Dataprint: A Novel Visualization Tool for Database Comparison

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

[1]Janssen Research & Development, LLC, Titusville, NJ [2]OHDSI collaborators, Observational Health Data Sciences and Informatics (OHDSI)

ABSTRACT

This study developed a dataprint to provide a method for visualizing differences in characteristics between databases in the Observational Health Data Sciences and Informatics (OHDSI) distributed data network. The dataprint displays the probability density function of record counts by sex and age at observation for each Observational Medical Outcomes Partnership (OMOP) common data model (CDM) table. Other descriptive information about the records in each table are also displayed, including the quantiles, records per person, and percentage of overall records. Dataprints help quickly summarize database characteristics.

BACKGROUND

- Databases differ by patient representation, data capture process, and availability of longitudinal information.
- Differences give each database a unique set of characteristics, which we call a "dataprint."
- Dataprints can facilitate data transparency and inform decisions about which database(s) provide the best evidence for a study

OBJECTIVE

- Illustrate domain specific distributions of patient characteristics in a CDM v5 database.
- Increase data transparency and evidence to inform decisions about CDMs and their applications.
- Enable broad data network visualization in addition to individual CDM insights.

METHODS

- Developed a minutia (figure 1) visualization incorporating a pyramid plot and additional basic descriptive statistics.
- Reproduced minutia for 11 databases:
  - Truven MarketScan's Commercial Claims & Encounters (CCAE), Medicare (MDCR) and Medicaid (MDCD)
  - OptumInsight SES
  - Premier
  - UK's Clinical Practice Reseach Database (CPRD)
  - IMS France, Australia, Germany
  - Japan Medical Data Center (JMDC)
  - National Health & Nutrition Examination Survey (NHANES)
- For each database, extracted number of records and individuals with data from every domain represented across 10 OMOP CDM (v5.0.1) tables
- Developed an R script to query each CDM database and generated JavaScript Object Notation (JSON) files summarizing number of records in 5-year age and sex strata.
- Developed an HTML based visualization leveraging d3js to render data using minutia (figure 1) to represent the overall data network dataprint (figure 2)

RESULTS

A minutia visualization is produced for each domain in each database. When displayed in a full row this forms the database's dataprint. The rows are then stacked to form a complete network visualization.

In interpreting the Conditions domain of the Truven CCAE database, there are:

- 4.18 billion condition records, which represents 15.6% of the 26.8 billion records for the 134 million unique subjects in the database
- 40 records per person and 78% of people contribute records
- 25th, 50th, and 75th quantiles at ages 29, 45, and 56, respectively

General patterns observed in 10 tables across the 11 datasets include:

- Populations of varied ages are clearly depicted across the dataprint (figure 2)
- In Truven MDCD, the majority of records for visits, conditions, drug exposures, and procedures occur among children and women. Deaths are more likely for men aged ≥40
- Most procedure records in IMS Germany occur among children and women of child-bearing age
- 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

MINUTIA

In biometrics and forensic science, minutiae are major features of a fingerprint, using which comparisons of one print with another can be made. This theme has been adapted for our dataprint visualization in the form of a tornado or pyramid plot in combination with other descriptive measures.

- Y-axis represents age at the record, with solid lines at each 10-year increment
- X-axis illustrates the number of records by sex (female - yellow male – blue)
- Solid white rectangles represent 25th, 50th, and 75th quantiles, respectively
- Additional statistics are presented to the right of each minutia (figure 1)

Figure 1. Introducing the Minutia

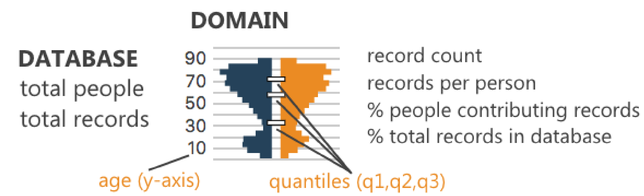
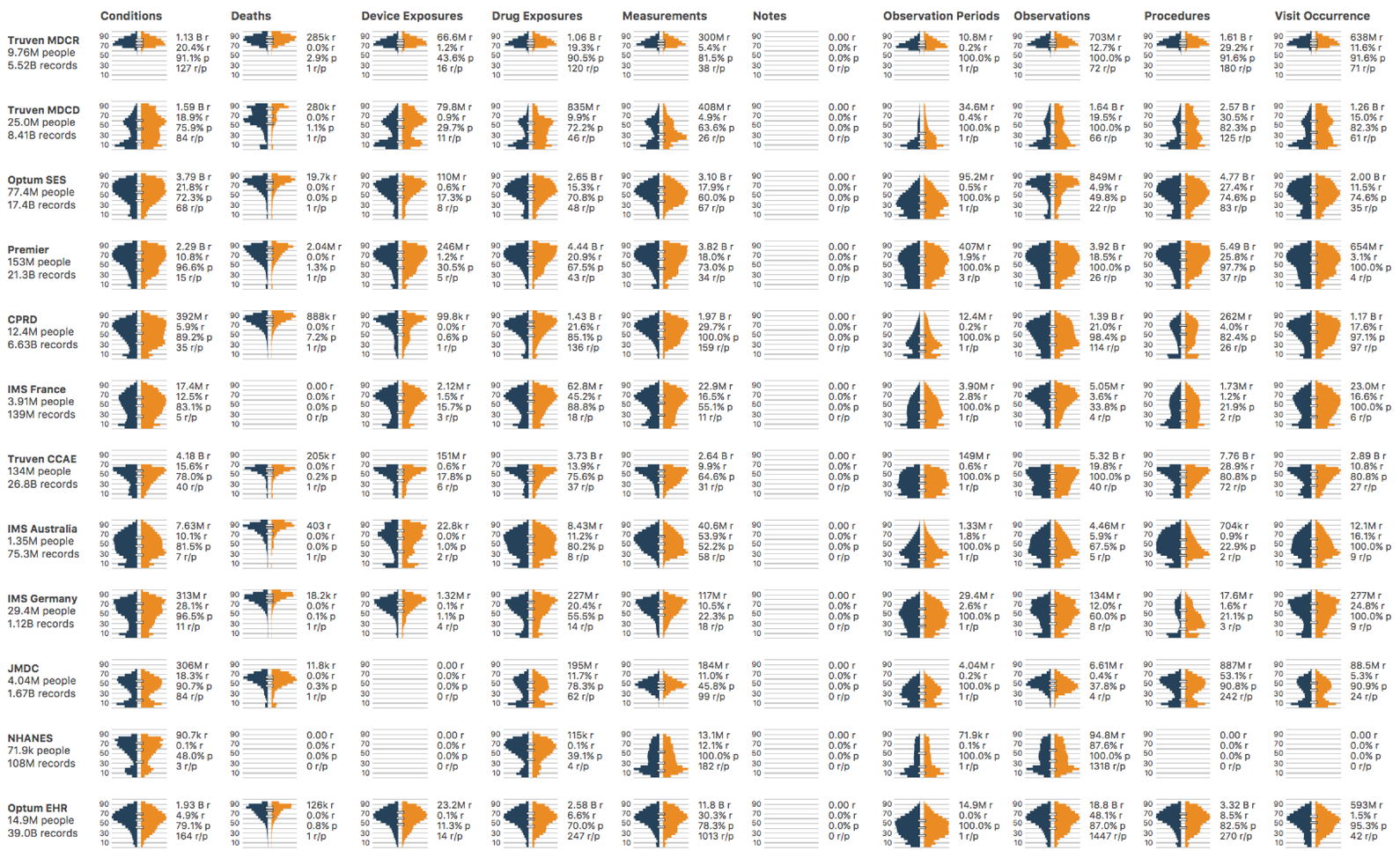


Figure 2. Dataprint showing 10 CDM domain tables across 11 databases available in the OHDSI network



CONCLUSIONS

- This study illustrates how dataprints can provide a quick summary of database characteristics in a large data network.
- This visualization can be made available to the OHDSI community to inform study design and facilitate data transparency across all OHDSI data assets.

NEXT STEPS

- Base data extract on Achilles / FeatureExtraction
- Share Dataprint repository to OHDSI community
- Apply Dataprint to cohort specific characterization

CONFLICT OF INTEREST STATEMENT

All authors are full time employees of Janssen Research and Development, a unit of Johnson and Johnson. The work on this study was part of their employment. They also hold pension rights from the company and own stock and stock options.

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