Name:	**Frank DeFalco, Jill Hardin, Laura Hester**
Affiliation:	**Observational Health Data Sciences and Informatics (OHDSI), New York, NY; Janssen Research and Development, Raritan, NJ**
Email:	fdefalco@its.jnj.com
Presentation type (select one):	Poster

Dataprint: A Novel Visualization Tool for Database Comparison

Frank DeFalco[1,2], Jill Hardin, MS, PhD[1,2], Laura Hester, PhD[1,2]

[1]Observational Health Data Sciences and Informatics (OHDSI), New York, NY; [2]Janssen Research and Development, Raritan, NJ

Abstract *[Poster-100-200 words]*

Information on the heterogeneity in a single database and across a data network is critical for selecting the database best suited to address a research question. Visualizations provide a rapid, comprehensive method for capturing patterns of data heterogeneity that may be difficult to recognize in complex, tabular summaries. This study developed a novel data visualization technique called a dataprint, to illustrate distributions of patient characteristics across 10 domains and 11 observational databases. Dataprint's contribution is to increase data transparency and be a part of the evidence to inform decisions about which dataset is most useful for conducting a study.

Introduction:

OHDSI is a distributed data sharing and research network which aggregates large, international clinical databases to aid multisite observational studies for generating evidence that improves health care.[1,2] The databases available in the network differ in terms of patient representation, data capture process, and availability of longitudinal information. These differences give each database a unique set of characteristics or fingerprints, which we call a "dataprint." Effective use of a data network requires researchers to understand the dataprint of available databases, specifically the heterogeneity of the patient populations, which can inform decisions about which dataset is best suited to inform a research topic. Visualizations provide a facile method for capturing data heterogeneity in comparison to numerical tables. This study sought to develop a novel dataprint visualization to characterize data from 10 tables in the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) across 11 OHDSI databases to facilitate data transparency and to provide evidence for dataset selection for conducting a study.

Methods:

The dataprint visualization utilized 11 data sets: Truven MarketScan Commercial Claims and Encounters (CCAE), Medicare (MDCR), Medicaid (MDCD); OptumInsight SES; Premier; UK's Clinical Practice Reseach Database (CPRD); IMS France, Australia, Germany; Japan Medical Data Center (JMDC); and the National Health and Nutrition Examination Survey (NHANES). Ten tables from the OMOP CDM (version 5.0.1) were queried (condition, death, device_exposure, drug_exposure, measurement, notes, observation_period, observation, procedure, and visit_occurrence) for the 11 databases. R version 3.4.1 was used to generate JavaScript Object Notation (JSON) files, which summarized the number of records in 5-year age and sex strata. The JSON files were used as input for an html file that generated comparative, vertically mirrored continuous distribution functions for each CDM table based on the age-sex strata. The y-axis is the number of records, and the x-axis is the age at the record. The yellow curve below the null line reflects the distribution of record counts for females, and the blue curve above the null line reflects the distribution for males. The 25th, 50th, and 75th quantiles were calculated from continuous distribution functions based on weighted, binned data and are displayed as solid white circles. These plots were reproduced for the 11 databases.

Results:

Figure 1 displays a legend for a dataprint. The database name is displayed to the left of the plot, the lines below it display the total number of unique subjects and records contained in the database. The x-axis represents age at the record, with solid lines at each 10-year increment. The y-axis illustrates the number of records by sex (female - yellow male – blue). The 25th, 50th, and 75th quantiles, represented by the solid white circles. The right vertical presents the records per person, the percentage of people contributing records, and the percentage of the records across all records in a given database. The Truven CCAE condition dataprint shows that there are 134 million unique subjects in the database and 26.8 billion records in total. There are

4.18 billion condition records, which represent 15.6% of all records in the database. The condition records are available for subjects aged 0-70. The 25th, 50th, and 75th quantiles fall at ages 29, 45, and 56, respectively. The percentage to the right of the plots displays there are 40 records per person and 78% of people contribute records, and lastly that the percentage of condition records out of all records in the Truven CCAE database (15.6%).

Figure 2 displays the dataprints for the 10 tables across the 11 datasets. The dataprints display notable patterns. The Truven CCAE dataset has minimal data for patients aged 65+, while the Truven MDCR largely contains data for patients aged 65+. The majority of records for visits, conditions, drug exposures, and procedures in the Truven MDCD data occur among children and women, while deaths are more likely among men aged ≥40 years. Most procedure records in the IMS Germany data occur among children and women of child-bearing ages. Approximately half of IMS France records are drug exposures, while the majority of JMDC records are procedures. Electronic health record data (CPRD, Premier) have a lower percentage of records classified as conditions compared to claims-based datasets.

Figure 1. Dataprint legend.

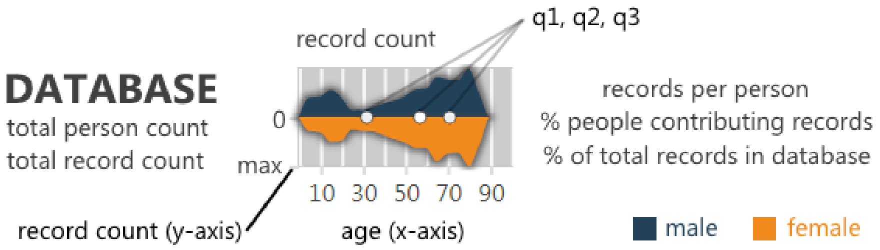
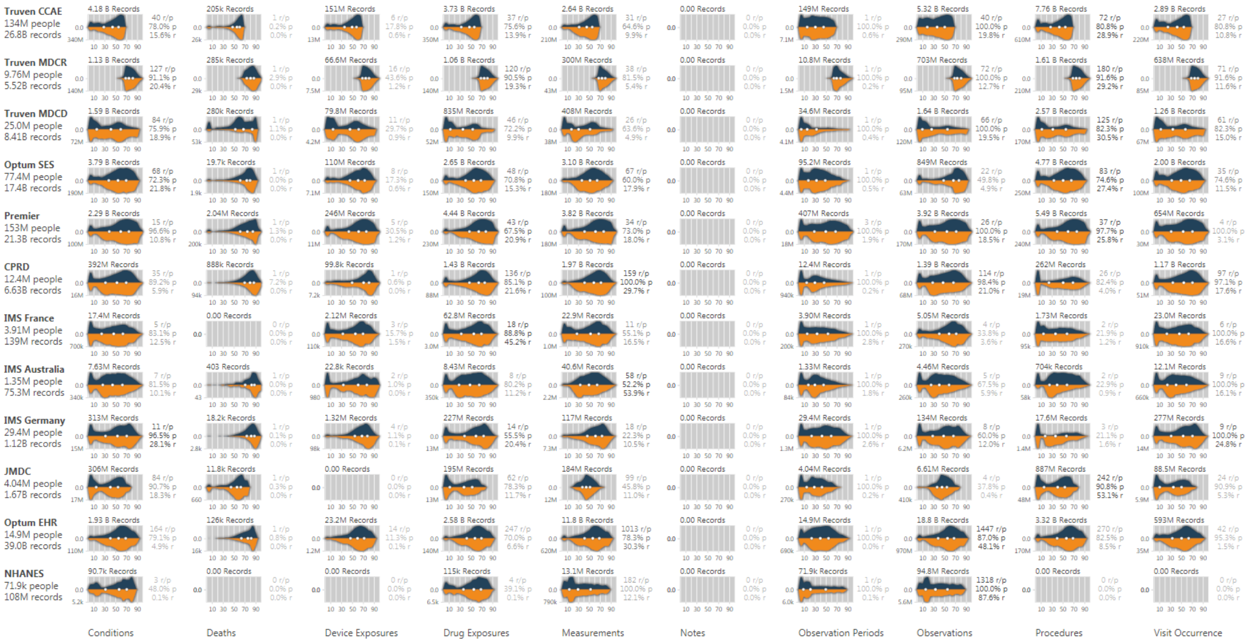


Figure 2. Dataprint showing 10 CDM domain tables across 11 databases available in the OHDSI network.



Conclusion

This study illustrates how dataprints can provide a quick summary of database characteristics in a large data network. Researchers can use this visualization tool to decide which databases have the most appropriate characteristics for studying a research question. This visualization can be made available to the OHDSI community to inform study design and facilitate data transparency across all OHDSI data assets.

References

1. Huser, V, DeFalco, FJ, Schuemie, M, Ryan, PB, Shang, N, Velez, M, et al. Multisite evaluation of a data quality tool for patient-level clinical datasets. eGEMs. 2016;4(1).
2. Kahn, MG, Brown, JS, Chun, AT, Davidson, BN, Meeker, D, Ryan, PB, et al. Transparent reporting of data quality in distributed data networks. eGEMs. 2015; 3(1):1052.