Untangling U.S. Medicaid Data: 30 Data Quality Metrics to Support Maternal Health Studies in Two Common Data Models

Judith C. Maro[1], Sarah K. Dutcher[2], David Moeny[2], Lucia Menegussi[2], Bradley G. Hammill[3], Michael Stagner[3], Lauren Zichittella[4], Justin Vigeant[4], Laura A. Shockro[4], Christine Halbig[4], Steve Lippmann[3], Julie M. Donohue[5], Almut G. Winterstein[6], Jon D. Duke[7], Emily R. Pfaff[8], D. Keith Branham[9], James Mork[10], Nick Williams[10]

[1] Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; [2] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; [3] Department of Population Health Sciences, Duke University School of Medicine, Durham, NC; [4] Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA; [5] Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA; [6] Department of Pharmaceutical Outcomes and Policy, Center for Drug Evaluation and Safety, University of Florida, Gainesville, FL,, [7] Georgia Tech Research Institute GA, USA; [8] North Carolina Translational and Clinical Sciences Institute (NC TraCS), University of North Carolina at Chapel Hill, Chapel Hill, NC'; [9] US Department of Health and Human Services (HHS), Washington, DC; [10] National Library of Medicine, Lister Hill National Center for Biomedical Communications, Bethesda, MD

Presented at the 2024 ISPE Annual Meeting

BACKGROUND

- National U.S. Medicaid and Children's Health Insurance Program (CHIP) data are an amalgamation of up to 53 jurisdictions with both fee-for-service (FFS) and managed care (CMC) plans, making a highly heterogeneous data source.
- Using this data source in pharmacoepidemiology studies can be challenging without extensive data quality assessment on a jurisdiction-year-plan type basis.

OBJECTIVES

- To develop 30 data quality metrics to characterize Medicaid jurisdiction-specific data that have been structured in the Sentinel and Observational Medical Outcome Partnership (OMOP) common data models (CDMs).
- To improve the data infrastructure of maternal health studies using Medicaid and CHIP data to contribute to the quantity, quality, and timeliness of safety research.

METHODS

- We recruited 5 technical experts with experience using Medicaid data and CDMs. Experts prioritized metrics that would assess a jurisdiction's fit-for-purpose qualities for maternal health studies.
- Following two rounds of discussion, prioritization, and voting, the panel selected 30 data quality metrics (Table 1) in five categories: demographics (9), enrollment (6), utilization (11), death (3), and data anomalies (1). These metrics were applied in two CDMs.

RESULTS

- While much of Medicaid's data was reliable and usable for pharmacoepidemiology studies, in a few jurisdictions we found negative lengths-of-stay for inpatient stays, enrollment for infants that pre-dated conception, changes in the volume of utilization across years suggesting data missingness, and notable under-capture of emergency department visits (Figure 1). Many issues were addressable with jurisdiction-specific remedies.
- In maternal health, coverage of pregnant persons was highly variable by jurisdiction, reflective of jurisdiction-specific policies, with a median of 9 months enrollment before live birth delivery (Figure 2) and 10 months enrollment postpartum.
- Jurisdictions that have not expanded Medicaid coverage via the Affordable Care Act have notably fewer adults represented.
- Application of the same data quality metrics to the two CDMs highlighted differences in measurement in the models, particularly in race and ethnicity capture (Figure 3) and utilization stratified by care setting. The latter occurs because the two CDMs approach the concept of a "medical encounter" differently.
- Race and ethnicity variables are combined in TAF. Patients designated Hispanic do not have other race data. In the OMOP CDM, Hispanic is a valid value for Race. However, in the SCDM, race and ethnicity are separate variables, captured per 1997 OMB regulations. Recently announced updates to these regulations will combine these variables into a multi-select category with implementation due in 2029.

CONCLUSIONS

- National U.S. Medicaid data is a heterogeneous but rich data source that requires jurisdiction-specific inclusion and exclusion criteria and cleaning to use it appropriately in pharmacoepidemiology studies.
- These data quality metrics provide useful cross-state and cross-year comparisons that can be used as benchmarks to identify outliers for further investigation.

ACKNOWLEDGEMENTS/DISCLOSURES

- The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.
- J.C.M, L.Z., J.V., L.A.S., and C.H. are employees of HPHCI, an organization which conducts work for government and private organizations, including pharmaceutical companies.
- This work was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF). The OS-PCORTF funding was made available to US Food and Drug Administration (FDA) by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) through Interagency Agreement 750121PE080007. This work was supported by Task Order 75F40119F19001 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).
- This work was supported in part by the Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health. ASPE Award Number: Memorandum of Understanding REMIS#: HP-21-002.

Table 1. Data Quality Metrics for Medicaid Data

Category (# of Metrics)	Description
Demographics (9)	• Number of patients with missing date of birth, sex, race, ethnicity • Proportion of in utero (<0), infants (0-1), pediatric (1-18), and patients of child-bearing age on the last day of data eligibility (i.e., snapshot) • Descriptive statistics of age at the time of live birth
Enrollment (6)	• Descriptive statistics on contiguous and cumulative enrollment • Descriptive statistics on lengths of gaps between periods of contiguous enrollment • Descriptive statistics of gap between date of birth and date of enrollment for infants • Descriptive statistics of the duration of enrollment preceding and following live birth delivery dates
Utilization (11)	• Proportion of patients with enrollment that lack healthcare utilization • Descriptive statistics on visits per person per year by setting (inpatient, emergency department, outpatient) and ratios among these visits • Descriptive statistics on inpatient length of stay • Descriptive statistics on dispensing records per patient per year
Death (3)	• Proportion of patients with death records among those discharged "expired", and proportion of patients with evidence of utilization after death among those that have died • Descriptive statistics on age at death
Data Anomalies (1)	• Proportion of encounters without any procedures or diagnosis codes among encounters

Figure 1. Ratio of Inpatient Hospitalizations to Emergency Department Visits in the Sentinel Common Data Model (SCDM) Transformation

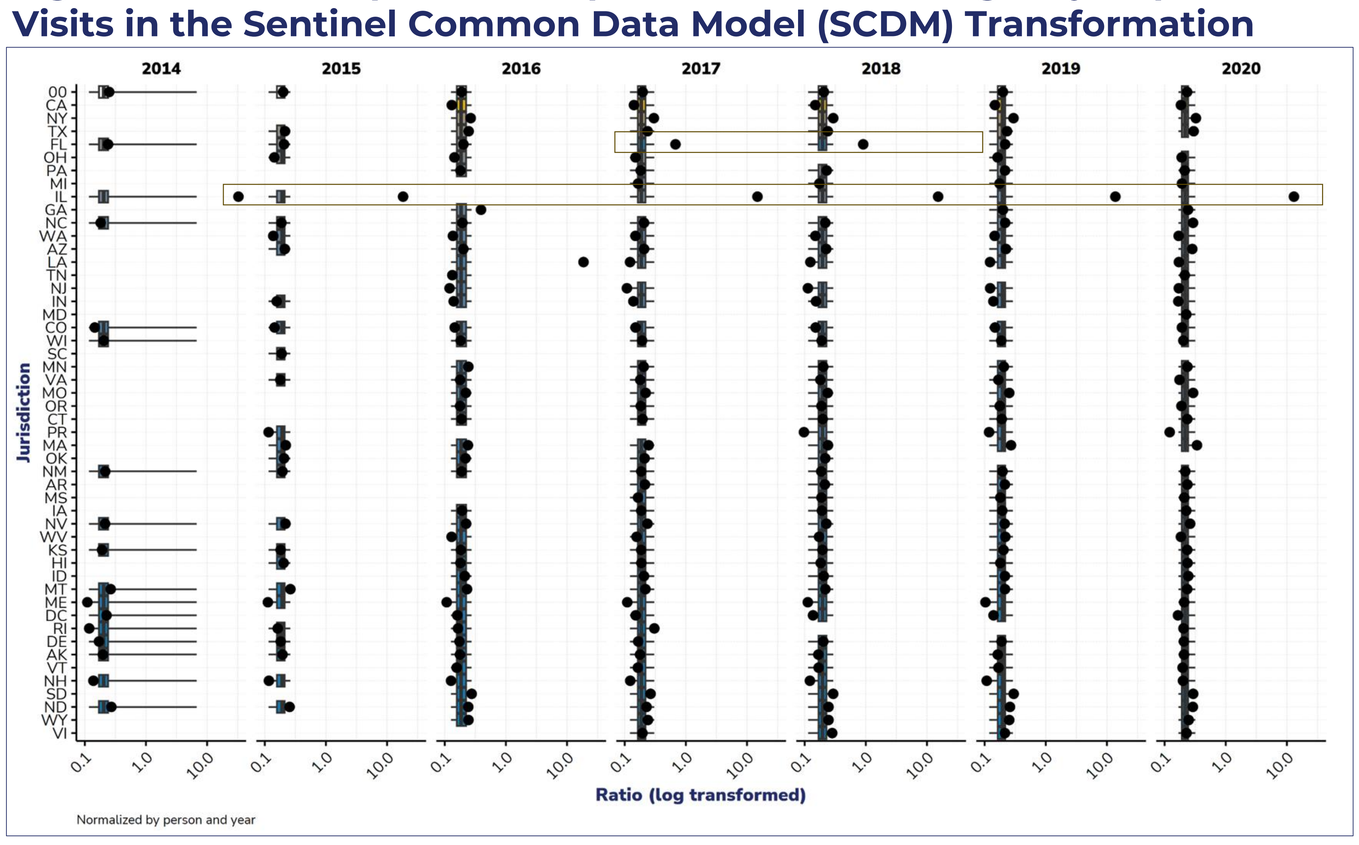

Figure 2. Duration of Pre-Live Birth Delivery Enrollment (SCDM)

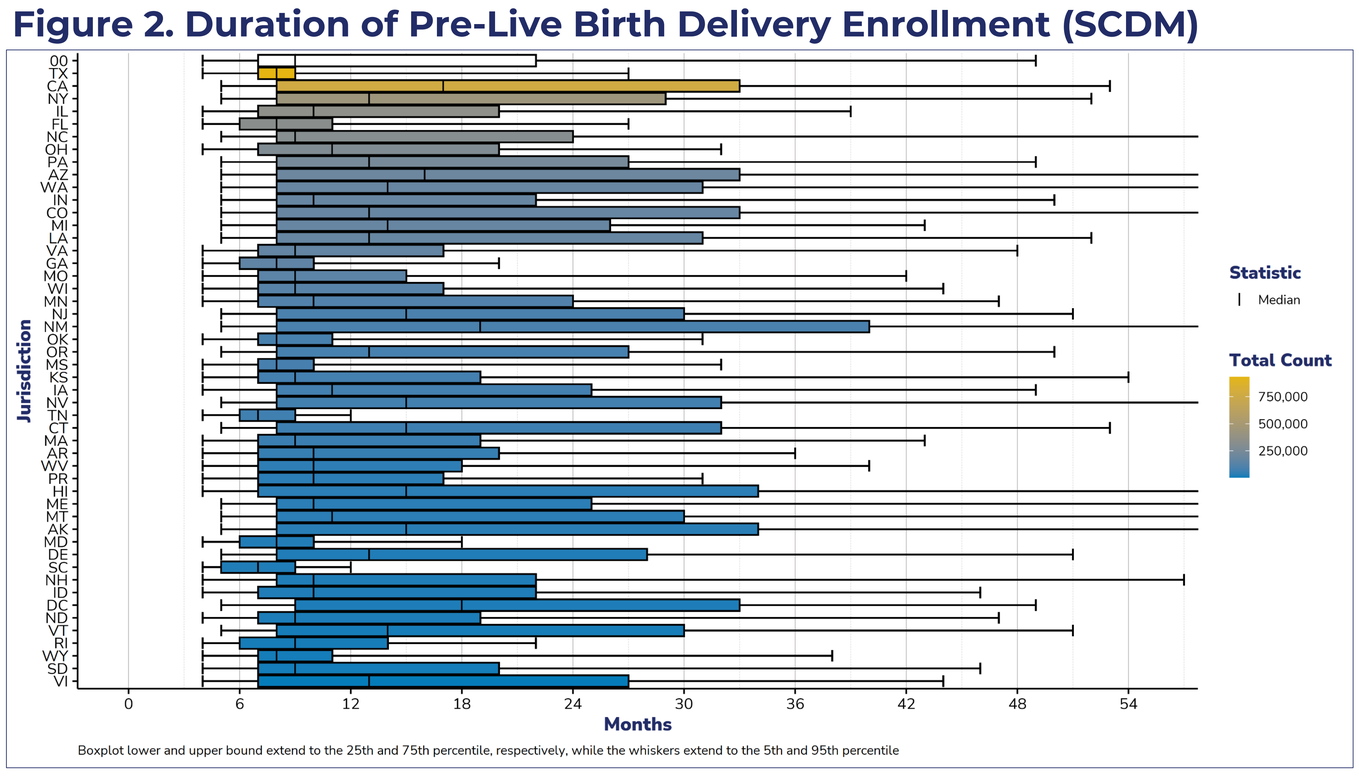

Figure 3. Missingness of Race and Ethnicity by Jurisdiction (OMOP CDM and SCDM)