Making Medicaid Data More Accessible Through Common Data Models and FHIR APIs

Final Report

Making Medicaid Data More Accessible Through Common Data Models and FHIR APIs

Final Report

Judith Maro, PhD,[1] Christine Lee Halbig, MPH,[1] Jennifer Noble,[1] Whitney Chlon, MPH,[1] Sarah Dutcher, PhD, MS,[2] David Moeny, RPh, MPH,[2] Lucia Menegussi, BSN, MS,[2] Terrence Lee, PhD, MPH,[2] Jamal Jones, PhD, MPH,[2] Jamila Mwidau, RN, MPH[2]

[1]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA

[2]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD

December 31, 2024

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise.

Preface

The views expressed in this report are those of the authors and do not necessarily represent the official position of the U.S. Department of Health and Human Services (HHS) or the Food and Drug Administration (FDA). The work discussed in this report was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF). The OS-PCORTF funding was made available to U.S. Food and Drug Administration (FDA) by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) through Interagency Agreement 750121PE080007. The Sentinel System contract was supported by Task Order 75F40119F19001 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Contents

Introduction

Project Overview and Objectives

Common data models (CDMs) have transformed the field of epidemiology and are increasingly used by public health researchers, government agencies, and others.[1] The growth of CDMs is a result of the demand for rapid evidence generation using multiple databases in combination to generate sample sizes large enough to study rare exposures, risk factors and outcomes.[2] By standardizing both the data structure and analytic approaches, evidence can be generated with greater efficiency, versatility, consistency, and scalability.

The Transformed Medicaid Statistical Information System (T-MSIS) has Analytic Files (TAF) and Research Identifiable Files (TAF RIFs)[3] that are suitable for research as part of a new research-optimized national Medicaid dataset, which begins in 2014 with full representation from all jurisdictions starting in 2016. TAF RIFs contain administrative claims data on Medicaid and Children's Health Insurance Program (CHIP) beneficiaries, including enrollment, demographics, service utilization, and payments. This project created open-source code to format TAF RIFs to the Sentinel CDM, in collaboration with the National Institutes of Health/National Library of Medicine's (NIH/NLM) formatting of these data to the Observational Medical Outcomes Partnership (OMOP) CDM,[4] with the aims of improving data access, accelerating analyses, and enabling multi-database studies.

Data quality metrics were developed to characterize each CDM-formatted version. A mother-infant linkage was created to support analyses on maternal health, illustrating the benefits of CDM transformation. A patient-centered outcomes research (PCOR) demonstration project was conducted using these CDM-formatted data, including the mother-infant linkage. The feasibility of using the Fast Healthcare Interoperability Resources (FHIR) specification to link electronic health record (EHR) data with administrative claims sources was explored. Given the overall project focus on Medicaid data, a theoretical linkage between T-MSIS TAF RIF and EHR data was used as a motivating example. Lastly, a recorded training series was developed to disseminate major project findings and educate researchers who are using, or plan to use, Medicaid data about the new data transformation tools and resources.

This joint agency project involving the U.S. Food and Drug Administration (FDA) and NIH/NLM addressed the PCOR priority to expand data capacity or data infrastructure for conducting research that informs decisions about the effectiveness of health interventions used in the Medicaid and CHIP.

Background and Problems Addressed

T-MSIS is a large, rich, and valuable data source, but one that is complex and challenging to use, making it an ideal target to standardize into CDMs for improved data

and analytic tool access, leveraging years of investment in pre-existing CDMs.[5,6,7,8] Further, as T-MSIS is a resource that aggregates data streams from 53 jurisdictions, there is a need for additional data quality efforts to enable a better understanding of the heterogeneity of the data contained within it.[9] Converting T-MSIS data into CDMs associated with sophisticated analytic infrastructure, like Sentinel and Observational Health Data Sciences and Informatics (OHDSI), is critical to maximizing Medicaid's benefit and advancing the field of PCOR.[10] Further, conversion of TAF RIF data into one of several popular CDMs enable combination with other data sources and comparison with individuals that have higher socioeconomic status, adding statistical power and enabling disparities research. The use of standardized toolkits allows data analyses with Medicaid data to be performed at scale, obviating the need for *de novo* study-specific analytic code for each analysis, which enables an overall greater volume of PCOR work. In addition to the benefits of CDMs and their associated analytic toolkits, Medicaid has a special role to play in maternal health research because of the large number of pregnancies that are publicly insured. With a standardized mother-infant linkage platform embedded in the Sentinel Common Data Model (SCDM), T-MSIS is a logical target data source to enhance and enable infant outcomes research following *in utero* exposures. This project seeks to use CDMs to improve the research infrastructure for the PCOR community.

The Medicaid population is important to the study of patient-centered outcomes for a variety of reasons:

- Medicaid provides healthcare coverage for approximately 80 million low-income Americans, including many with complex and costly needs for care.[10]
- Medicaid covers nearly half of all births, 83% of poor children, 48% of children with special healthcare needs and 45% of non-elderly adults with disabilities (including developmental disabilities such as autism, cerebral palsy, traumatic brain injury, serious mental illness and Alzheimer's disease).[10] Medicaid covers 60% of nursing home residents and is the principal source of long-term care coverage for Americans.[10]
- Medicaid is the largest insurer for adults with human immunodeficiency virus (HIV), covering more than 40% of adults.[11]
- Medicaid finances nearly one-fifth of all personal healthcare spending in the United States.[10]
- Medicaid is the single largest payer for mental health services, covering 12 million emergency care visits involving mental health disorders and substance abuse problems in 2007.[12]

A key part of the strategic vision for this project was to enhance T-MSIS data quality characterization by applying learnings from a completed Patient-Centered

Outcomes Research Trust Fund (PCORTF) project funded by the Assistant Secretary for Planning and Evaluation (ASPE) titled, *Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data*.[8] This prior project focused on how to establish new data quality metrics; a variation was used in this project intended to characterize the impact of transformation into the CDMs and assess the completeness of electronic health data and fitness for use. The supplemental data quality metrics builds on and enhances Centers for Medicare & Medicaid Services' (CMS) existing Data Quality (DQ) Atlas Tool[14] that classifies data quality on a tiered scale, providing researchers with quantitative, study-specific metrics to assess fitness for purpose.

A second part of the strategic vision was to improve the data infrastructure for studies on maternal and infant health topics. A major federal task force has cited the need to increase the quantity, quality, and timeliness of research on safety and efficacy of therapeutic products used by pregnant and lactating women.[15] To illustrate the research opportunities of the CDM-formatted TAF RIF data and the mother-infant linkage, maternal health and birth defects experts from the FDA, the Centers for Disease Control and Prevention (CDC), the National Institutes of Health (NIH), the Health Resources and Services Administration (HRSA), and the Office of the National Coordinator for Health Information Technology (ONC) collaborated on a study to answer important public health questions related to screening and treatment of prenatal and congenital syphilis.[16]

The third part of the strategic vision was to improve the depth and accessibility of T-MSIS data through linkages with EHR data using Fast Healthcare Interoperability Resources with an application programming interface **(**FHIR APIs). The rationale for exploring FHIR is based on the 21st Century Cures Act which requires that certified EHRs support standardized APIs allowing for both bulk (population-level) data and individual data sharing via FHIR.[17] The ability to link EHR data from large hospital systems, academic medical centers, or Federally Qualified Healthcare Centers to Medicaid through FHIR APIs would help realize major goals within ONC's National Health IT Priorities for Research: A Policy and Development Agenda,[18] which seeks to leverage EHR data for research and advance a health information technology (IT) infrastructure to support scientific discovery. To this end, this project explored the technical requirements, data privacy, and data governance considerations regarding linking patient healthcare data between private sector healthcare entities and government operated federal research enclaves, using a theoretical linkage between TAF RIF and EHR data as an example use case.

This project advances four OS-PCORTF functionalities to enable the use of T-MSIS data and build data capacity for PCOR.[19]

1. Standardized Collection of Standardized Clinical Data: Using freely available and publicly posted code developed in this project, researchers will be able to transform CMS' TAF data into the Sentinel and OMOP CDMs. CDMs allow for

standardizing healthcare data based on common data element standards across research projects and networks, thereby facilitating aggregation of data across data sources.

2. Linking Clinical and Other Data for Research: Because TAF data are sourced from administrative claims, researchers will be able to follow patients across the care continuum over time. A white paper describes the feasibility of using FHIR APIs to link TAF data with EHR data, a linkage that can provide researchers with improved depth and breadth of data to address a wider variety of PCOR questions.
3. Use of Clinical Data for Research: Use of routinely collected healthcare data is valuable for conducting PCOR studies in the unique population enrolled in Medicaid and CHIP. For example, the SCDM-formatted TAF RIF data are one of the contributing data sources in the FDA's Sentinel System, which conducts observational studies assessing medical product safety. If FHIR APIs can be used to link EHR data to TAF RIF datasets, this may enable PCOR researchers to access richer clinical data including laboratory data, genetic information (e.g., breast cancer gene mutations) and patient-reported outcomes (PROs) found in EHR systems.
4. Use of Enhanced Publicly Funded Data Systems for Research: Researchers using TAF RIF data will be able to format their data into two CDMs, better enabling them to leverage these data infrastructures and their associated toolkits in their own research, link it to other data sources, and/or aggregate it with other publicly or privately-funded data sources.

Major Project Tasks

This project included six major task areas: 1) create freely available code to transform TAF RIFs into the Sentinel CDM; 2) develop data quality metrics to characterize these data once in the Sentinel and OMOP CDMs; 3) link mothers and infants to create a data resource for infant outcome studies; 4) conduct a patient centered outcomes research study on maternal health as a demonstration project; 5) develop a white paper to explore considerations for linking EHR data via a FHIR API to T-MSIS data; 6) conduct activities for stakeholder engagement and sustainability, including development of a training webinar series. In this section, we describe each project task in detail.

Task 1. Develop Freely Available Code to Format the TAF RIF Data into the Sentinel Common Data Model

As mentioned above, CDMs provide a standardized structure that can be analyzed using analytic tools and allow PCOR projects to be completed at scale. They

also allow for integration with other data sources that are already formatted into the same CDM. For this task, a rigorous process was undertaken to identify data for inclusion for transformation in the SCDM that met a minimum level of data quality standards. Select variables or fields in the source TAF RIF data were deemed critical, largely based on the principle of their relevance to "complete capture." For most PCOR studies, a high priority is placed on complete or near-complete capture of healthcare utilization during a specific time period for a patient. Variables that directly contribute to this assessment were deemed the most critical for initial inclusion in the data resource. These consist of number of enrollment spans, length of enrollment gaps, overlapping enrollment spans, dual eligibility codes, restricted benefits codes, service users, claims volume, Comprehensive Managed Care (CMC) plan encounters, admission and discharge dates, diagnosis and procedure codes, and types of service. These variables had to not be deemed "Unusable" per CMS' Data Quality Atlas to merit inclusion. Each unique jurisdiction (e.g. state, territory)-year-plan type (e.g., comprehensive managed care plans, fee-for-service) was assessed independently according to the same criteria. Following this broad inclusion/exclusion criteria, individual beneficiaries were excluded if they had limited benefits enrollment and during periods of enrollment when their dual Medicare-Medicaid eligibility status was missing. This is because Medicare is the primary payer for dual-eligible beneficiaries, and the emphasis is on complete capture of data. Thus, all individuals included were required to have comprehensive coverage. The final criterion excluded capitated or supplemental payment claims specifically that do not indicate a service was rendered. Following that initial assessment documented in a publicly posted memo detailing TAF data characteristics, the project team created a programming specification document to map the data elements in the TAF RIF into the SCDM, and then performed the programming and quality control procedures to transform the data, as well as developed a user guide for other researchers.

Task 2. Leverage the Office of the Assistant Secretary for Planning and Evaluation Funded Data Quality Metrics Model to Characterize Data Quality

The project team convened a Technical Expert Panel (TEP) to inform the development of a comprehensive quality assurance plan to help investigators evaluate if the CDM-transformed data were fit for the purpose of maternal health studies (Figure 1). TEP members were experts in Medicaid data or CDMs. After two rounds of voting, 30 data quality metrics were selected by the TEP and categorized by topic area: Demographics (9), Enrollment (6), Utilization (11), Death (3), and Orphan Records (1). See Table 1 below for a description of each metric category. Metrics in italics particularly pertain to PCOR studies focused on maternal health.

The 30 data quality metrics were chosen to be complementary to resources already available including the DQ Atlas, the Sentinel Quality Assurance program, and the Observational Health Data Sciences and Informatics (OHDSI) Data Quality Dashboard. They also considered best practices put forth in the prior ASPE project, *Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data*,[20] and were guided by literature[21]. SAS code was written to quantify these metrics against the SCDM-transformed TAF RIF data and results were reviewed by the project team and TEP. In parallel, NLM also evaluated these metrics OMOP CDM-transformed TAF RIF data and results were compared.

Figure 1. *The Many Layers of Quality Assurance*

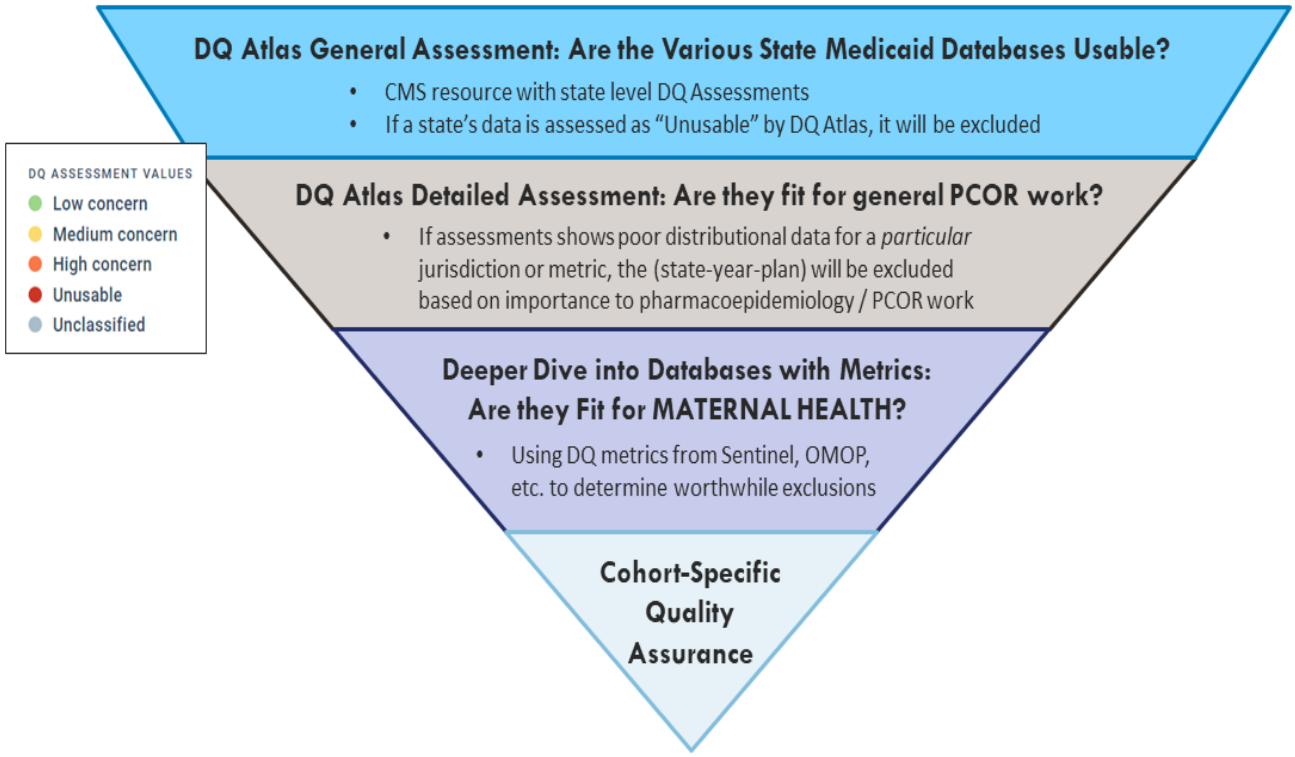


Table 1. *Data Quality Metrics for Medicaid Data*

Category (Number 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
Orphan Records (1)	• Proportion of encounters without any procedures or diagnosis codes among encounters

Task 3. Develop Freely Available Code to Create a Mother-Infant Linkage in the Sentinel Common Data Model

Data models such as the SCDM that include mother-infant linked information can evaluate maternal and infant outcomes in relation to medical product use or other exposures or events that occur during pregnancy. Thus, programming specifications

were developed, along with the completed program and users guide, to enable linkage of mothers and live-born infants in the transformed TAF RIF dataset.

Task 4. Patient Centered Outcomes Research Demonstration Project

To demonstrate the usability of the newly transformed, quality checked TAF RIF data (Tasks 1-3), including the mother-infant linkage, a workgroup was formed and included subject matter experts from academia, FDA, CDC, NIH, HRSA, and ONC to design and conduct the study. The study workgroup selected the important public health topic of congenital syphilis and assessed syphilis screening and treatment during pregnancy among publicly and commercially insured pregnant women in the US. A protocol was published to describe the study and then was implemented in the newly transformed CDM-formatted dataset. A manuscript for submission to a peer-reviewed journal was drafted to describe the project, methodology, results, and conclusions.

Task 5. Develop a White Paper to Assess the Feasibility of Using FHIR APIs to Link Electronic Health Record Data to Transformed Medicaid Statistical Information System

A study workgroup consisting of academia, FDA, NLM, and ONC members conducted a feasibility assessment on the ability to use FHIR APIs to link EHR data with administrative claims sources maintained by payers. Linking administrative claims with EHR data expands the type and scope of research questions that can be answered with each source individually, supplementing the "complete capture" of medically attended events found in claims with the granular clinical information recorded within the EHR like vital signs, laboratory results, and inpatient medication administrations. Given the overall project focus on Medicaid data, the white paper includes an example use case of a linkage between T-MSIS TAF RIF files and EHR data to discuss key considerations on the feasibility of linkage using FHIR APIs. Although the T-MSIS TAF RIF and EHR data linkage is described as an example, the general process outlined in the white paper could be used with other administrative claims resources provided both organizations could send and receive FHIR-formatted data via FHIR APIs.

Task 6. Stakeholder Engagement and Sustainability

For this task, the aforementioned TEP selected the data quality metrics in Task 2 and provided non-binding guidance on potential PCOR demonstration project topics in Task 4 using the new CDM-formatted TAF RIF dataset built in Tasks 1 and 3. Once a demonstration project was identified by the multiagency workgroup, as discussed in the Task 4 section above, the TEP was consulted for feedback on the study design and parameters. The TEP also provided guidance on the training materials that were developed under Task 6 to help ensure that major findings were optimally disseminated to the Medicaid research community. Dissemination efforts were primarily focused on a

recorded webinar series aiming to promote greater utilization of the CDM translation code and use of TAF RIF data for PCOR.

Accomplishments by Final Deliverables

Each project task outlined in the previous section has associated deliverables that are accessible to the public. Here we discuss the final project deliverables, how to access them, and the work accomplished for each.

Summary of Final Deliverables

Table 2 below provides a summary of the final deliverables for this project and instructions on how the public can access them. The subsequent sections provide details on the work accomplished in the completion of each deliverable.

Table 2. *Summary of Final Deliverables*

Final Deliverable	How to Access Final Deliverable
Data processing code (Code Pack) to transform five separate TAF RIF files into a harmonized SCDM representation. **Task 1.**	Access the Code Pack materials for SCDM v8.1.0 here on the Sentinel GIT repository.
User's guide to transform five separate TAF RIF files into a harmonized SCDM representation. **Task 1.**	Access the User Guide for SCDM v8.1.0 here on the Sentinel GIT repository.
Recorded, publicly available presentation on the major DQ Metrics findings, including how it helps to build data capacity. **Task 2.**	View the presentation on DQ Metrics here.
Data processing code (Code Pack) to create a mother-infant linkage for the TAF RIF data formatted in the SCDM. **Task 3.**	Access the Code Pack materials for SCDM v8.1.0 here on the Sentinel GIT repository.
User's guide detailing procedure to create a mother-infant linkage for the TAF RIF data formatted in the SCDM. **Task 3.**	Access the User Guide for SCDM v8.1.0 here on the Sentinel GIT repository.
An article that describes the PCOR demonstration project methods, results, and conclusions for submission to a peer-reviewed journal. **Task 4.**	Journal article is being finalized for submission at the time of writing this report. View the protocol for the PCOR demonstration project here.
A white paper containing major findings of the feasibility assessment of using FHIR APIs to link Medicaid T-MSIS to EHR data. **Task 5.**	View the final White Paper here on the Sentinel Initiative website.

Final Deliverable	How to Access Final Deliverable
Publicly available series of video recorded trainings to describe how to use the new data infrastructure tools developed and communicate major project findings to the Medicaid research community. **Task 6.**	View the recorded trainings here.
Dissemination of training materials and project results at scientific conferences, such as the Sentinel Annual Public Workshop. **Task 6.**	**Presentations:** • A brief overview explaining how Sentinel's public health surveillance efforts would benefit from the addition of CMS Medicaid data was presented during the 14th Annual Sentinel Initiative Public Workshop in November 2022. View the presentation here. • *Transforming Medicaid Data into the Sentinel Common Data Model* was presented at the Maternal Health Consortium and the OS-PCORTF webinar in January 2023. View the presentation here. • *Variation in Mother-Infant Linkage Rates by Jurisdiction in U.S. Medicaid Data* was presented at the 39th International Conference on Pharmacoepidemiology & Therapeutic Risk Management in August 2023. View the presentation here. **Posters:** • *Diversifying the FDA's Sentinel System with Rigorous Quality Inclusion Rules for the U.S. Medicaid Population* was presented at the 39th International Conference on Pharmacoepidemiology & Therapeutic Risk Management in August 2023. View the poster here.

Final Deliverable	How to Access Final Deliverable
	• *Prenatal Syphilis in the US: Characterizing Screening and Treatment During Pregnancy in Publicly and Commercially Insured Individuals* was presented at the 2024 ISPE Annual Meeting in August 2024. View the poster here. • *Untangling U.S. Medicaid Data: 30 Data Quality Metrics to Support Maternal Health Studies in Two Common Data Models* was presented at the 2024 ISPE Annual Meeting in August 2024. View the poster here.
A Final Report that is 508-compliant summarizing the project methods, findings, and information about the deliverables including a nontechnical description, the audience, how to access it. **Task 6.**	This report meets this deliverable.

Deliverable 1: Data processing code (Code Pack) to transform five separate TAF RIF files into a harmonized Sentinel Common Data Model representation.

Data inclusion criteria were applied to the TAF RIF data to ensure that only those data that met a standard minimum level of quality were included in the transformation. After initial data inclusion and characterization was complete, SAS code was written to transform the TAF RIF data into the Sentinel Common Data Model (SCDM) format. This code was informed by previous code created by project team to transform Medicare fee-for-service data into the SCDM format and includes crosswalks for unique patient identifiers across multiple years. This transformation code also maps the data from the TAF RIF source files to the SCDM variables and values. One example was transforming portions of numeric Medicaid billing codes to the setting that the patient encounter took place, which is a series of values in SCDM such as Ambulatory, Emergency, or Inpatient stays. This SAS code is freely downloadable and can be used by any researcher that has obtained the necessary data use agreements and approvals to obtain the TAF RIF data, either in physical hard copy or in CMS's Virtual Research Data Center (VRDC). Data in

VRDC is available in both SAS and non-SAS formats, but most researchers use the default SAS-based files, and thus the SAS programming code can be utilized. Several researchers have used similar Medicare transformation codes in their studies. Notably, this includes three industry-funded studies awarded to the Health Data Collaborations for Safety Effectiveness Research (HDC-SER) program at the Harvard Pilgrim Health Care Institute, as well as research supported by the Reagan-Udall Foundation for the FDA.[22-24]

Deliverable 2: User's guide to transform five separate TAF RIF files into a harmonized Sentinel Common Data Model representation.

The project team published a user's guide for the developed SCDM transformation code. This document walks through each SAS package in the Code Pack and its purpose and provides comprehensive instructions for execution of the code and user-specifications that need to be applied in the specific source data. It is intended for researchers that will convert TAF RIF files into the SCDM using the code described in Deliverable 1.

Deliverable 3: Recorded, publicly available presentation on the major Data Quality Metrics findings, including how it helps to build data capacity.

A presentation focused on the Data Quality Metrics findings was created as part of the recording training series for this project (see Deliverable 8). The full training series is accessible to the public via YouTube and the Sentinel Initiative website.

In addition, a technical specification of the 30 metrics and how they were calculated and a SAS-based program which can be used against SCDM-transformed data is publicly posted.

Deliverable 4: Data processing code (Code Pack) to create a mother-infant linkage for the TAF RIF data formatted in the SCDM.

A SAS-based program was developed to identify deliveries and live-born infants and to perform and quality check a linkage. This program has requirements for inclusion of mothers and infants based on appropriate timing, encounters, and enrollment spans. These requirements ensure inclusion of records with a minimum level of quality and enough observation time in the data to be able to assess infant outcomes following an event or exposure (e.g., use of a medical product) during pregnancy. The mothers and infants were subsequently linked together using the encrypted T-MSIS case number (MSIS_CASE_NUM variable) which is a jurisdiction-assigned unique identifier which, in Medicaid data, often acts as family-level identification. Later, access was obtained to de-encrypted TMSIS data to improve linkage. Researchers can use this SAS code in the same way as described in Deliverable 1 to perform pregnancy or maternal health related studies.

Deliverable 5: User's guide detailing procedure to create a mother-infant linkage for the TAF data formatted in the Sentinel CDM.

Additional information to support users of the Code Pack for the mother-infant linkage was added to the user guide in Deliverable 2 and publicly posted.

Deliverable 6: An article that describes the PCOR demonstration project methods, results, and conclusions for submission to a peer-reviewed journal

A manuscript was developed for the PCOR demonstration project that characterized the screening and treatment of prenatal and congenital syphilis in the US based on a protocol that was publicly posted. Transformed TAF RIF data for 35 jurisdictions that met data quality standards were used for this project. The project team identified a cohort of pregnancies with continuous insurance enrollment for individuals aged 10-54 years using a validated list of diagnosis and procedure codes indicating a pregnancy outcome event (live birth, stillbirth, or miscarriage). A validated algorithm was used to estimate the gestational age (EGA) at the time of the pregnancy outcome and the last menstrual period. This allowed the study team to calculate the trimester of syphilis screening and/or treatment. Medicaid-insured patients were compared with commercially-insured individuals, demonstrating the principal advantage of using a CDM for analysis that allows multiple data sources to be analyzed in concert and with subgroup comparisons.

Three study questions were included in the study:

- Question 1: What proportion of pregnant women are tested for syphilis in pregnancy?
- Question 2: What proportion of syphilis-diagnosed pregnant women are treated for syphilis in pregnancy, and when?
- Question 3: Are infants born to pregnant women with syphilis diagnosis in pregnancy tested and treated in the first 30 days of life?

To analyze infant outcomes (Question 3), only the live birth-infant pairs identified as part of the mother-infant linkage (Task 3) were included.

Deliverable 7: A white paper posted on the Sentinel and HHS ASPE website containing major findings of the feasibility assessment of using FHIR APIs to link Medicaid T-MSIS to EHR data

The Task 5 study workgroup developed a white paper describing the feasibility of using the FHIR specification and APIs to enable the linkage of EHR data with administrative claims sources maintained by payers. The white paper was publicly posted on the Sentinel Initiative website and includes the following information: a brief

background of key terms and definitions related to T-MSIS, FHIR, and legislation related to data sharing; considerations for the scope of additional EHR data to be included, technical aspects of obtaining and linking data, and aspects of data governance that concern the use of those data; and factors that impact the use of a linked claims-EHR dataset. The use case presented in the white paper focused on a theoretical linkage between CMS/T-MSIS TAF-RIF and EHR data, but the same considerations also apply to linkages with data from other health insurance payers. Investigators would benefit from understanding some of the benefits and challenges of pursuing such a linkage to enhance PCOR studies.

Deliverable 8: Publicly available series of video recorded trainings to describe how to use the new data infrastructure tools developed and communicate major project findings to the Medicaid research community.

To educate researchers who are using, or plan to use, Medicaid data about the new data transformation tools and disseminate major project findings, the project team created a publicly available recording training series with six distinct chapters. Each chapter covers a specific topic area addressed within the major project tasks; Table 3 below provides an overview of each chapter in the series. The recordings can be found here on YouTube and can also be accessed here on the Sentinel Initiative website.

Table 3. *Chapters Included in the Publicly Available Recording Training Series*

Recorded Training Series Chapter Title	Topic Area Covered
Chapter 1: *Transforming Medicaid and Children's Health Insurance Program Data for use with the U.S. Food and Drug Administration's Sentinel Common Data Model*	The benefits of TAF RIF data transformed into the Sentinel CDM with guidance on how to leverage the data processing code to transform the TAF RIF files.
Chapter 2: *Transforming Medicaid and Children's Health Insurance Program Data for use with the Observational Medical Outcome Partnership (OMOP) Common Data Model*	The benefits of TAF data transformed into the OMOP CDM with guidance on how to leverage the data processing code to transform the TAF RIF files.
Chapter 3: *Creating a Mother-Infant Linkage using TAF Data in the Sentinel Common Data Model*	How to develop a mother-infant linkage with TAF RIF data in the Sentinel CDM.
Chapter 4: *Data Quality Metrics in US Medicaid Data: Results from Sentinel's Medicaid Data Mart*	Results of the 30 data quality metrics developed and applied to the TAF RIF data transformed into the Sentinel CDM.

Recorded Training Series Chapter Title	Topic Area Covered
Chapter 5*: Lessons Learned: Feasibility of Using the Fast Healthcare Interoperability Resources (FHIR) Specification to Enable EHR Data Linkage with Administrative Claims*	Lessons learned from the white paper describing the feasibility of FHIR-based claims-EHR data linkages, using a potential linkage between TAF RIF and EHR data as an example use case.
Chapter 6: *Prenatal and Congenital Syphilis in the US: Characterizing Screening and Treatment*	Results from the PCOR demonstration project leveraging TAF RIF transformed into the Sentinel CDM.

Deliverable 9: Dissemination of training materials and project results at scientific conferences, such as the Sentinel Annual Public Workshop.

Major project findings and lessons learned were disseminated in multiple forums. A project overview highlighting how FDA's public health surveillance efforts would benefit from Sentinel's inclusion of CMS Medicaid data was featured as part of the 14th Annual Sentinel Initiative Public Workshop. This two day virtual webinar took place November 15-16, 2022, and was hosted by the Duke-Margolis Center for Health Policy under a cooperative agreement with the FDA. Event materials can be viewed here.

An overview of the Sentinel Medicaid DataMart in SCDM format was presented at the Maternal Health Consortium on January 17, 2023, and the OS-PCORTF webinar on January 23, 2023. Both a presentation and a poster were featured at the 39th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, which was held August 23-27, 2023, in Halifax, Nova Scotia, Canada. The presentation, titled *Variation in Mother-Infant Linkage Rates by Jurisdiction in U.S. Medicaid Data*, assessed mother-infant linkage rates by U.S. Medicaid jurisdiction and explored jurisdiction-specific reasons for differential results. The poster, titled *Diversifying the FDA's Sentinel System with Rigorous Quality Inclusion Rules for the U.S. Medicaid Population*, explored establishing jurisdiction-level, beneficiary-level, and record-level criteria for inclusion of TAF RIF data into a SCDM-compliant database and documenting the fit-for-purpose requirements of the TAF RIF data for Sentinel System regulatory needs (link).

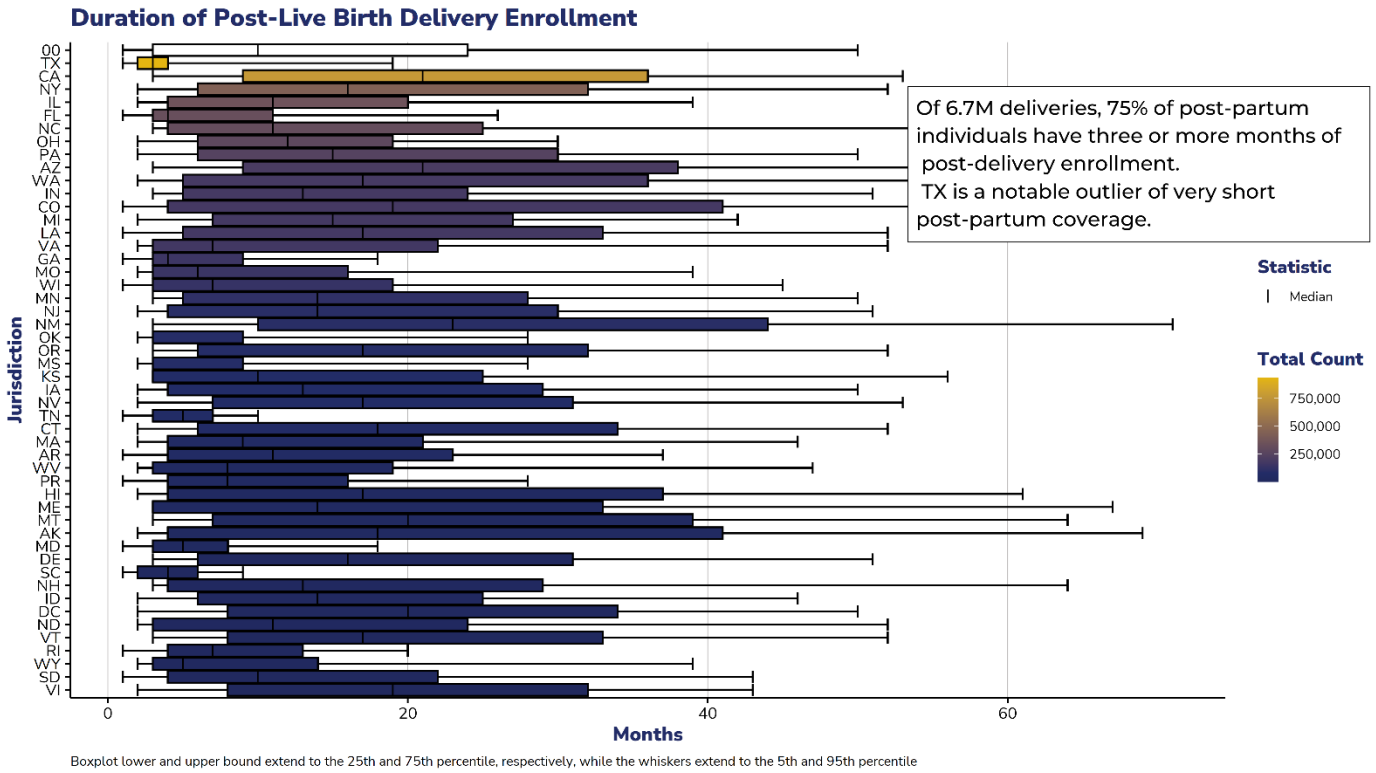
Two posters originating from this project were presented at the 2024 International Society for Pharmacoepidemiology Annual Meeting, held August 24-28, 2024, in Berlin, Germany. *Prenatal and Congenital Syphilis in the US: Characterizing Screening and Treatment* provided an overview of the PCOR demonstration project assessing syphilis screening and treatment during pregnancy among publicly and commercially insured pregnant women in the US. The second poster, titled

Untangling National Medicaid Data: 30 Data Quality Metrics to Support Maternal Health Studies in Two Common Data Models, discussed results of the data quality metrics assessments. This poster was selected to be showcased in a spotlight session at the event. Links to both posters can be found in Table 2.

Lessons Learned and Considerations for Future Work

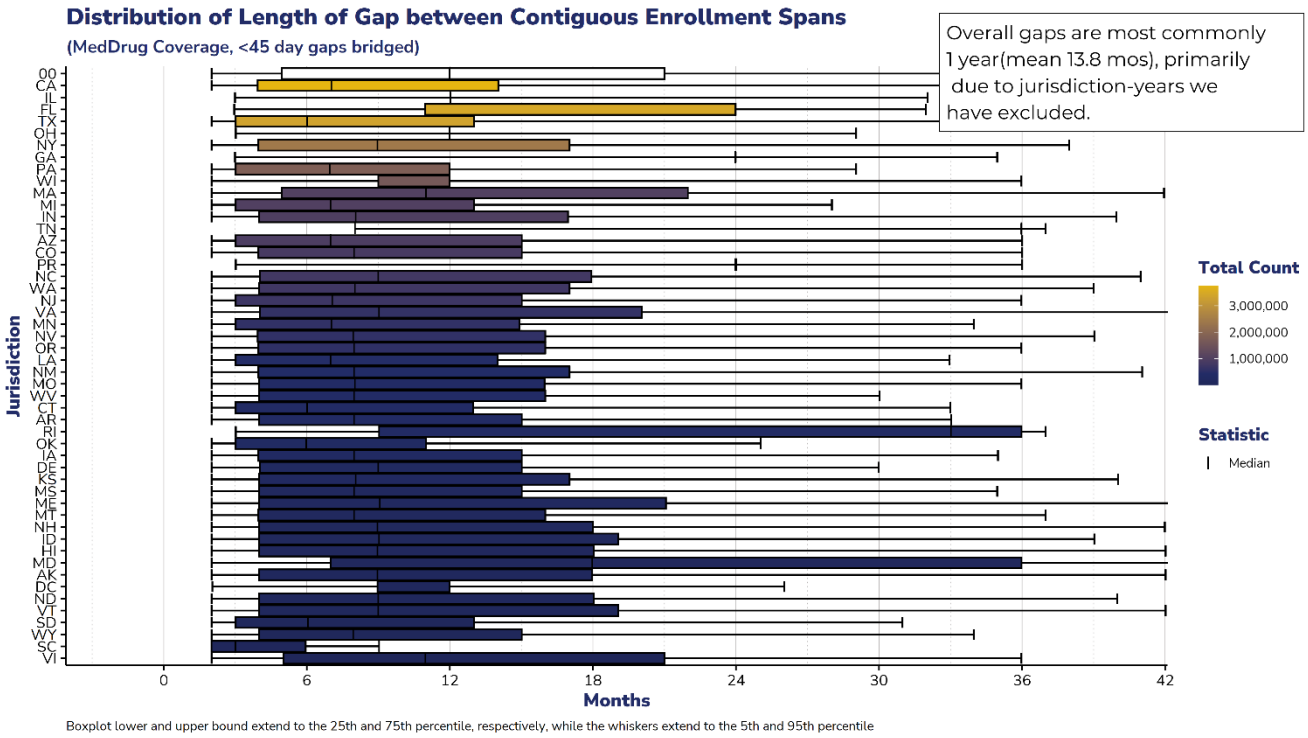
Heterogeneity in jurisdictions for data inclusion and quality year over year. Medicaid and State Children's Health Insurance Programs (SCHIP) are administered independently by each jurisdiction (e.g., state, territory) with data being reported centrally to CMS periodically. Resources like the DQ Atlas are an excellent starting point for assessment of data quality and the additional data quality metrics developed in this project can supplement this resource. The metrics developed in this project have a directed focus on PCOR studies, especially those on maternal health. Data must be fit for purpose when being used for a study, and understanding sources of data heterogeneity, or merely its existence, may prevent inappropriate generalizations or conclusions when attempting to use TAF RIF data. For example, Figure 2 below shows the duration of Medicaid enrollment following a live birth delivery by jurisdiction. While 75% of post-partum individuals have ≥3 months of post-delivery enrollment there is a wide range depending on jurisdiction. This information is important to understand when designing a PCOR study, especially for assessment of pregnancy outcomes.

Figure 2 2. *Duration of Post-Live Birth Delivery Medicaid Enrollment by Jurisdiction*



Another challenge with heterogeneity is continuity of care. Millions of individuals have gaps of medical and drug coverage under Medicaid which vary by jurisdiction. Reasons may be disenrollment from one state when moving to another, between child births, or due to changes with requirements at the state level.[25] Figure 3 below shows the distribution of the length of gaps between contiguous Medicaid enrollment periods. The variability in the longitudinal capture of patients' care impacts the ability for studies to include a representative and generalizable cohort of patients into PCOR studies, particularly those that may require longer observation periods.

Figure 3 3. *Distribution of Length of Gap Between Contiguous Enrollment Spans by State*



Inclusion/exclusion and imputation decisions can change the data interpretation significantly. In this project, data quality was strictly enforced to ensure complete capture of care for PCOR studies. However, for jurisdictions that have threats to complete capture over the time period covered (i.e., the data are deemed unusable in select years), these decisions can affect the overall longitudinal capture for patients in these jurisdictions. The most common observed gap between contiguous enrollment periods was 12 months (see Figure 3 above), likely often due to specific jurisdiction-years that were excluded for data quality reasons. Further, data were not imputed when they were not available or missing. Any decision regarding inclusion, exclusion, or imputation can change the interpretation of conclusions for any study, particularly the degree of generalizability to the entire Medicaid population. These decisions should always be disclosed in any use of these data for PCOR studies.

21st Century Cures act has increased adoption of FHIR but health data exchange for research purposes has distinct governance needs from health data exchange for normal healthcare operations. The gradual adoption of FHIR interoperability standards as promoted within the 21st Century Cures act defines a common standard for the industry. However, the current use cases are still primarily to support clinical care or required operational purposes such as prior authorization from insurers. Considerations for transmitting patient-identifiable data to external organizations for public health surveillance or research may still require changes to agreements for data use and IT infrastructure and oversight to ensure compliance. Costs

and time to address these governance considerations may be barriers for many public health surveillance and research use cases. Additionally, while CMS has data quality standards for aggregating data from multiple jurisdictions, each individual payer or EHR system may have their own data quality measures that differ from one institution to another. So, while FHIR solves the issue of reconciling unique data formats, these linked datasets still need to be assessed for quality and fit-for-purpose prior to use for each scientific question based on the source data and processes to aggregate (including de-duplication).

There is no "right answer" for data quality metrics. Most of the data quality metrics created in this project were ways to characterize datasets so that researchers could come to conclusions on their fit for purpose for PCOR or maternal health studies. Because of the project team's choices to exclude data that was not fit-for-purpose outright, the metrics provide ways to understand cross-jurisdiction and cross-year inconsistencies in the data being observed. For example, there were metrics that allowed easy visualization and detection that certain year-jurisdiction combinations had under-captured emergency department claims relative to inpatient claims. Negative lengths-of-stay were observed in some jurisdiction's inpatient data that later needed to be investigated and highlighted potential misclassification of certain data variables. Data quality metrics can highlight heterogeneity in the data, particularly given that TAF RIF reflects jurisdictions with different eligibility criteria and other reasons that might give rise to significant variation in healthcare utilization. For example, the number of low-income adults that are eligible for PCOR study in Medicaid data is highly dependent on various jurisdictions' adoption (or non-adoption) of the Affordable Care Act provisions. The data quality metrics developed in this project that subgroup the overall dataset by jurisdiction, year, and plan type can more easily depict the heterogeneity and provide a useful resource for researchers that might be making data quality decisions or conducting research around investigating heterogeneity. Thus, data quality characterization does not necessarily deem some data "bad" or "inaccurate" per se but identifies differences. These distinctions are important for researchers to recognize who seek to use Medicaid data as a whole for PCOR studies.

References

1. Schneeweiss S, Brown JS, Bate A, Trifirò G, Bartels DB. Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products. Clinical Pharmacology & Therapeutics. 2019;107(4):827-833. doi:https://doi.org/10.1002/cpt.1577
2. Maro JC, Toh S. Invited Commentary: Go BIG and Go Global-Executing Large-Scale, Multisite Pharmacoepidemiologic Studies Using Real-World Data. Am J Epidemiol. 2022 Jul 23;191(8):1368-1371. doi: 10.1093/aje/kwac096.
3. Centers for Medicare & Medicaid Services. Introduction to the Transformed Medicaid Statistical Information System (T-MSIS) Analytic Files (TAF). Available from: https://www.medicaid.gov/medicaid/data-and-systems/downloads/macbis/taf-introduction.pdf.
4. Williams N. Building the observational medical outcomes partnership's T-MSIS Analytic File common data model. Informatics in medicine unlocked. 2023;39:101259. doi:https://doi.org/10.1016/j.imu.2023.101259
5. Rai A, Maro JC, Dutcher S, Bright P, Toh S. Transparency, reproducibility, and replicability of pharmacoepidemiology studies in a distributed network environment. Pharmacoepidemiology and Drug Safety. 2024;33(6). doi:https://doi.org/10.1002/pds.5820
6. Lyons JG, Shinde MU, Maro JC, et al. Use of the Sentinel System to Examine Medical Product Use and Outcomes During Pregnancy. Drug safety. 2024;47(10):931-940. doi:https://doi.org/10.1007/s40264-024-01447-z
7. Brown JS, Mendelsohn AB, Nam YH, et al. The US Food and Drug Administration Sentinel System: a national resource for a learning health system. Journal of the American Medical Informatics Association. Published online September 12, 2022. doi:https://doi.org/10.1093/jamia/ocac153
8. Connolly JG, Wang SV, Fuller CC, et al. Development and application of two semi-automated tools for targeted medical product surveillance in a distributed data network. Current epidemiology reports. 2017;4(4):298-306. doi:https://doi.org/10.1007/s40471-017-0121-0
9. Centers for Medicare & Medicaid Services. Introduction to the Transformed Medicaid Statistical Information System (T-MSIS) Analytic Files (TAF) Transcript. Available from: https://www.medicaid.gov/medicaid/data-and-systems/downloads/macbis/taf-introduction-transcript.pdf
10. KFF. 10 Things to Know About Medicaid. Available from: https://www.kff.org/medicaid/issue-brief/10-things-to-know-about-medicaid/.

11. KFF. Medicaid 101. Available from: https://www.kff.org/health-policy-101-medicaid/?entry=table-of-contents-introduction.
12. Centers for Medicare & Medicaid Services. Behavioral Health Services. Available from: https://www.medicaid.gov/medicaid/benefits/behavioral-health-services/index.html
13. Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data. Available from: https://aspe.hhs.gov/standardization-querying-data-quality-metrics-characteristics-electronic-health-data.
14. Centers for Medicare & Medicaid Services. DQ Atlas. Available from: https://www.medicaid.gov/dq-atlas/welcome.
15. Task Force on Research Specific to Pregnant Women and Lactating Women. Report Implementation Plan. Available from: https://www.nichd.nih.gov/sites/default/files/inline-files/PRGLAC_Implement_Plan_083120.pdf.
16. Suarez EA, Maro JC, Hague C, et al. Prenatal and Congenital Syphilis in the US: Characterizing Screening and Treatment. Sentinel Initiative; 2024. Available from: https://www.sentinelinitiative.org/sites/default/files/documents/TMSIS_Task_4_Protocol_v2.0_1.pdf.
17. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. Available from: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification.
18. U.S. Department of Health and Human Services, Office of the National Coordinator for Health Information Technology. National Health IT Priorities for Research: A Policy and Development Agenda. Available from: https://www.healthit.gov/topic/scientific-initiatives/national-health-it-priorities-research-policy-and-development-agenda.
19. Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Building the Data Capacity for Patient-Centered Outcomes Research: The 2021 Annual Report. Available from: https://aspe.hhs.gov/reports/building-data-capacity-pcortf-2021-annual-report.
20. Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Building the Data Capacity for Patient-Centered Outcomes Research: United State, 2019: Available from: https://aspe.hhs.gov/standardization-querying-data-quality-metrics-characteristics-electronic-health-data

21. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;4(1):18. doi:https://doi.org/10.13063/2327-9214.1244

22. 1.Rai A, Nam YH, Mendelsohn AB, et al. Utilization, user characteristics, and adverse outcomes of insulin glargine originators and follow-on drug in patients with diabetes in the United States. Journal of Managed Care & Specialty Pharmacy. 2023;29(7):842-847. doi:https://doi.org/10.18553/jmcp.2023.29.7.842

23. 0844: Effectiveness and Safety of the Recombinant Zoster Vaccine in Patients ≥18 Years of Age with Systemic Lupus Erythematosus or Multiple Sclerosis. ACR Convergence, 2024. Published November 16, 2024. Accessed November 8, 2024. https://acr24.eventscribe.net/fsPopup.asp?PresentationID=1509420&mode=presInfoasdf

24. Reagan-Udall Foundation. Master protocol and parallel approach to analyze angioedema in patients with heart failure identified in an integrated care delivery system compared to administrative claims. Society for Epidemiologic Research, 2021 Annual Meeting, June 22-25, 2021. Available from: https://www.reaganudall.org/sites/default/files/2021-08/20210602%20SER%20Poster%20PA%20ACEi.pdf

25. KFF. Status of State Medicaid Expansion Decisions: Interactive Map. Available from: https://www.kff.org/affordable-care-act/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/