research and Development, Titusville, New Jersey

Background

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has significantly advanced clinical research by enabling the harmonization and interoperability of federated healthcare data across diverse institutions [1-4]. However, determining the eligibility of institutions for studies in a federated network addressing specific clinical questions remains challenging without detailed insights into the quality and completeness of their data. The database diagnostic (dbDiagnostic) profile addresses this need by creating detailed profiles of individual databases converted to the OMOP CDM, including aggregated summary statistics and data quality results [5]. Sites can expose a profile of their data without putting the row level patient health information at risk. Profiling can be used to assess whether a database has the necessary elements to answer specific clinical questions, ensuring the reliability and validity of multicenter studies [6]. By examining the database profile results, researchers can identify discrepancies and similarities between datasets, enhancing the robustness of clinical research. The profiles provide indicators of a database's readiness for specific analyses, helping researchers assess its suitability for studies [5]. This study aims to examine the database profile results from two institutions, identify the similarities and differences, and explain their implications for future research within the OHDSI community.

Methods

This study used DbDiagnostics, the OMOP CDM database diagnostics utility R package (v. 1.3.1) from the OHDSI software suite to compare OMOP CDM databases from two large tertiary hospitals: Johns Hopkins School of Medicine and Washington University School of Medicine in St. Louis [5]. The SITE A dataset included EHR data converted to OMOP CDM from December 2016 to June 2023, while the SITE B dataset covered data from January 2000 to March 2024. A comprehensive evaluation of the OMOP CDM datasets was conducted using Achilles (v1.7.2) and the Data Quality Dashboard (v2.6.0) [5]. The evaluation involved generating summary statistics for various CDM domains, including persons, visits, conditions, procedures,

measurements, medications, observations, and device exposures, to assess the overall data quality and consistency. The comparison focused on common concepts, data density, data completeness, and data quality to provide insights into the characteristics and reliability of the datasets.

Results

Our study analyzed data from 2.1 million patients at Site A and 8 million patients at Site B. The diagnostic profiles provided a comprehensive view of patient demographics, including age, sex, race, and ethnicity. We identified data quality issues in age distribution and found gaps in sex, race, and ethnicity, with differences in the granularity of race between the institutions. The visit table enabled us to pinpoint key visit types essential for clinical research, such as inpatient (9201), outpatient (9202), and emergency room visits (9203) (Table 1). Overall, the condition concept IDs followed SNOMED CT codes at both institutions. However, differences were observed in the vocabulary for common procedures such as total knee replacement, cholecystectomy, and laparoscopic appendectomy between the institutions. For measurements, Site A used body mass index (BMI) [ratio], while Site B used BMI [percentile]. Additionally, Site A and Site B used different vocabularies for respiratory rate and different concepts for heart rate (Table 1). Both institutions had essential domains such as person, visit occurrence, condition occurrence, procedure occurrence, drug exposure and measurement (Table 2). Care site information was not present in the Site A OMOP CDM due to its intentional exclusion from the specific OMOP instance analyzed by the dbDiagnostic package. The condition concept IDs showed similar trends and higher frequencies for chronic diseases like hypertension, hyperlipidemia, and type 2 diabetes across institutions. However, differences were found in the top five measurement concepts. Site A recorded more vital signs, while Site B focused on lab values and body weight. For medications, both institutions had high frequencies of ondansetron 2mg/mL injections. However, Site A recorded more entries for potassium chloride and COVID-19 vaccines, while Site B more frequently included various doses of sodium chloride (Table 3). By examining overlapping patient data across various domains, we identified gaps in data density and areas needing improvement for study conduct (Figure 1).

Conclusion

The comparative analysis of the database profiles for OMOP CDM datasets from Site A and Site B demonstrated the crucial role of database profiling in proactively evaluating data quality and completeness for multicenter studies. This study highlighted both strengths and limitations in these datasets, offering significant insights into patient demographics, vocabularies, and common concepts across various domains. We identified discrepancies, particularly in the vocabularies and concepts used in procedures and measurements. For future research, addressing these discrepancies is expected to provide clearer insights into each database's characteristics, support more effective network studies, and facilitate collaborative research within the OHDSI community.

Table 1. Summary of Common Concepts Across Two Institutions

Concept Category	Site A (Concept ID)	Site B (Concept ID)
Race		
White	8527	8527
Black or African American	8516	8516
Asian	8515	8515
American Indian or Alaska Native	8657	8657
Other Pacific Islander	38003613	38003613
Visit		
Inpatient	9201	9201
Outpatient	9202	9202
Emergency room visit	9203	9203
Intensive care visit	32037	32037
Condition		
Essential hypertension	320128	320128
Type 2 diabetes mellitus without complication	4193704	4193704
Obesity	433736	433736
Gastroesophageal reflux disease without esophagitis	4144111	4144111
Procedure		
Cesarean section	2110316	4015701
Total knee replacement	2105103	43531648
Cholecystectomy	2109368	4242997
Laparoscopic Appendectomy	2109144	4243973
Measurement		
Body height	3036277	3036277
Body weight	3025315	3025315
Body mass index	40762636	3038553
Body temperature	3020891	3020891
Systolic blood pressure	3004249	3004249
Diastolic blood pressure	3012888	3012888
Respiratory rate	3024171	4313591
Heart rate (Pulse rate)	3027018	4301868

Table 2. Key Data Domains Across Institutions

Domain	Site A	Site B
Person	Y	Y
Visit Occurrence	Y	Y
Condition Occurrence	Y	Y
Procedure Occurrence	Y	Y
Measurement	Y	Y
Drug Exposure	Y	Y
Observation	Y	Y
Care Site	N	Y
Device Exposure	Y	N

Table 3. Top 5 Most Common Concepts by Domain and Concept Ids Across Institutions

Domain	Site A	Site B
Condition	Essential hypertension (320128)	Patient encounter procedure (4203722)
	Hyperlipidemia (432867)	History of event (1340204)
	Type2 diabetes mellitus without complication (4193704)	Essential hypertension (320128)
	Gastroesophageal reflux disease without esophagitis (4144111)	Hyperlipidemia (432867)
	Postoperative state (438485)	Type2 diabetes mellitus without complication (4193704)
Measurement	Heart rate (3027018)	Hemoglobin [Mass/volume] in Blood (3000963)
	Diastolic blood pressure (3012888)	Body weight (3025315)
	Systolic blood pressure (3004249)	Platelets [#/volume] in Blood by Automated count (3024929)
	Respiratory rate (3024171)	No matching concept (0)
	Oxygen saturation in Arterial blood by Pulse oximetry (40762499)	Respiratory rate (4313591)
Medication	Potassium chloride 0.004 MEQ (19135374)	No matching concept (0)
	SARS-COV-2 (COVID-19) vaccine (724906)	1000 ML sodium chloride 9 MG/ML Injection (40220357)
	SARS-COV-2 (COVID-19) vaccine (724907)	10 ML sodium chloride 9 MG/ML Prefilled Syringe (19127213)
	oxycodone hydrochloride 5 MG Oral Tablet (40232756)	100 ML sodium chloride 9 MG/ML Injection (40221385)

2 ML ondansetron 2 MG/ ML Injection (35605482)	2 ML ondansetron 2 MG/ML Injection (35605482)

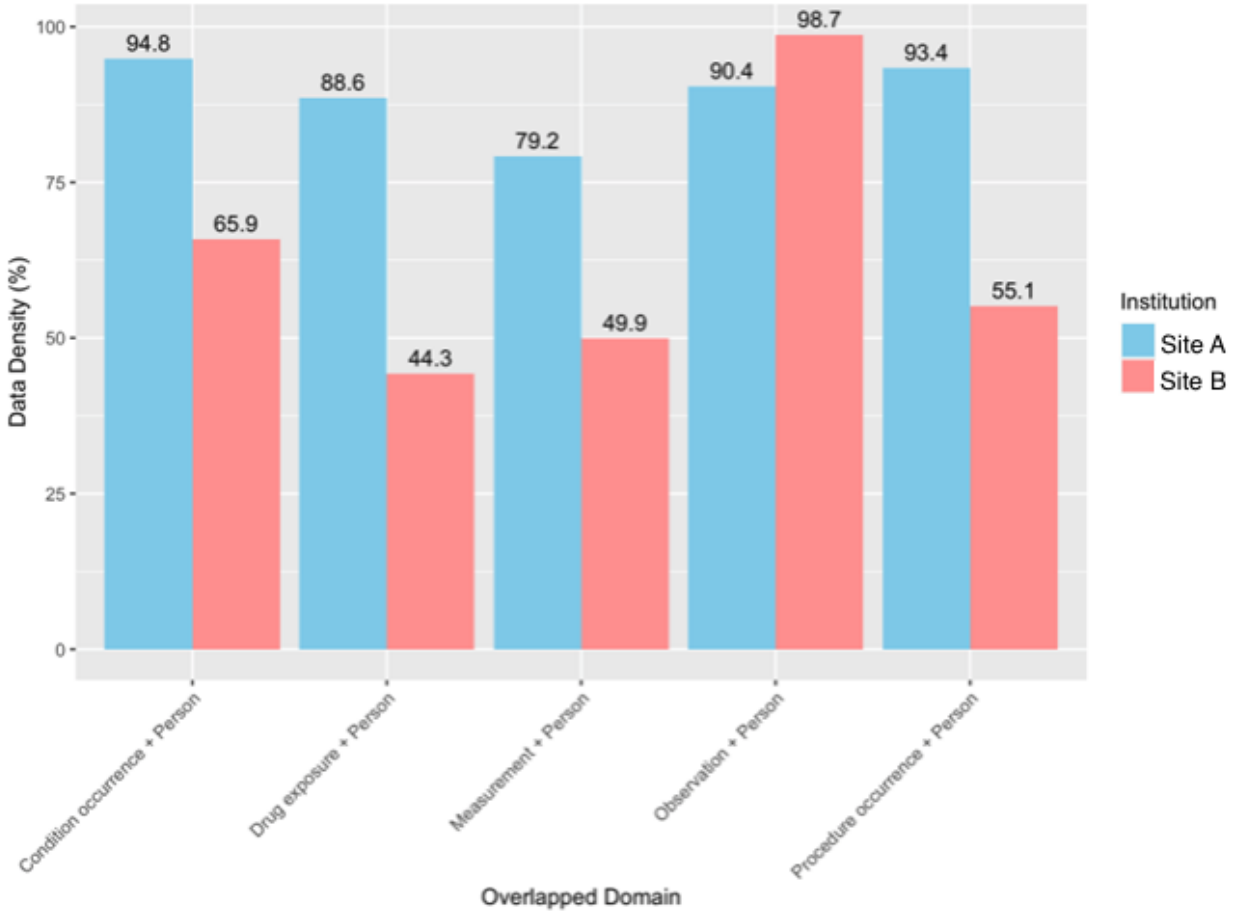

Figure 1. Comparison of Data Density for Different Domains Between Institutions

References

1. Chai Y, Man KK, Luo H, et al. Incidence of mental health diagnoses during the COVID-19 pandemic: A multinational network study. *Epidemiology and Psychiatric Sciences*. 2024;33. doi:10.1017/s2045796024000088
2. Naderalvojoud B, Curtin CM, Yanover C, et al. Towards global model generalizability: Independent cross-site feature evaluation for patient-level risk prediction models using the OHDSI network. *Journal of the American Medical Informatics Association*. 2024;31(5):1051-1061. doi:10.1093/jamia/ocae028
3. Reich C, Ostropolets A, Ryan P, et al. OHDSI standardized vocabularies—a large-scale centralized reference ontology for International Data Harmonization. *Journal of theAmerican Medical Informatics Association*. 2024;31(3):583-590. doi:10.1093/jamia/ocad247
4. Klann JG, Joss MA, Embree K, Murphy SN. Data Model Harmonization for the all of us research program: Transforming I2B2 data into the OMOP Common Data Model. *PLOS ONE*. 2019;14(2). doi:10.1371/journal.pone.0212463
5. Blacketer C, DeFalco F (2024). *DbDiagnostics: OMOP CDM Database Diagnostics Utility*. R package version 1.3.1.
6. de Ridder MA, de Wilde M, de Ben C, et al. Data Resource Profile: The Integrated Primary Care Information (IPCI) database, the Netherlands. *International Journal of Epidemiology*. 2022;51(6). doi:10.1093/ije/dyac026