Feasibility of Converting the Medicare Synthetic Public Use Data Into a Standardized Data Model for Clinical Research Informatics

Mark D. Danese, MHS, PhD[1]; Erica A. Voss, MPH[2]; Jennifer Duryea, MPH[1]; Michelle Gleeson, PhD[1]; Ryan Duryea[1]; Amy Matcho[2]; Donald O'Hara, MS[3]; William E. Stephens[4]; Adler Perotte, MD, MA[5]; Lee Evans[6]; Christian Reich, MD, PhD[7]

[1]Outcomes Insights, Westlake Village, CA; [2]Janssen Research and Development, Raritan, NJ; [3] ComputerAid, Allentown, PA; [4] Paxata, Redwood City, CA; [5] Columbia University, New York, NY; [6] LTS Computing, West Chester, PA; [7] IMS Health, Burlington MA

INTRODUCTION

- A challenge in health care informatics is the availability of actual patient data
- Virtually all longitudinal data sources require a security plan to ensure patient privacy
- Therefore, it is challenging to develop tools and share best practices that are effective on "real-world" health data
- Recently, the Centers for Medicare and Medicaid Services (CMS) released de-identified synthetic Medicare public use data files (SynPUF) to facilitate the development of data analysis tools
- To date, the usability of SynPUF for clinical research informatics has not been reported

STUDY AIMS

- The assess the feasibility of using the SynPUF data for clinical research informatics by developing specifications for converting to the Observational Medical Outcomes Partnership (OMOP) Common Data Model version 5.0 (CDMv5)
- Also, to compare SynPUF to the actual CMS 5% sample limited data set (LDS) to determine the limitations to applying the specification to actual CMS data

METHODS

- A multidisciplinary, cross-organizational team developed an open-source specifications to extract, transform, and load (ETL) SynPUF into CDMv5
- The Observational Health Informatics and Data Science (OHDSI) group has created a number of open-source data analysis tools for researchers that work on any dataset stored in OMOP's common data model
- The following OHDSI tools were used:
  - White Rabbit tool to characterize the raw data
  - Rabbit-In-A-Hat to create mapping documentation
  - OMOP Vocabulary to translate source codes into standardized terminologies
- Each SynPUF variable was mapped to the appropriate table(s) into CDMv5, including variables about people, visits (including hospitalizations), conditions, procedures, deaths, and drugs
- Fields available in SynPUF were compared to those normally available in the CMS 5% sample LDS

RESULTS

Figure 1: Achilles SynPUF Data Overview

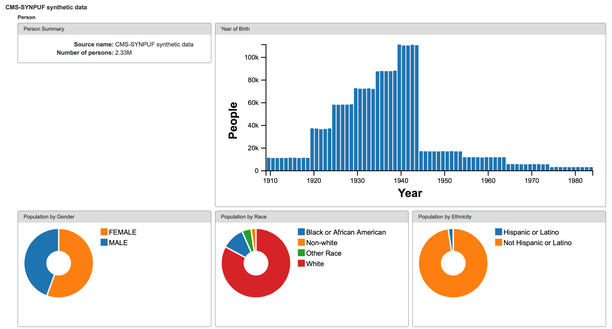


Figure 1 shows an overview of the population characteristics

Figure 5: ATLAS

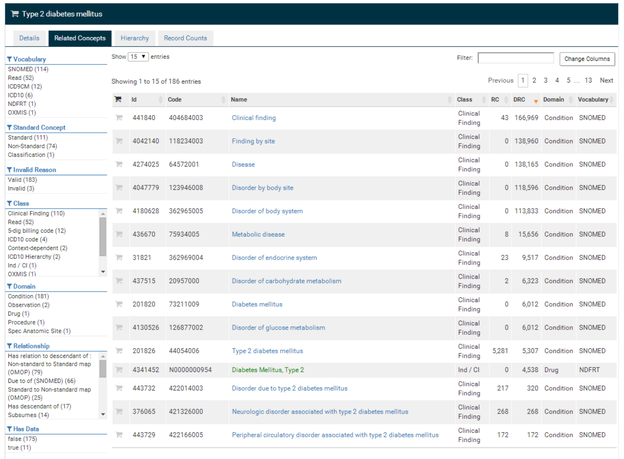


Figure 5 shows the frequency of specific concepts in the SynPUF data, allowing the researcher to identify the frequency of related concepts to help identify their possible relevance

- *RC: Row count for a specific concept*
- *DRC: Row count for a concept and all descendant concepts*

Figure 2: Achilles Drug Exposure Overview for SynPUF

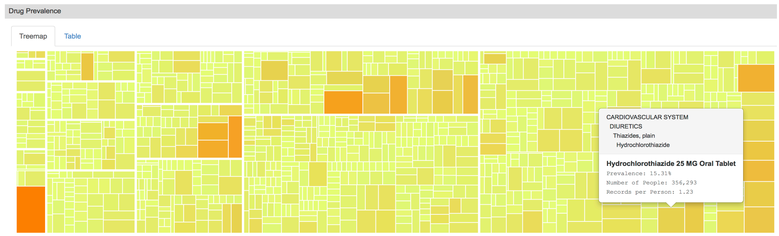


Figure 3: Achilles Condition Overview for SynPUF (1k Sample)

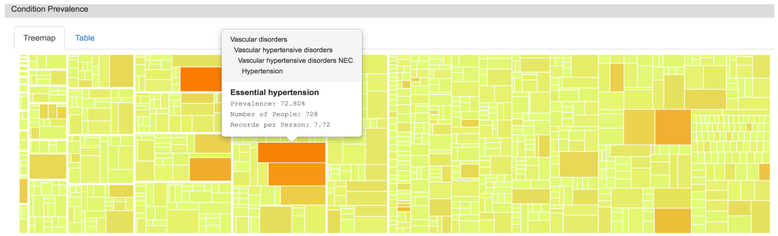
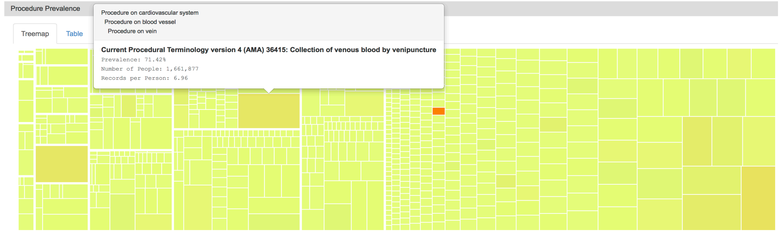


Figure 4: Achilles Procedure Overview for SynPUF



Figures 2 - 4 are heat maps of drug exposure, conditions, and procedures in SynPUF data after conversion into OMOP CDMv5

Complete reports are available online:

- *Go to http://www.ohdsi.org and follow links*
- *Access the SynPUF dashboard directly: http://bit.ly/1M82qV4*

- We were able to map or use 282 of 294 (96%) non-empty data fields in the raw SynPUF data to populate at least one table in CDMv5
- Medical conditions, procedures, visits, drug records and costs accounted for most of the available data
- Fields not mapped included
  - 9 annualized cost fields (identical to more detailed cost information already mapped)
  - 1 length of stay field (admission and discharge date already mapped)
  - 1 date field (inpatient "claim through" date selected and not "discharge" date)
- Figures 1-4 show data visualizations and Figure 5 shows the integration with the ATLAS tool (under development)
- Limitations of SynPUF compared to the LDS included
  - Missing enrollment dates
  - Limited data on location of care
  - Lack of charged amounts, no modifiers for procedure claims
  - No specialized laboratory-reporting fields
  - No revenue codes
  - No durable medical equipment, home healthcare, or hospice care data
  - The file structure also differed across years within LDS, and also between LDS and SynPUF

CONCLUSIONS

- The SynPUF data represents a usable and valuable tool for clinical research informatics
- The CDMv5 specification and data will enable researchers to
  - Explore the SynPUF data in CDMv5
  - Experiment with open-source OHDSI tools
  - Leverage the OMOP standard vocabulary
  - Develop new tools
- Because of the missing variables and different data structures, SynPUF does not facilitate a complete "plug and play" process for other CMS datasets including the 5% sample LDS
- Researchers using SynPUF for development would need to do additional work to adapt to the CMS datasets

DISCLOSURES

- SignetAccel (WES), LTS Computing (LTS), ComputerAid (DO), IMS Health (CR), and Outcomes Insights (MDD, JD, MG, and RD) all provide services related to working with electronic health data including the OMOP CDM
- EAV and AM are full time employees of Janssen Research and Development, a unit of Johnson and Johnson and hold pension rights from the company and own stock and stock options