MINI-SENTINEL COORDINATING CENTER DATA CORE

YEAR 1 COMMON DATA MODEL (CDM) REPORT

REPORT OF DATA CORE ACTIVITIES, OCTOBER 2009 – SEPTEMBER 2010

Prepared by: Lesley H. Curtis, PhD,[1] Mark G. Weiner, MD,[2] Nicolas U. Beaulieu, MA,[3] Robert Rosofsky,[4] Tiffany S. Woodworth, MPH,[3] Denise M. Boudreau, PhD,[3] William O. Cooper, MD, MPH,[4] Gregory W. Daniel, MPH, PhD,[5] Vinit P. Nair,[6] Marsha A. Raebel, PharmD,[7] Jeffrey S. Brown, PhD[3]

Author Affiliations: 1. Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, USA. 2. Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA. 3. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA. 4. Health Information Systems Consulting LLC, Boston, MA, USA. 5. Group Health Institute, Seattle, WA, USA. 6. Department of Pediatrics and Preventive Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA. 7. HealthCore, Inc., Wilmington, DE, USA. 8. Humana, Louisville, KY, USA. 9. Kaiser Permanente Colorado Institute for Health Research, Denver, CO, USA.

March 16, 2011
Revised June 2011
Revised October 2011

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Coordinating Center Data Core

Report of Data Core Activities, October 2009 – September 2010

I. INTRODUCTION 4

A. OVERVIEW OF THE MINI-SENTINEL PROGRAM 4
B. MINI-SENTINEL SCIENTIFIC OPERATIONS CENTER 4
1. *Responsibilities of the Data Infrastructure Division* *4*
C. MINI-SENTINEL COORDINATING CENTER DATA CORE 5
1. *Overview* *5*
2. *Roles and Responsibilities* *5*
3. *Members* *6*
4. *Members' Terms and Selection* *6*
5. *Data Partners* *6*
D. DISTRIBUTED DATA APPROACH 6
1. *Need for a Common Data Model* *7*

II. DEVELOPMENT OF THE MINI-SENTINEL COMMON DATA MODEL 8

A. MINI-SENTINEL COMMON DATA MODEL GUIDING PRINCIPLES 8
1. *Initial Priorities and Approach* *9*
B. DATA PARTNER DATA AVAILABILITY 9
1. *Claims/Administrative Data* *9*
2. *Electronic Health Records (EHR)* *10*
3. *Registries* *12*
C. OVERVIEW OF EXISTING COMMON DATA MODELS 13
1. *HMO Research Network Virtual Data Warehouse* *13*
2. *Vaccine Safety Datalink* *14*
3. *Meningococcal Vaccine Study* *14*
4. *Post-licensure Rapid Immunization Safety Monitoring (PRISM)* *15*
5. *Observational Medical Outcomes Partnership (OMOP)* *16*
6. *Other Data Models Reviewed* *16*
7. *Summary of Lessons Learned from Implementation of Existing CDMs* *17*
D. DEVELOPMENT OF SPECIFICATIONS 17
1. *Drafting Process* *17*
2. *Review and Revision Process* *17*
E. MINI-SENTINEL COMMON DATA MODEL V1.1 19
1. *Overview and Structure* *19*
2. *Advantages* *21*
3. *Disadvantages* *22*

III. IMPLEMENTATION OF THE MINI-SENTINEL COMMON DATA MODEL 22

A. MINI-SENTINEL COMMON DATA MODEL IMPLEMENTATION AND DOCUMENTATION 22
B. DATA QUALITY ASSURANCE AND CHARACTERIZATION 23
1. *Overview* *23*
2. *Data Characterization Specifications* *24*
3. *Development and Testing* *27*
4. *Implementation and Reporting* *27*
5. *Additional Information* *28*

C. Technical Assessment 28
D. Initial Queries 29
1. *Pediatric Distribution in Mini-Sentinel Distributed Database 29*
2. *Acute Myocardial Infarction Protocol Development and Validation 29*
E. Modular Programs 29
F. Distributed summary table Query Tool and Portal 30
G. Lessons Learned 33

IV. FUTURE WORK 34

V. REFERENCES 36

VI. APPENDICES 38

A. Appendix A: EHR Discussion Outline 38
B. Appendix B: Mini-Sentinel Common Data Model Initial ETL Report Template 44
1. *Data Partner Information 44*
2. *SAS Technical Environment 44*
3. *Preliminary Assessment of Key Variables 45*
4. *Report on Mini-Sentinel Common Data Model ETL 48*
5. *Example Report (Enrollment) 53*
C. Appendix C. Preliminary ETL Report Summary Results* 54
D. Appendix D: Example of Data Checking Report Questions Identified in Year 1 56

I. INTRODUCTION

A. OVERVIEW OF THE MINI-SENTINEL PROGRAM

Mini-Sentinel is a pilot program sponsored by the U.S. Food and Drug Administration (FDA) as a part of its Sentinel Initiative to inform and facilitate development of a fully operational active surveillance system for monitoring the safety of FDA-regulated medical products, i.e., the Sentinel System. Mini-Sentinel is a major element of the Sentinel Initiative, FDA's response to a Congressional mandate to create an active surveillance system using electronic health data for 100 million people by 2012.

Initially, the Mini-Sentinel program will focus on three major types of activities: (1) prospective evaluation of accumulating experience about specific medical products and outcomes; (2) evaluation of the impact of FDA actions (e.g., labeling changes) on medical practice and health outcomes, and (3) rapid assessment of available data in response to FDA questions about specific medical products and outcomes.

A wide range of Collaborating Institutions enable access to data environments and provide other resources to support meeting the epidemiologic requirements of Mini-Sentinel. In addition, representatives of the Collaborating Institutions provide ongoing scientific, technical, and methodological expertise by participating in the Planning Board, the Safety Science Committee, the three Mini-Sentinel Coordinating Center Cores (Data, Methods, and Protocol), project-specific workgroups, and other developmental activities.[i]

B. MINI-SENTINEL SCIENTIFIC OPERATIONS CENTER

The Mini-Sentinel Scientific Operations Center oversees the data infrastructure and epidemiologic aspects of the overall program. It supports the scientific work of the Methods, Protocol, and Data Cores and all Mini-Sentinel project workgroups. The Scientific Operations Center is the central point of contact for the FDA and all Collaborating Institutions regarding scientific aspects of Mini-Sentinel.

The Data Infrastructure Division oversees data development and data source documentation, as well as evaluation implementation activities of Mini-Sentinel. Individuals working within this Division possess expertise in database design, implementation, and analysis. Data Infrastructure Division staff are members of the MS Data Core and support and work closely with the FDA, the Data Core, and Data Partners on these Mini-Sentinel activities (see **Figure 1**).

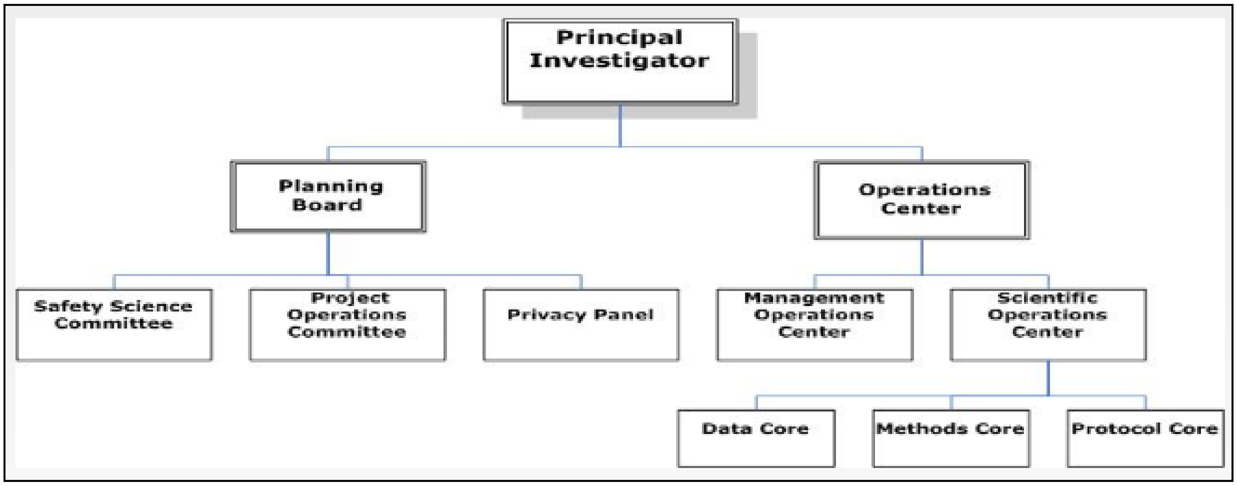
1. Responsibilities of the Data Infrastructure Division

- Coordinate and support the activities of the Data Core
- Coordinate and oversee development and implementation of the Mini-Sentinel distributed data approach and common data model
- Document data sources and characteristics

[i] For additional information, please see www.mini-sentinel.org.

- Assess data quality
- Develop reusable analytic tools
- Develop standard operating procedures for writing distributed programs
- Coordinate Mini-Sentinel data activities and projects to ensure use of available tools and adherence to programming standards
- Lead programming to support workgroups and analyses, as necessary
- Develop and manage Mini-Sentinel public website and private secure communications systems

Figure 1. Mini-Sentinel Coordinating Center



C. MINI-SENTINEL COORDINATING CENTER DATA CORE

1. Overview

The Mini-Sentinel Coordinating Center Data Core leads development and implementation of the Mini-Sentinel Common Data Model (MSCDM), distributed data approach, and related data standards and quality measures. The Data Core establishes additional workgroups as needed and interacts regularly with the Methods and Protocol Cores. A key responsibility of the Data Core is to facilitate communication across the Data Partners and manage the creation of the Mini-Sentinel Distributed Database, the data held and maintained by the Data Partners in the MSCDM format. The Data Core also serves as the main conduit for communication among Data and Academic Partners, project workgroups, and other parties interested in data-related aspects of Mini-Sentinel activities.

2. Roles and Responsibilities

- Develop, implement, and manage a scalable and extensible common data model to meet the needs of Mini-Sentinel
- Incorporate national data standards, as appropriate, into development of the MSCDM and data analysis
- Create and update Mini-Sentinel distributed datasets that conform to the MSCDM
- Establish and implement data quality measures
- Lead data development strategic planning

- Establish data workgroups
- Oversee and review data workgroup activities
- Develop, coordinate, and conduct data-related reviews and training for the FDA and Mini-Sentinel affiliate organizations
- Collaborate with Methods Core, Protocol Core, Operations Center, and FDA staff
- Communicate with external stakeholders as directed by FDA

3. Members

- Data Core Leaders
- Scientific Operations Center Director
- Data Infrastructure Division Deputy Director
- Representatives from each Data Partner
- Representatives from FDA
- Additional analytical and technical staff as needed

4. Members' Terms and Selection

Member terms are one year and are renewable. Data Core Leaders are selected by the Mini-Sentinel Principal Investigator and approved by the Planning Board. Data Partners and FDA representatives are chosen by their respective institutions.

5. Data Partners

Mini-Sentinel Data Partners involved in the initial implementation of the MSCDM include HealthCore, Inc. (working with WellPoint data), the HMO Research Network, Humana, Kaiser Permanente Center for Effectiveness and Safety Research, and Vanderbilt University (working with Tennessee Medicaid data). These Data Partners have access to the data elements needed to contribute to Version 1 of the MSCDM, such as health plan administrative and claims data.

The Mini-Sentinel includes other Collaborating Institutions that have access to other data sources of interest for medical product safety surveillance, including laboratory data, electronic health record (EHR) data, inpatient systems, and disease and device registries. Efforts to incorporate these data areas into the MSCDM will be the focus of activities in subsequent years.

D. DISTRIBUTED DATA APPROACH

In principle, the goals for the Mini-Sentinel program could be accomplished by implementing either a distributed model or through creation of a large centralized data repository.[2] A centralized system stores all patient and clinical data in one central database that is accessible to all authorized users. In the centralized model, all data partners (e.g., health plans, medical clinics) send their data to a central location; all the network data are physically stored together outside the physical control of the data partner. In this model, data analyses are conducted by the entity that controls the data warehouse. In a distributed, or decentralized, system each data partner maintains physical control of their data behind their firewalls, protected by their security processes and rules. Analysis in a distributed model can involve distributing the analyses (i.e., executable programs) to the data partners for processing and return or distributing a protocol for local interpretation, programming, implementation, and return.

A mixed model can be used on a case-by-case when evaluations require person-level intermediate analytic datasets, for example, when performing multivariate analyses.[2,3] A mixed model uses a distributed approach for all analyses or cohort specifications that can be conducted in a distributed manner (e.g., incidence rates, safety surveillance, identification of specific cohorts) and only transfers person-level data for combined analysis (e.g., case-control or cohort approach) if necessary. If so then only the minimum necessary data are transferred which typically includes 1 row per person with highly summarized aggregate information such as age in an age range, number of prior hospitalizations, and total days exposed to a treatment.

Mini-Sentinel uses a distributed data approach in which Data Partners maintain physical and operational control over electronic data in their existing environments.[1-7] By allowing data partners to maintain control of their data and its uses, the distributed model avoids or reduces many of the security, proprietary, legal, and privacy concerns of data partners, including those related to the Health Insurance Portability and Accountability Act (HIPAA). This approach also incorporates the need to have local content experts maintain a close relationship with the data. For example, only a local expert can easily and effectively trouble-shoot an unexpected finding or anomaly. The distributed model also allows data partners to accurately assess, track, and authorize query requests, or categories of requests, on a case-by-case basis, and ensure that only the minimum data necessary are shared with MSOC or FDA.

1. Need for a Common Data Model

There are two options for conducting analyses in a distributed data environment:

1. Create an evaluation or assessment protocol or other written instructions and ask each data partner to implement the protocol locally by creating the analytic programs, or

2. Create an evaluation or assessment protocol, centrally develop the analytic code (i.e., query) and distribute the code to each data partner to run against the data they have stored in a common format.

Mini-Sentinel has chosen the second approach because it reduces the potential for analytic inconsistency, ensuring that the results from the data partners are comparable. Comparability is achieved through use of a common data model (CDM) and the uniform implementation of shared programs that analyze the distributed data. This approach avoids differences in interpretation of protocols by separate data partners and it maximizes efficient use of analyst effort. Additionally, it is difficult to identify implementation differences without detailed and timely investigation and data checking.

Most implementations of common data models require each data partner to transform its data into a common model, either virtually or physically. Physical transformations are referred to as extract, transform, and load (ETL) processes. Virtual transformations use an intermediate software layer that maps local data partner concepts and data elements to the common model and require either real-time transformation of the data for querying or transformation of the query to allow it to run on the local data warehouse.

This distributed approach has been used successfully in several multi-institutional projects including the Vaccine Safety Datalink (VSD) project, the HMO Research Network (HMORN), the Meningococcal

Vaccine Safety study, the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) project focusing on H1N1 vaccine safety, and the Observational Medical Outcomes Partnership (OMOP). Each of these projects implemented a common data model to allow a single analytic program to run identically in each data environment.

Additional details of different database models can be found in the report titled "*Evaluating Possible Database Models to Implement the FDA Sentinel Initiative".*[3]

II. DEVELOPMENT OF THE MINI-SENTINEL COMMON DATA MODEL

Development of the MSCDM was based on Mini-Sentinel Common Data Model Guiding Principles, the FDA Sentinel Initiative Reports,[5] and an evaluation of other common data models used or proposed for distributed pharmacoepidemiology studies of medical products. Revisions and enhancements to the MSCDM are expected, including the addition of clinical information obtained from EHRs, laboratory systems, and registries.

A. MINI-SENTINEL COMMON DATA MODEL GUIDING PRINCIPLES

The design and implementation of the MSCDM strives for a high level of cross-institutional and longitudinal consistency and requires that data comparable in format and meaning are stored at all sites. The following principles guide the development, maintenance, and use of the MSCDM within the context of the Mini-Sentinel distributed data approach:

- Data Partners have the best understanding of their data and its uses; valid use and interpretation of findings requires input from the Data Partners.
- Distributed programs should be executed with no site-specific modification of the analytic code after appropriate testing.
- The MSCDM accommodates all requirements of Mini-Sentinel surveillance activities and may evolve to meet FDA objectives.
- The MSCDM is able to incorporate new data types and data elements as needs indicate.
- Development of the initial MSCDM and all enhancements require input from and acceptance by the Mini-Sentinel Data Partners.
- Documentation of Data Partner specific issues and caveats that may impact use and interpretation of the data is necessary for the effective operation of Mini-Sentinel activities.
- The MSCDM design is transparent, intuitive, well-documented, and easily understood by analysts, investigators, and stakeholders. It is easy for experienced analysts and investigators to use; special skills or knowledge beyond those commonly found among pharmacoepidemiologists and professional analytic staff is not necessary.
- The MSCDM leverages evolving healthcare data standards.

- The MSCDM captures values found in the source data. When necessary, mapping to standard vocabularies is transparent. Validated mappings should be used whenever available.
    - With the exception of specific laboratory values (e.g., absolute neutrophil count [ANC] and international normalized ratio [INR]), calculated variables are not included in the MSCDM.
- Only the minimum necessary information should be used and shared with authorized staff of the Mini-Sentinel Coordinating Center.
- Personal identifiers are never shared with the MSOC or FDA.
- Data Partners may include “site-specific” variables in their implementation of the MSCDM.

1. Initial Priorities and Approach

The overall initial goal was to build the foundation for Mini-Sentinel to support active surveillance and the ability to respond to FDA queries. The aggressive time line, in turn, required a highly focused yet readily extensible model. Initial functionality relies on claims and administrative data with additional functionality and data areas to be added in subsequent years. Version 1 of the MSCDM:

- Reflects the guiding principles;
- Focuses on claims and administrative data;
- Leverages the cumulative experience of the Data Partners;
- Relies on existing and standardized coding standards (e.g., ICD-9-CM, HCPCS/CPT, and NDC) to minimizing the need for ontology mapping; and
- Is compatible with claims-based components of existing multi-site CDMs.

B. DATA PARTNER DATA AVAILABILITY

The Mini-Sentinel Scientific Operations Center is pursuing a 3-stage approach to building and then incorporating data into the MSCDM. The first priority, accomplished in Year 1 of the contract, is the inclusion of claims and administrative data to enable analyses of medical product exposures and outcomes including specific diagnoses, procedures, or significant events such as a hospitalization. The second and third stages of development will be targeted in subsequent years. The second priority is the inclusion of clinical data (e.g., laboratory results, vital signs, immunization data, and smoking status) from EHRs and other sources. The third priority is incorporation of data from freestanding registries. Registry information will be most useful when it is possible to link the registry information to health plans’ additional exposure and outcome data, since their data are usually limited to the clinical focus of the registry. The FDA Sentinel Initiative contract reports (http://www.fda.gov/Safety/FDAsSentinelInitiative) informed the data availability assessment.

1. Claims/Administrative Data

Administrative and claims data from public (e.g., Medicare) and private (e.g., health insurers) payers are widely used for medical product safety studies. The characteristics, advantages, and disadvantages of these data for medical product safety surveillance studies are well understood. The basic data elements available in administrative and claims databases are known and relatively consistent across payers due to standardized billing submission systems. However, variation in administrative and claims data coding and availability across Mini-Sentinel Data Partners exists, and a keen understanding of local nuances is essential. For example, some Mini-Sentinel Data Partners have access to data elements not available to other Mini-Sentinel partners, and those unique data elements may prove valuable for medical product safety surveillance activities. Additionally, some but not all Data Partners have information on data elements such as the primary reason for an emergency department visit, diagnosis-related group (DRG) for an inpatient stay, and facility location. Some Partners have local codes for drug, diagnosis, and procedure that can help identify exposures and outcomes. To assess available administrative and claims data and cross-site variation, each Mini-Sentinel Data Partner provided comprehensive information regarding their internal data systems and capabilities as part of development of the MSCDM. Section D: Development of Specifications describes the process for identifying and standardizing the claims and administrative data across Data partners.

2. Electronic Health Records (EHR)

We conducted a confidential assessment of selected EHR information available from the Mini-Sentinel Data Partners. The purpose was to develop an initial inventory of the Data Partners' ambulatory and hospital electronic health record (EHR) systems in order to understand the breadth and depth of information available and some of the limitations of the data available for queries on medical product safety. The inventory will be used to improve understanding of the kinds of data available for use and help guide discussion with FDA and the Data partners regarding possible future directions for incorporating clinical data into the MSCDM.

The initial EHR assessment involved three phases. In the first phase, a template for the database profile was developed and refined in consultation with the Mini-Sentinel Scientific Operations Center. In the second phase, representatives for the Data Partners provided feedback on the template through a series of three group conference calls. Further revisions were made and the template was disseminated to the Data Partners for completion. In the third phase, confidential responses were returned, screened for completeness and compiled into a single spread sheet file constituting the final inventory.

The inventory contains information on the number of patients, breadth and the longitudinal scope of the data, the potential for data linkage of systems across settings and sites, and availability of clinical data subsystems (e.g., cardiology, radiology, pharmacy, oncology, pulmonary, pathology), laboratory data, vital signs, social history, over-the-counter (OTC) medications, nutritional supplements, and medical devices. Information on any previous validation studies and the use of standardized clinical vocabularies are also included. Please see **Appendix A** for additional information. Based on agreement with the Data Partners, information collected on detailed clinical data availability is for internal use by MSOC and FDA to help guide the development of additional data areas within the MSCDM.

Overall, most Data Partners have access to at least some clinical data for all or a portion of their populations and can link information for members across data systems. Each of the partners with clinical data has previously used the data for public health surveillance and research purposes. As expected, there is variability in the range of data elements available, the years the data are available, and the ease at which the data can be extracted and used.

As part of our background Data Partner inventory, we also specifically assessed data availability within Data Partners with inpatient data. Each Data Partner with access to inpatient data was asked questions such as:

- Number of inpatient facilities with full access to data
- Number of admissions in 2008 or most recent year available?
- Number of adult (≥18 years), pediatric (3-17 years), and neonate admissions
- Ability to retrieve full text inpatient medical records
- Ability to link to ambulatory records
- Ability to identify medical products/procedures
    - Drugs administered in operating rooms
    - Drugs administered in ICUs
    - Drugs administered in EDs
    - Drugs administered in dialysis units
    - Blood products administered
    - Allografts administered
    - Implants administered
        - type
        - manufacturer
        - model
        - serial number

Overall, the Mini-Sentinel Data Partners have access to detailed inpatient hospital data from 88 teaching and community hospitals. In summary:

Kaiser Permanente owns and manages 35 hospitals, with 600,000 admissions per year. Inpatient dispensing data have been captured since the early 1990s, with implementation of an inpatient EMR in 2009. Inpatient data capture now includes all interventions in all departments, including drugs, blood products, and allografts.

Oregon Health & Science University's (OHSU) hospitals care for 29,000 inpatient admissions per year.

Partners Healthcare System (PHS) in Boston includes data from 4 hospitals (Brigham and Women's Hospital, Massachusetts General Hospital, Newton-Wellesley Hospital, and Faulkner Hospital). The Partners Research Patient Data Repository (RPDR) contains rich clinical data collected via their electronic medical record system. The RPDR contains information since 1988, including 120,000 inpatient admissions in 2008.

The **Pediatric Health Information System** (PHIS) database contains ~3 million discharges from 43 non-profit tertiary care pediatric hospitals affiliated with Child Health Corp. of America (www.chca.com/owner_hospitals).

The **University of Illinois Medical Center at Chicago** (UIC) has 9 years of detailed inpatient medication and laboratory data for approximately 20,000 yearly hospital admissions.

The **University of Pennsylvania Health System** (UPHS) includes 3 acute care hospitals: the Hospital of the University of Pennsylvania, the Presbyterian Medical Center, and Pennsylvania Hospital. Electronic clinical information for about 84,000 admissions per year has been utilized in multiple studies by Mini-Sentinel investigators.

3. Registries

A broad range of device and disease registries and registry expertise is accessible through Mini-Sentinel Collaborating Institutions. These include orthopedic registries of total shoulder, total hip, and total knee replacements, cardiovascular registries focused on heart failure, acute coronary syndromes, coronary artery disease, and stroke, and state-based immunization registries. Other registries of various sizes are maintained by the Mini-Sentinel Collaborating Institutions. At the direction of FDA, the Mini-Sentinel Scientific Operations Center will investigate the availability and content of specific registries.

As with the inpatient data inventory, our background Data Partner registry inventory included questions such as:

- Registry name and description.
- Describe data access to the registry (full, partial…)
- Linkage capacity to other data sources (Y/N)
- If so, what type of data sources could be linked to
- Inception of registry (year)
- Are new patients being entered
- Total number of patients (most recently available data)
- What exposures (e.g., outpatient medications, vaccinations; inpatient procedures) are collected
- What outcomes are collected (inpatient diagnoses/procedures, outpatient diagnoses/procedures)
- For devices, can you collect info on (Y/N):
  - Type of device
  - Manufacturer
  - Model
  - Serial #
- For medications/ biologics, can you also collect info on lot number (Y/N)

Overall, the Mini-Sentinel Data Partners have access to a variety of disease registries as summarized below.

The Duke Clinical Research Institute had a lead role in developing 6 national cardiovascular registries with the American College of Cardiology National Cardiovascular Disease Registry (ACC NCDR), American Heart Association (AHA), and the Society of Thoracic Surgeons (STS): ACC NCDR ACTION (~300,000 patients with acute coronary syndrome), ACC NCDR CathPCI (~2M patients with cardiac catheterizations for diagnosis of intervention), AHA Get-With-The-Guidelines (GWTG) CAD (~500,000 inpatients with coronary artery disease), AHA GWTG Stroke (~800,000 stroke inpatients), AHA GWTG HF (~300,000 inpatients with heart failure), and STS Database (~3 million adult cardiac surgery including coronary artery bypass, aortic valve repair, and mitral valve repair).

Kaiser Permanente's Inter-Regional Implant Registry has complete information since 2001 on cardiac and orthopedic implants, including heart valves, implanted cardiac defibrillators, coronary artery stents, and pacemakers, total knee replacements, total hip replacements, and anterior cruciate ligament (ACL) procedures. Three manufacturers have provided data on cardiac valves and defibrillators, totaling more than 100,000 devices, linkable to other KP data sources. Six KP regions also participate in the national Total Joint Replacement Registry, with ~85,000 total joint replacement procedures (~55,000 total knees, ~30,000 total hips).

Outcome Sciences, Inc. maintains multiple national registries such as stroke and Type 2 diabetes registries. The stroke registry with ~1 million patients includes medications, procedures, and complications during the first 30 days. Inpatient therapy such as tissue plasminogen activator can be evaluated for early adverse events, and linked to administrative datasets for long term outcomes. Outcome captures data in about 2,000 hospitals and several thousand physician offices, enabling rapid development of registries to verify new signals.

The **Weill Cornell CERT** maintains a growing 10,000 procedure registry of total joint replacement for hips, knees and shoulders, developed to study device safety and effectiveness. It features completeness of patient participation, data collection, clinical information, and a protocol for follow up, including both clinical and patient reported outcomes.

C. OVERVIEW OF EXISTING COMMON DATA MODELS

Below is an overview of several of the antecedent distributed data models, all of which rely on a distributed data model that protects the confidentiality of person-level data. This overview is not intended as a complete review of each of the models described, rather, the information is provided as background within the context of the Mini-Sentinel CDM development process. Each of the models was reviewed during development of the Mini-Sentinel CDM. In addition, several members of the Mini-Sentinel Coordinating Center and all of our Data Partners have been involved in the antecedent projects and have extensive experience in designing distributed data models and executing multi-institutional medical product safety projects using a distributed approach. Their knowledge and experiences in those models informed the development of the MSCDM (described below).

1. HMO Research Network Virtual Data Warehouse

The HMO Research Network (HMORN; http://www.hmoresearchnetwork.org) is a consortium that includes 15 U.S. health care delivery systems with integrated research divisions, offering a sample of geographically diverse, population-based health care data for more than 11 million people in the U.S. [8] The HMORN Partners have created several independent yet related networks based on the ready availability of data from multiple health plans within the network. For example, the HMORN Partners have created the HMORN Cancer Research Network funded by the National Cancer Institute, the Cardiovascular Research Network funded by the National Heart Lung and Blood Institute, and Center for Education and Research on Therapeutics funded by the Agency for Healthcare Research and Quality.

These multi-center projects increasingly use the HMORN's Virtual Data Warehouse (VDW), a distributed data network comprised of dataset standards and automated processes to facilitate multi-site research. Each health plan (site) maintains local data files that conform to VDW standards. These files are derived from administrative, claims, EHRs, and internal/external registries. The VDW data model includes enrollment, demographics, outpatient pharmacy dispensings, medical utilization (encounters, diagnoses, and procedures), vital signs, census information (using geo-coded data), death, and laboratory results tables. The HMORN VDW also supports a cancer registry table that standardizes cancer registry information across the sites. The death tables are based on the same tables in the Vaccine Safety Datalink distributed database described below. The data are updated at least annually; many sites update monthly. The VDW tables are designed to support a wide range of observational studies, including medical product safety, comparative effectiveness, health economics, outcomes research, cancer surveillance, and quality.

The HMORN VDW data model balances the desire to maintain source data granularity with the standardization necessary to facilitate efficient multi-site collaboration. The model maintains the granularity and local coding standards (e.g., ICD-9-CM, HCPCS, CPT) of the most important data elements and also allows sites to include "local variables" that are not included in the data model (e.g., prescription co-payment amount) but that improve the value of the files locally. The data files do not include calculated variables, but rather, the HMORN has developed a suite of standard programming macros that perform routine calculations (e.g., creation of a continuous eligibility periods and flexible medication treatment episodes) on an as-needed basis as specified by the a specific project.

The development and maintenance of the HMORN VDW demonstrates the ability of a large number of health plans to extract data into a common data model, use a distributed data model to conduct studies across a range of topics, and to update data frequently. The structure of the HMORN's oversight of the VDW – through committees responsible for the maintenance and enhancement of the VDW – and the data quality checking procedures employed by the HMORN also illustrate an approach for oversight of a data environment like Mini-Sentinel's.

2. Vaccine Safety Datalink

The Vaccine Safety Datalink (VSD; www.cdc.gov/vaccinesafety) Project is a collaboration between CDC's Immunization Safety Office (ISO) and eight HMORN health plans.[9-10] VSD investigators conduct active, "near real-time" surveillance (referred to as Rapid Cycle Analysis) and full scale retrospective epidemiologic studies of immunization safety. The VSD analyzes data through a distributed data model that was developed in 2002. As with the HMORN, the VSD's common data model specifies separate tables with defined structures for demographics, enrollment, immunizations, inpatient and outpatient diagnoses and procedures, and death/cause of death. Selected exposure and outcome tables are updated weekly, allowing near real-time surveillance for select vaccine adverse events.

The structure and format of the VSD tables are similar to the HMORN VDW data files. Major differences in the common data models include the VSD's capture of detailed information on vaccine administration extracted from electronic health record data, the separation of inpatient and outpatient diagnoses, the level of granularity in enrollment files, and the lack of a dispensing file.

With respect to Mini-Sentinel, the VSD project demonstrates the ability of health plans to use a distributed data model to conduct near real-time vaccine safety surveillance. Near real-time surveillance is a key element of the Mini-Sentinel project and the ongoing VSD activities are expected to inform implementation of active surveillance projects within Mini-Sentinel.

3. Meningococcal Vaccine Study

The Meningococcal Vaccine Study is a post-marketing epidemiological evaluation of Guillain-Barré syndrome (GBS) following administration of a new meningococcal conjugate vaccine, Menactra.[3] The study represents a new collaboration that uses the data and other resources of five health insurers with an analyzable population of approximately 50 million members from 2005 through 2008. A common data model is used to support the distributed analysis of a cohort of approximately 10 million adolescents who belong to five health plans.[3] Each of the five participating data partners extracted information from its data systems, and loaded the data into a common data model that includes 5 files: demographics, enrollment, vaccination, outcome, and other medical encounter data. These files are maintained by the data partner. The study coordinating center leads the development of analytic programs and distributes the programs to the study sites. The study sites execute the program and return tables, summary data and derived data elements for the common analysis. Each site is responsible for obtaining full text medical records from providers or medical facilities that submitted the claims associated with the diagnosis code indicative of GBS.

This study provides an additional example of a multi-site medical product safety study that successfully used a distributed data approach with a common data model. The study also successfully incorporated a process for requesting, abstracting, and adjudicating hundreds of medical charts from facilities throughout the country.

4. Post-licensure Rapid Immunization Safety Monitoring (PRISM)

The PRISM H1N1 project used large health plan claims databases to perform active near real-time surveillance of potential adverse events from the administration of H1N1 vaccines during the 2009-2010 influenza season.[11] Four of the five PRISM health plans are engaged with Mini-Sentinel. PRISM was initiated as a standalone activity, but has become a part of the Mini-Sentinel program and has been expanded to evaluate other non-influenza vaccines. PRISM included linkage of data from nine state immunization registries, also referred to as immunization information systems, to membership and claims data from five health plans, with adjustment for data latency.

Health plans extracted their data from their operational systems and transformed it into a common PRISM data model consisting of specially formatted statistical datasets in SAS. There were 5 data tables: demographics, vaccinations based on claims, vaccinations based on state immunization data, inpatient diagnoses, and outpatient diagnoses. All patient-level and encounter-level data tables remained at health plans, while aggregated data, without identifiers, were sent to the PRISM Coordinating Center for analyses.

Multiple data checking programs were executed by the health plans after each of their data updating cycles, in order to ensure quality of individual data structures, allowable values, and consistency across data values. Additionally, summary data profiles (counts and proportions) were generated and shared with the PRISM analysts. These profiles enabled checks on both “reasonableness” of the data within and between health plans, and provided counts that could be compared to results in aggregated files used in analyses.

This data model of the PRISM project demonstrated the ability of health plans to rapidly and regularly assemble data from their own files, and to link those with information from state immunization registries. Generation of aggregate files occurred approximately every two weeks, using statistical SAS programs developed by the PRISM coordinating center and distributed to the health plans.

5. Observational Medical Outcomes Partnership (OMOP)

The OMOP project (http://omop.fnih.org) has created a network of data contributors using a common data model appropriate for medical product safety studies. Review of the OMOP model was based on several OMOP documents (e.g., "Points to Consider in Developing a Common Semantic Data Model and Terminology Dictionary for Observational Analyses", the OMOP common data model specifications, the ETL mappings, and the vocabulary documentation) and feedback from Mini-Sentinel Partners engaged in the OMOP project.

The core structure of the OMOP CDM includes 7 tables that capture detailed information regarding demographics, enrollment (observation time), drug exposures (dispensing and prescribing), diagnoses, procedures, and medical encounters. The model also can capture other medical encounter information such as laboratory results. The 7 core OMOP CDM tables are similar to the HMORN VDW and Mini-Sentinel data tables.

The OMOP NETWORK includes health insurer administrative and claims data, and EHR data from integrated delivery systems and health information exchanges. These systems have different data capture mechanisms, often use different coding standards to define the same or similar concepts, and capture data at varying levels of granularity. For example, health plans typically use National Drug Codes (NDCs) to capture outpatient pharmacy dispensing and a "days supplied" and "amount dispensed" fields are used to calculate exposure days and treatment episodes. EHR systems often capture medication exposure and treatment episode information based on prescriptions written and number of refills recorded in the EHR. To address this type of mismatched data capture, OMOP mapped data from different sources to standardized data concepts. For example, OMOP mapped drug dispensings identified in claims and prescriptions identified in EHRs to a unified nomenclature based on drug ingredient (RxNorm-based Clinical Drugs and Branded Drugs; http://www.nlm.nih.gov/research/umls/rxnorm/). Original codes (NDCs, for example) are retained in the data model and standardized concepts are added as new variables. Also implemented in the data model is the mapping of ICD-9-CM diagnosis codes to SNOMED-CT codes and an OMOP-specific set of clinical concept identifiers.

In addition to the 7 core data tables, the OMOP CDM includes derived tables describing "drug eras" (equivalent in concept to treatment episodes) and "condition eras" based on diagnosis codes. For example, the drug era table includes a specification that bridges exposure gaps of less than 30 days into a single era, and a second specification that does not bridge exposure gaps. Creating a drug era for an exposure gap of 7 or 14 days requires creation of a new drug era table. Many of the OMOP programs rely on the era tables rather than the core data tables.

6. Other Data Models Reviewed

The Mini-Sentinel team also investigated several other distributed data models, including i2b2 (www.i2b2.org),[12] ePCRN (www.epcrn.bham.ac.uk),[13] and the Teradata (www.teradata.com) health care data model. The "Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative"[2] report to the FDA includes a description of these and several other data models.

7. Summary of Lessons Learned from Implementation of Existing CDMs

Combined, the antecedent distributed data models have demonstrated the viability of several operational approaches. Specific lessons learned include the following:

- *Data Distribution:* It is feasible for multiple Data Partners to assemble patient-level files of their own data, following a common data structure.
- *Distributed Control:* Health plans can retain complete control of their data while working toward common objectives that use identical computer programs to generate data checking reports and aggregated files.
- *Data Mapping:* It is necessary to carefully evaluate all coding schemes used by each Data Partner to ensure that variability is documented and addressed.
- *Throughput:* As a result of their technical environments and operational efficiencies, health plans vary in their ability to refresh data frequently.
- *Analytical Outcomes:* With common programs used for generating aggregated analytical data files from patient-level files that remain with the health plans, analytical imperatives can be met using a distributed model.

D. DEVELOPMENT OF SPECIFICATIONS

1. Drafting Process

The MSCDM Guiding Principles, review of other distributed models, and expected analytical needs were used to design the MSCDM. The Mini-Sentinel Data Partners experience with the HMORN VDW, VSD, and PRISM common data models led them to prefer that the MSCDM be consistent in structure and format with those models. The HMORN Data Partners noted that consistency with these models would minimize effort needed to participate in Mini-Sentinel and allow Mini-Sentinel to leverage the data characterization, quality checking, and data domain oversight and enhancement efforts they had previously undertaken.

2. Review and Revision Process

a. Data Partner Review and Comment

Each Data Partner was asked to provide specific feedback on each data element in every table of the proposed CDM, including information on the specificity of the variable definition and whether or not the Data Partner could populate the variable as specified. Data Partners gave feedback and recommendations on the contents, structure, and relationships of the MSCDM. There were approximately three months of review (late February through mid-May, 2010) by Data Partners and the Data Core staff. An example set of questions posed to Data Partners during their review:

- Can your site implement the proposal? What changes would you make and why?
- What effort will be required to implement the proposal at your site?

- Are the definitions of tables and variables specific enough?
- Are there other important data elements not included in the proposal?
- Do you have any other comments on the design?

Major issues addressed included the extent to which the Partners' data could be used to populate the various data elements of the MSCDM, technical considerations of the data transformations that would be required from source data to MSCDM format, and design considerations to maximize performance of running analytic programs. Data Partners identified many issues including challenges in uniquely identifying an "encounter," ascertainment of death outside of institutional settings, various coding sets for diagnoses and procedures, questions about transforming their data to the values in the CDM, inclusion of care provided to individuals who are not health plan members in the utilization tables, and inclusion of denied or rejected claims. These questions resulted in more precise definitions of the tables and of the variables included in the table.

Data Partners placed a high priority on a data model that minimized their need to routinely calculate derived variables or to maintain secondary data tables, such as drug eras. This was especially important in Mini-Sentinel because Data Partners must refresh their data frequently and quickly. Data Partners expressed concern about the effort required to maintain up to date complex clinical mappings. Specifically, the dynamic nature of the underlying data systems would likely require new mappings with each data refresh. The Mini-Sentinel Data Partners expressed a clear preference for a model that reflected as closely as possible the data stored in their source files in both granularity and concept, and to avoid inclusion of concepts that are not intrinsically part of the Data Partners' original data systems.

Decisions were based on the Mini-Sentinel Guiding Principles, balancing the needs of Data Partners with the needs of the Mini-Sentinel project. At the conclusion of this process, data partners verified that (1) the necessary data elements were available in source databases and (2) the requirements were consistent with their expectations.

b. FDA Review and Comment

In parallel with the Data Partner review, selected FDA staff provided input on the drafts of the CDM. FDA comments and requests centered on operational issues, ability to incorporate national standards for health data, interest in ensuring that the CDM captured specific types of data (e.g., race categories, types of facilities), how multiple diagnoses and procedures at a patient visit or hospitalization will be captured, and inclusion of all claims (whether or not they were approved or denied). Many comments related to the collection of drug dispensing information, such as knowledge about numbers of refills, routes of administration, availability of lot numbers, and NDC code changes.

c. Additional Review and Comment

FDA staff shared drafts of the MSCDM with various collaborators at OMOP and CMS's contractor, Acumen LLC (www.acumenllc.com). Both provided written feedback that was reviewed by the Mini-Sentinel Coordinating Center Data Core and discussed with FDA.

E. MINI-SENTINEL COMMON DATA MODEL V1.1

1. Overview and Structure

The MSCDM V1.1 includes 8 tables that represent specific data domains that focus on the initial target of the CDM, namely administrative and claims-type data necessary to accomplish medical product safety surveillance evaluations.[ii] Each table serves a specific purpose, and the overall structure is designed to facilitate data access while preserving the granularity and nature of the source data. For example, the data tables keep similar clinical concepts together, and whenever possible keep separate "data streams" separate so that tables can be updated individually at different intervals if necessary. For example, outpatient pharmacy dispensings are kept separate from other claims sources so that the pharmacy table can be updated without impacting other tables in the data model. Details of the tables and each individual variable are available at www.mini-sentinel.org: Overview and Description of the Mini-Sentinel Common Data Model V1.1. A unique person identifier is included in all tables to allow linkage across the tables and comprehensive view of patient care during an enrollment period. The unique person identifier is not a true identifier (e.g., SSN), but rather a health plan generated alpha-numeric string that is unique to each person in the data files. Each health plan maintains a crosswalk between the unique person identifier and the true identifier is retained by the Data Partner. The person identifier is unique within a health plan and is not shared outside of the health plan with either the MSOC or the FDA. The tables are briefly described below.

Enrollment. The ability to ascertain who is eligible to receive specific kinds of care at any particular time is a requirement for most Mini-Sentinel investigations. For many medical product safety evaluations, it is important to know when it is reasonable to expect the capture of relevant medical utilization, so that the absence of medical care is meaningful and not a result of non-membership. That is, confidence in the absence of care is often as important as the observation of a medical event.

The enrollment table contains records for all individuals who were health plan members during the period included in the data extract. The table includes the unique person identifier, the starting and ending dates of coverage, and flags for medical and pharmacy coverage. Patients can have multiple periods of coverage that are continuous or disjointed. Continuous periods of coverage are joined together into one period. For example, if a coverage period that ends on December 31 and is followed by another that begins on January 1, the two periods are joined. A change in any variable, such as the drug coverage flag, in the enrollment table generates a new record even if the coverage is continuous. Disjointed periods of coverage –those that are separated by more than 1 day - are listed as separate records. Data Partners are not required to "bridge" gaps of more than 1 day in coverage; when appropriate, bridging will be incorporated into analysis programs based on the specific needs of the evaluation.

Most Mini-Sentinel evaluations will use the enrollment table to verify that patients identified in other tables (e.g., exposed to a specific medication) are health plan members during the evaluation period.

[ii] As the MSCDM is revised, newer versions will replace the older documents. MSCDM V1.1 is the current version. MSCDM V1.1 incorporates clarifications and editing changes, the core model and variables did not change.

The table structure is a simplification of the HMORN VDW enrollment table structure and similar in structure to the other common data models evaluated.

Demographic. The demographic table includes unique person identifier, sex, birth date, race and an ethnicity marker. However, only a subset of the Data Partners collects meaningful race and ethnicity information. The demographic table includes everyone found in the Data Partner database and is not limited to members included in the enrollment table. For example, everyone in the enrollment and pharmacy tables must be in the demographic table, but the reverse is not true.

Dispensing. The dispensing table represents outpatient pharmacy dispensing captured by the Data Partners. Each outpatient dispensing picked up by the patient is captured in the table. The table includes a unique record that lists the unique person identifier, dispensed date, dispensed NDC (in 11 digit format), the days supply listed on the dispensing, and the amount dispensed on the dispensing record. Each record is a unique combination of the unique person identifier, dispensed date, and dispensed NDC. Data Partners are instructed to process source transactions to remove rollback transactions and other adjustments before populating the dispensing table. This typically requires summation of dispensing information by unique person identifier, dispensing date, and dispensed NDC. No negative days supplied or amounts dispensed appear in the table and no corrections are made for values that are “out of range” such as days supplied of 900 days.

Individual dispensings can be linked to create treatment episodes based on any algorithm or specification necessary for the evaluation. For example, dispensings with out-of-range values can be cleaned or removed, and treatment episodes can be created on a case-by-case basis depending on the specific drug dispensed, patient cohort or any other criteria as specified by the evaluation team.

Medications dispensed at discount pharmacies (e.g., WalMart, Target) may or may not be included in the table, depending on whether or not the pharmacy submits the claim to the health plan and whether the drug benefit includes dispensings at pharmacies external to the health plan. Similarly, the purchase of over-the-counter medications is only included in the dispensing table if the transaction is submitted via the pharmacy to the health plan (and this is rarely the case). An analysis of pharmacy dispensing data for 11 HMORN health plans found a OTC medications accounts for 2% to 9% of all outpatient dispensing between 2000 and 2007; although this rate of capture is likely to be a small portion of all OTC use.[14] Infused medications, vaccinations, and medications (e.g., injections) provided directly by medical providers are captured in the separate **procedures table** because those administrations are considered “procedures” within the existing medical coding nomenclature and are captured by the Data Partners in a separate data stream. A very small percentage (less than 0.1%) of outpatient dispensings represent NDCs for procedures.[14] Similarly, medications dispensed in the inpatient setting are captured in a separate data stream and are not included in the Dispensing Table.

Encounter. Each time a patient sees a provider in the ambulatory setting (including emergency department care), or is hospitalized, a record is entered into the encounter table. Each record within the table is a unique combination of person, admission/encounter date, provider, and care setting. For example, if a patient sees a primary care physician who then sends the patient to the emergency department, and is later admitted to a hospital, there are three records in the encounter table. Additional information on this table includes discharge date of the hospitalization, provider code, facility code, 3-digit provider zip code for the facility, Diagnosis Related Group assigned to the admission, the admitting source, the discharge status, and the discharge disposition.

Diagnosis. Each encounter, whether inpatient or ambulatory/outpatient, is associated with at least one diagnosis. Therefore, the diagnosis table is linked to the encounter table in a one-to-many relationship so that all of the associated diagnoses are recorded in the diagnosis table. The diagnosis table includes 1 row for every unique diagnosis recorded during an encounter. The table also includes a flag for whether the diagnosis was recorded in the primary diagnosis field on the encounter (applies only to care in the inpatient setting), an indicator for the care setting the diagnosis was recorded, and an indicator for the type of diagnosis code. This table structure - "long and thin" – facilitates searching for specific diagnosis codes in large tables.

The diagnosis table can be used to identify disease cohorts or health outcomes of interest. The structure makes it easy to apply cohort algorithms such as identifying patients with at least one inpatient diagnosis or two outpatient diagnoses of bipolar disease, or a primary inpatient diagnosis of stroke.

Procedure. Similar to diagnoses, each inpatient and ambulatory/outpatient encounter is associated with one or more procedures. Therefore, the procedure table is linked to the encounter table in a one-to-many relationship so that all of the associated procedures are recorded in the procedure table. The procedure table includes 1 row for every unique procedure recorded during an encounter. The table includes the unique person identifier, the procedure code, an indicator for the care setting in which the procedure was recorded, and the specific type of procedure recorded (e.g., ICD-9 CM, CPT-4, HCPCS). There are currently many coding standards used to record procedures including ICD-9 CM procedure codes, CPT-4 codes, and HCPCS codes, and the table allows capture of any existing or future coding standards. This table structure - "long and thin" – facilitates searching for specific procedure codes in large tables.

The procedure table can be used to identify patients who have undergone specific surgical procedures (e.g., hip replacement surgery), received certain outpatient infusions, or were administered a specific vaccination.

Death. The Data Partners have various mechanisms for acquiring information about an enrollee's death. If a patient dies while in the hospital, the death is recorded in association with a related discharge disposition. However, many patients die outside the clinical setting and the only clue to the death is the cessation of health utilization activity. Therefore to confirm the death, many of the Data Partners link to local (state) death registries to update the death status of their members. This update is performed relatively infrequently – about once a year for most Data Partners. As a result, a 2-year lag in death data is not uncommon. Within the death table, the death date is recorded, along with imputation method if the exact date is not known.

Cause of Death. Since each death can be associated with one or more contributing conditions, the death table is linked to a separate cause of death table that records diagnosis code reflecting the underlying condition, along with coding dictionary used, type of contribution to the death, and the source of the information.

2. Advantages

The MSCDM contains many of the data elements needed for medical product safety evaluations. The MSCDM has a flexible structure that enables multiple Data Partners to use their disparate source data systems to transform to a common format. This common format enables a set of computer programs to be created at a single location and run in a distributed fashion in the multiple locations of the Data Partners. The structure can accommodate new data domains, typically through the addition of new tables. For example, adding laboratory test results or vital signs will entail creating the structure and adding new tables to the existing model; there is no need to impact any of the other tables in the model. Importantly, the structure maintains the information found in the source data, does not require substantial mapping to external standards, and can accommodate nearly any algorithmic approach to defining exposures, outcomes, and evaluation eligibility criteria.

3. Disadvantages

Version 1.1 contains no clinical information about vital signs or diagnostic test results, such as laboratory, pathology, or imaging tests. Some of these will be added in the next version.

This common data model approach does not readily accommodate stand-alone electronic health records (i.e., electronic health records that are not linked to an insurer system) as those systems might lack enrollment information, and many use other coding standards.

III. IMPLEMENTATION OF THE MINI-SENTINEL COMMON DATA MODEL

A. MINI-SENTINEL COMMON DATA MODEL IMPLEMENTATION AND DOCUMENTATION

Each Data Partner implemented the MSCDM using an extract-transform-and-load (ETL) process designed for their source data. This process involved translating their local data source to the MSCDM structure and format. As expected, many implementation questions arose during the ETL process. Weekly Data Partner calls and one-on-one teleconferences were used to share questions with other Data Partners, Mini-Sentinel Operations Center, and Data Core. The Data Partners also used weekly calls to discuss ETL approaches and lessons learned. Data Partners participating in HMORN’s VDW worked together to create a single ETL process that they shared across sites. Most Data Partners created new Mini-Sentinel datasets (i.e., physical tables) and a few others created “views” (i.e., virtual tables)[iii] of their source data in the MSCDM format. Using “views” obviates the need to create separate physical files in MSCDM format for their entire membership, thereby saving storage space. For the partners that use “views”, distributed programs execute against a MSCDM view of their source data. For partners with “views” the distributed program will include all data available in the source files that are the target of the “views”, including recent data that may not have been data checked. Using “views” does not change the data checking process; the data checking programs execute on the data available at the time the program is executed.

[iii] Using a MSCDM “view” of the source data obviates the need to store and maintain two sets of tables – the MSCDM and the local source files. In practice, when a query runs against a view of the MSCDM, the host system executes a mapping to the MSCDM before executing the distributed code, often resulting in longer execution times. Data Partners that use “views” will at any point in time have information (e.g., the most recent utilization data) in their sources files that have not been checked by the MSOC but are nonetheless available for querying through the “views”. Based on the purpose of any specific query, if appropriate, the most recent “un-checked” utilization information can be excluded for a specific request.

Once the ETL was finished, each Data Partner completed a detailed ETL Report (**Appendix C**) that included information on the local processes used to transform their data into the MSCDM and information regarding any site-specific information for each of the tables and data elements. For each table the Data Partners provided extensive detail on how they created the tables (e.g., from which local sources) and if there were any issues that arose during the transformation. The types of source files vary across sites as does the point in the claims adjudication/medical utilization recording process at which the files are available to the Data Partners. This could introduce variation that is observable but beyond the control of the Partners. For example, some Partners may not have access to the raw enrollment files but rather a processed version that bridges enrollment gaps based on the health insurer's internal standards. In this event, the Data Partner would note that the file includes application of a local enrollment bridging rule. Other examples include local procedures for defining an inpatient episode of care and applying local business rules for "cleaning" of claims files.

The ETL report also included a series of high-level demographic (e.g., race and sex distributions) and data characteristics for review by the Data Partners. This information was designed to highlight initial errors or concerns with the transformation that would be immediately obvious to the Data Partners based on their knowledge of their local data. For example, the Data Partners know the total enrollment and race distribution of their populations, so any transformation errors would be immediately apparent upon review of the ETL report.

Completion and review of the ETL report identified several Data Partner specific instances of errors or concerns (e.g., a date range that was inconsistent with the source data) that were corrected before submission of the ETL report to the MSOC. The ETL was considered complete – but unchecked – once the Data Partner sent their completed ETL report to the MSOC. The MSOC combined the ETL Reports for review as shown in **Appendix C**. The information in the ETL Report and the combined table was used to identify potential problems with the ETL process and the MSCDM, and to target MSCDM areas for more comprehensive data checking.

B. DATA QUALITY ASSURANCE AND CHARACTERIZATION

1. Overview

Transformed data were checked through the use of standard programs developed by the Mini-Sentinel Operations Center and refined through feedback from the Data Partners. The programs generated a series of output tables for review. The data characterization programs were run by each of the Data Partners on their local implementation of the MSCDM. The programs generated tables and reports used by the Data Partners to identify data issues. The data review procedure included a series of steps culminating in detailed documentation of the data available at each Data Partner and an agreement on the next steps for data development, including required corrections to the ETL and planned revisions for the subsequent ETL. The specific steps were:

1. Development and Testing
   a. Develop data characterization approach and programming specifications (described in the document titled "*Data Quality and Characterization Procedures and Findings*" (document will be available at www.mini-sentinel.org/data_activities)
   b. Develop distributed code to implement the data characterization specifications
   c. Data Partner testing of data checking code, feedback to the MSOC

    d. Revision of distributed code based on Data Partner feedback
2. Implementation and Reporting
    a. Data Partner execution of data characterization code
    b. Review of data characterization output, revise ETL as necessary, re-run data characterization code
    c. MSOC review of data characterization output, within and across sites
    d. MSOC data characterization report provided to Data Partners for review and comment
    e. MSOC and Data Partners review and discuss the data characterization report, agree to any necessary changes and the timeline for changes
3. Acceptance of the year 1 ETL

2. Data Characterization Specifications

The Mini-Sentinel program will heavily rely on the comprehensiveness and quality of the data available in the Mini-Sentinel Distributed Database (MSDD). Prior to using the MSDD to answer queries, sufficient data characterization and review must be implemented for all tables and variables to characterize the data, identify anomalies, and ensure cross-partner data extract consistency.

To have a better understanding of the structure and information available from each Data Partner, the MSOC worked closely with each Partner to assess the quality and completeness of their MSDD data and to identify any caveats for use. To ensure the MSDD data meet quality expectations, the MSOC developed a series of measures to check data quality and to characterize the breadth and depth of the data available for querying. The specifications address areas such as missing data, invalid values, invalid date ranges, and internal inconsistencies. The design and the scope of the data characterization programs take into account the following considerations:

- The way in which Mini-Sentinel Data Partners access the administrative and claims data and the electronic health record information can vary across Data Partners, possibly leading to variation in data capture and completeness. .

- Since the MSOC does not have access to the MSDD datasets it is vital that tables created match the defined Mini-Sentinel requirements.

The data characterization programs written by MSOC staff and distributed to the Data Partners will be run on the MSCDM data (i.e., presented in the MSCDM data dictionary format). The data quality activities for Year One are organized into three levels, based on the type of checks being performed. A description of the data characterization approach and the findings will be available on the Mini-Sentinel website in a separate document titled: "*Data Quality and Characterization Procedures and Findings.*"

a. Level 1 Data Characterization

The Level 1 assessments review completeness and content of each variable in each file to ensure that the required variables contain data and conform to the formats specified by the MSCDM data dictionary. For each MSCDM variable, data characterization verified that data types, variable lengths, and SAS formats are correct and reported values are within the specified range. For example, in the demographic table the date of birth must be a SAS numeric data type, with a length of 4 bytes. Additionally, the date of birth must be in the range of January 1, 1885 through the date in which the demographic table was created. Categorical variables must include only the values specified in the data dictionary. The table below illustrates several of the Level 1 data characterization items for the dispensing table.

Table 1. Level 1 Data Characterization: Example for the Dispensing Table

	Variable Name	Rule	Error Code
1	PatID	Must be character data type	Dis1.1
	PatID	Must be non-missing	Dis1.2
	PatID	Must be left justified	Dis1.3
2	RxDate	Must be a SAS date value of numeric data type	Dis1.4
	RxDate	Must be of SAS length 4	Dis1.5
	RxDate	Must be non-missing	Dis1.6
3	NDC	Must be character data type	Dis1.7
	NDC	Must be exactly 11 characters in length	Dis1.8
	NDC	Must be non-missing	Dis1.9
	NDC	Must only contain digits from 0-9 (i.e., no space or other characters)	Dis1.10
4	RxSup	Must be a SAS date value of numeric data type	Dis1.11
	RxSup	Must be of SAS length 4	Dis1.12
	RxSup	Must be non-negative	Dis1.13
5	RxAmt	Must be a SAS date value of numeric data type	Dis1.15
	RxAmt	Must be of SAS length 4	Dis1.16
	RxAmt	Must be non-negative	Dis1.17

b. Level 2 Data Characterization

Level 2 characterizations assess the logical relationship and integrity of data values within a variable or between two or more variables within and between tables. For example, it is permissible for the unique personal identifier to occur more than once in the enrollment table, as there can be more than one span of enrollment for an individual. However, in the demographic table, the person identifier variable should occur once. Further, the person identifier variable in the enrollment table must have a corresponding value in the demographic table. This ensures that for all patients for whom enrollment spans are created, there is corresponding demographic information. The table below illustrates several of the Level 2 data characterization items for the enrollment table.

Table 2. Level 2 Data Characterization: Example for the Enrollment Table

	Variable Name	Rule	Error Code
1	PatID	Can occur more than once in the file	Enr2.1
	PatID	Must have a corresponding value in the demographic table	Enr2.2
2	Enr_Start	Must be earlier than or equal to Enr_End	Enr2.3
	Enr_Start	In combination with PatID, MedCov, and DrugCov, must occur only once in the file	Enr2.4
3	Enr_End	In combination with PatID, MedCov, and DrugCov, must occur only once in the file	Enr2.5

Level 1 and 2 data quality characterizations will generate a set of tables that are sent to MSOC for review. During Year One, the MSOC staff (1) manually inspected the level 1 and level 2 reports, (2) identified data anomalies and reported them back to Data Partners, and (3) began discussion of the potential for developing ranges of acceptable error threshold rates. We reported all anomalies to the Data Partners for discussion and response to determine whether the issue we identified could be fixed or if it was part of the underlying data. If necessary, a plan for remedying the anomalies was developed – this typically entailed a correction in the subsequent data extract – or the anomaly was documented so it wouldn't signal an alert in the next data checking process.

c. Level 3 Data Characterization

In contrast to the Level 1 and Level 2 data checks, the Level 3 data assessments "profile" the data, focusing on characterizations that do not have an expected outcome or True/False finding. . Rather, the expectation is for some level of inconsistency across partners and over time for some assessments and some level of consistency for other assessments. Periods of sharp increases or decreases in trends are also unexpected. These characterizations generate counts and proportions and show the spread of values within each relevant field across Data Partners and time. This profiling characterizes specific data fields for each Data Partner and aggregates the information for cross-institutional comparisons. The Level 3 data characterizations also evaluate trends to help identify data gaps and unusual patterns. Examples of trends include outpatient pharmacy dispensing per member per month, hospital admissions per member per month, total dispensing per month, and total encounters by encounter type per month. Other Level 3 data characterization topics include counts of procedures per encounter by encounter type and year and diagnoses per encounter by encounter type and year. This approach has been used successfully by the HMO Research Network, the Vaccine Safety Datalink, and other distributed networks to identify issues within their distributed databases. Several Level 3 data characterizations for the dispensing table are listed below.

- Overall table statistics
  - Number of records in the table
  - Number of unique PatIDs (includes number/percent with missing, if any)
- Distribution of dispensing date (RxDate)
  - Dispensings by month and year
- Average number of prescriptions per PatID
  - By year
- Distribution of days supplied (RxSup)

  - All years
  - Overall
- Distribution of dispensed amount (RxAmt)
  - All years
  - Overall

By examining the counts and proportions, both Data Partners and the Operations Center are able to ensure that the data are reasonable within Data Partners and consistent across Data Partners. For example, age in years is profiled in the following ranges: 0-1, 2-4, 5-9, 10-14, 15-18, 19-21, 22-44, 45-64, 65-74, 75+. If a Data Partner's Level 3 data showed an unusually large proportion of any one age range, this would indicate that there may be an issue with how the MSCDM was populated. Or, if the age proportions at one Data Partner are substantially different from the other Partners, it may indicate a difference in the underlying populations. The Level 3 data characterizations are designed to identify areas where variation within and across sites represents a potential concern to be further evaluated. Active participation from the Data Partners is essential to addressing unexplained variability. We note that this level of data check is not intended to find the "needle in the haystack" data anomaly, but rather to assess metrics that can be readily checked and flagged for explanation. Detailed, topic-specific data checking is required for every Mini-Sentinel query as review of specific data areas or patient cohorts may uncover anomalies not identified in the initial data checking activities.

3. Development and Testing

The Data Partners tested the data characterization code and raised several concerns regarding the efficiency of the programming, the number of output files generated, the transfer of potentially proprietary data, and the total volume of data initially requested by the MSOC. This feedback led to several rounds of revision and testing to arrive at a process acceptable to the MSOC and the Data Partners. These revisions included combining several output files, removing some of the frequencies by exact date, and incorporating a two-step checking process that obviated the need to send large files such as the frequency of every NDC, diagnosis, and procedure code to the MSOC. The two-step process involved Data Partners maintaining control of several large files (e.g., listing of every NDC with frequencies of use by year) generated by the data characterization programs that were later evaluated using a second set of distributed programs. The output of the second program generated summary findings that could be returned to the MSOC for review.

4. Implementation and Reporting

The data characterization code was executed by Data Partners after review of their ETL Summary Report. Several Data Partners identified errors upon review of the data characterization results and corrected the errors before sending any results to the MSOC. Upon receipt of the data characterization reports from the Data Partners, the MSOC assessed data quality and completeness within sites. Results of the data characterization activities were shared with the Data Partners. **Appendix D** provides a select list of data reporting issues identified during the data characterization process. The tables in the Appendix include findings from multiple Data Partners. Two companion documents - the "*Data Quality and Characterization Procedures and Findings Report"* and the "*Mini-Sentinel Distributed Database Year 1 Summary Report* "- provide details of the data checking and characterization activities and results. These reports will be available on the Mini-Sentinel public website (www.mini-sentinel.org/data_activities).

5. Additional Information

Data characterization and checking is a continuous process. MSOC will update the data checking process on a regular basis, including enhancing the level 1 to level 3 checks and adding new types of checks. Planned additions to the data checking activities include characterizations focusing on specific exposures (e.g., rate of beta-blocker use among members with a diagnosis of hypertension; rate of broken ankles by age, sex, and year; HPV vaccination rate by age and year) and outcomes (e.g., rate of inpatient AMI diagnoses and knee replacement surgery by age, sex, and year) for cross-site and longitudinal comparison. These characterizations will focus on specific exposures and outcomes to help further identify cross-site and longitudinal variation.

C. TECHNICAL ASSESSMENT

The MSOC performed a confidential assessment of the Data Partners' technical environments. The assessment, "The Mini-Sentinel Technology Assessment and Recommendations Project," was performed by an external consultant working under supervision of the MSOC. The assessment covered the technical environments, operational issues, institutional technical and operational constraints, and staff capacity for responding to requests. The table below provides an overview of the size of the data tables includes in the Mini-Sentinel Distributed Database.

Data Partners reported minimal serious impact of running Mini-Sentinel standard programs. The Technical Assessment report contains technical and organizational recommendations on how the MSOC can work more closely with the Data Partners to manage queries against the MSCDM. The technical challenges related to data formats/storage, local infrastructure upgrade cycles, data processing capacity, and SAS licensing costs. Organizational challenges related to team composition needed to effectively respond to queries, project scope, communication of activities and tasks, and standard operating procedures. The full confidential report was provided to FDA and the Data Partners.

Table 3. Summary of Current Mini-Sentinel Dataset Sizes

Data Partner	Total Rows	Largest Table Row Count	Estimated Size on Disk
1	2.5 B	1.2 B	350 GB
2	2.3 B	1.3 B	320 GB
3	944.3 M	517.9 M	131 GB
4	347.1 M	169.1 M	48 GB
5	231.1 M	110.6 M	32 GB
6	206.4 M	81.0 M	29 GB
7	191.0 M	63.2 M	27 GB
8	189.4 M	91.0 M	27 GB
9	157.9 M	79.8 M	22 GB
10	157.2 M	70.4 M	22 GB
11	94.4 M	42.1 M	13 GB
12	90.7 M	45.0 M	13 GB
13	82.7 M	37.5 M	12 GB

D. INITIAL QUERIES

1. Pediatric Distribution in Mini-Sentinel Distributed Database

In response to an FDA question, the MSOC distributed a query that collected counts of pediatric patients (aged ≤19 years of age as of January 1, 2009). The query accessed only the demographic table in the MSCDM, and the time frame of the evaluation period varied by Data Partner depending on the data available. At the time of the request the MSOC had not yet developed and tested standard coding for rapid response to FDA queries. Therefore, the request required new programming and testing. The program was distributed to all Data Partners, and executed successfully at each of the Data Partners that had implemented the MSCDM. All Data Partners responded within one week of receiving the distributed SAS program from the MSOC.

2. Acute Myocardial Infarction Protocol Development and Validation

In response to a request from the Mini-Sentinel workgroup developing a protocol for acute myocardial infarction and oral hypoglycemic agents, the Mini-Sentinel Operations Center developed, tested, and distributed queries (1) describing use of certain prescribed anti-diabetic agents and (2) estimating incidence of AMI hospitalizations among diabetic versus non-diabetic population. The query program was tested with two Data Partners, and then distributed to all potential participants in the planned protocol. The distributed program accessed Enrollment, Demographic, Dispensing and Diagnosis tables. All Data Partners responded within 11 days.

Similar assistance was provided to the AMI diagnosis validation workgroup to identify hospitalized acute AMI cases and the associated inpatient facilities. A distributed SAS program was designed in collaboration with both the AMI validation and surveillance workgroups, then prepared, tested, and finalized by the Mini-Sentinel Operations Center staff and two Data Partners. The query program was tested by two data partners, then distributed to all participating Data Partners. All Data Partners responded within 11 days.

E. MODULAR PROGRAMS

The Mini-Sentinel Operations Center has developed four modular programs to facilitate rapid response to common queries. Each program has several required input parameters (e.g., exposures and/or outcomes) and the output contains summary-level counts (e.g., number of members exposed to a drug, number of members with a specific diagnosis/condition) stratified by various parameters (e.g., age group, sex, year). Documentation for each of the modular programs is available on the Mini-Sentinel website (Data Activities), including a description of the program and the SAS code.

- ***Modular Program 1*** ***(medication use):*** characterizes the use of specified products (or groups of products) in the outpatient pharmacy dispensing table defined by National Drug Codes (NDC). Example: use of statins by age group and sex over time.

- ***Modular Program 2*** ***(medication use by condition):*** characterizes the use of specified products (or groups of products) in the outpatient pharmacy dispensing table defined by National Drug Codes (NDC) among individuals with a specified condition defined by ICD-9-CM diagnosis codes in the diagnosis table. Example: use of asthma medications among those with an asthma diagnosis by age group and sex over time.

- ***Modular Program 3*** ***(incident use and outcomes):*** evaluates the rate of specified outcomes (defined by ICD-9-CM diagnosis codes) among those with incident use of specified products (or groups of products) in the outpatient pharmacy dispensing table (defined by National Drug Codes (NDC) with or without a pre-existing condition (defined by ICD-9-CM diagnosis codes in the diagnosis table). Example: rate of inpatient AMI diagnoses after incident anti-diabetic product use among those with a diabetes diagnosis.

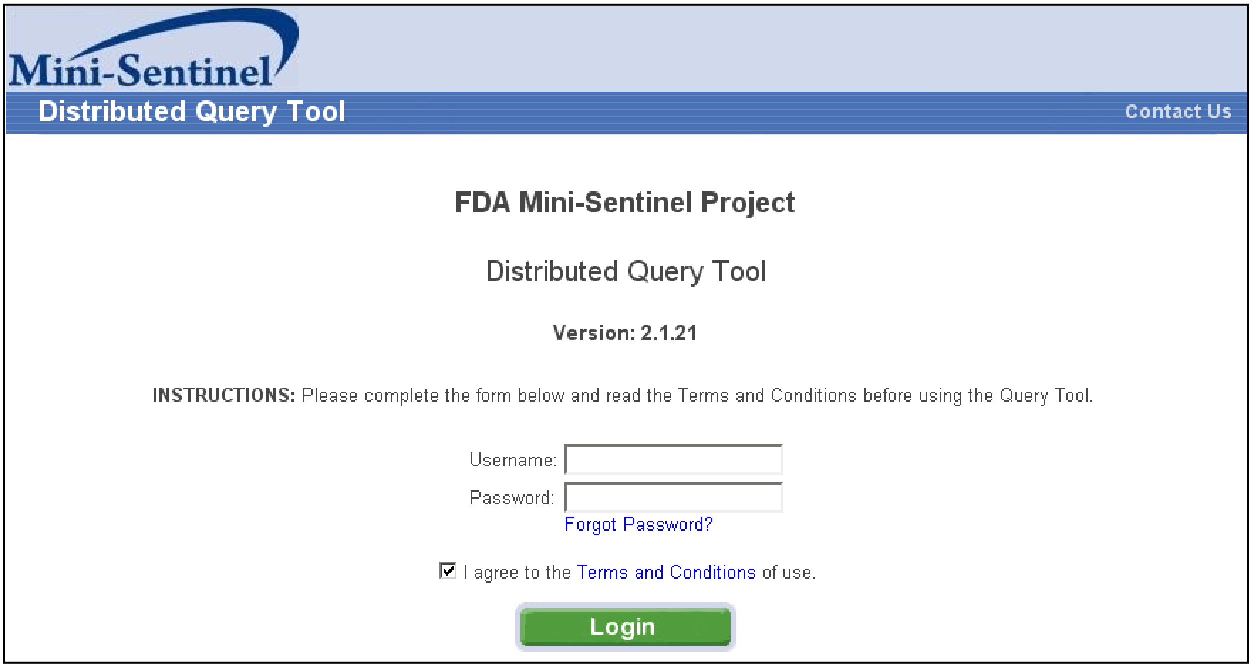
- ***Modular Program 4*** ***(concomitant medication use):*** characterizes concomitant use of products (or groups of products) in the outpatient pharmacy dispensing table (defined by National Drug Codes (NDC) among those with incident use of specified products with or without a pre-existing condition (defined by ICD-9-CM diagnosis codes in the diagnosis table). Example: characterization of atypical antipsychotic drug use among those with a diagnosis of depression and incident use of SSRI products.

These modular programs will be updated and revised based on feedback from FDA and the Data Partners, and additional modular programs will be developed on an ongoing basis in collaboration with FDA.

F. DISTRIBUTED SUMMARY TABLE QUERY TOOL AND PORTAL

The FDA Mini-Sentinel Distributed Summary Table Query Tool and Portal is allow Mini-Sentinel Operations Center to create and securely distribute “queries” to Data Partners and to have Data Partners review, execute, and securely return the results of those queries via a secure web Portal. Data Partners maintain control of their data, and they have the ability to review all queries before they are executed locally, and to review all query results before the results are transferred securely back to the Portal. The system is designed with multiple manual steps to allow partners to review and approve all requests according to their local processes and procedures. The system allows different levels of query automation that can be set at the discretion of the Data Partners. The network is hosted in a private cloud environment in a Federal Information Security Management Act of 2002 (FISMA) compliant TIER III data center.

Figure 2. Distributed Query Tool Login Page



FDA Mini-Sentinel Project

Distributed Query Tool

Version: 2.1.21

INSTRUCTIONS: Please complete the form below and read the Terms and Conditions before using the Query Tool.

Username:

Password:

☑ I agree to the Terms and Conditions of use.

The Mini-Sentinel distributed query tool (**Figure 2**: screenshot of the login screen above) currently allows rapid distributed querying of pre-processed summary tables. Using pre-processed summary tables speeds the querying process because it 1) obviates the need to access person-level data, thereby avoiding local privacy and patient confidentially data release authorization procedures; 2) allows use of a simple menu-driven querying tool interface that can be used by non-technical staff at MSOC; 3) allows non-technical Data Partner staff to execute and return results; and 4) avoids the need to specify, create and validate new SAS programming codes to answer simple questions. The expected response time for these queries is 48 hours. The system currently supports 9 query types that represent prevalence counts of diagnoses, procedures, and drug exposures. For diagnoses and procedures, the system also generates prevalence rates per 1000 enrollees, events per 1000 enrollees, and the number of events per person. For drug queries the system generates users per 1000 enrollees, dispensing per 1000 enrollees, days supply per dispensing, and dispensing per user. A sample 4 digit ICD-9 CM query result and a sample generic name query are provided below (using fake data). The Mini-Sentinel Distributed Query Tool Investigator's guide, a description of the Mini-Sentinel Summary Tables, and additional documentation is available on the Mini-Sentinel website and has additional details on the summary tables and a description of how to create and distribute queries.

Table 4. Sample aggregated result for a 4-digit ICD-9 CM code query using Mini-Sentinel Distributed Query Tool (fake data)

Age Group	Sex	Period	DXCode	DXName	Setting	Events	Members	Total Enrollment in Strata (Members)	Prevalence Rate (Users per 1000 enrollees)	Event Rate (Events per 1000 enrollees)	Events Per member
0+	All	2018	2502	DIABETES WITH HYPEROSMOLARITY	AV	9	9	63429	0.1	0.1	1
0+	All	2018	2503	DIABETES WITH OTHER COMA	AV	102	15	63429	0.2	1.6	6.8
0+	All	2018	2504	DIABETES WITH RENAL MANIFESTATIONS	AV	186	9	63429	0.1	2.9	20.7
0+	All	2018	2508	DIABETES W/OTH SPEC MANIFESTATIONS	AV	201	18	63429	0.3	3.2	11.2
0+	All	2018	2509	DIABETES W/UNSPECIFIED COMPLICATION	AV	384	30	63429	0.5	6.1	12.8

Table 5. Sample aggregated result for a generic name query using Mini-Sentinel Distributed Query Tool (fake data)

Age Group	Sex	Period	Generic Name	Disp.	Days Supply	Members	Total Enrollment in Strata (Members)	Prevalence Rate (Users per 1000 enrollees)	Dispensing Rate (per 1000 enrollees)	Days Per Dispensing	Days Per user
10-14	All	2018	AMITRIPTYLINE HCL	4	120	4	4636	0.9	0.9	30	30
10-14	All	2018	IMIPRAMINE HCL	16	440	8	4636	1.7	3.5	27.5	55
15-18	All	2018	AMITRIPTYLINE HCL	32	960	8	4104	1.9	7.8	30	120
15-18	All	2018	IMIPRAMINE HCL	4	48	4	4104	1	1	12	12
19-21	All	2018	AMITRIPTYLINE HCL	24	720	8	2664	3	9	30	90
22-44	All	2018	AMITRIPTYLINE HCL	648	22164	172	23672	7.3	27.4	34.2	128.9
22-44	All	2018	IMIPRAMINE HCL	24	660	8	23672	0.3	1	27.5	82.5
45-64	All	2018	AMITRIPTYLINE HCL	1908	67996	392	20588	19	92.7	35.6	173.5
45-64	All	2018	IMIPRAMINE HCL	148	7320	36	20588	1.7	7.2	49.5	203.3
65-74	All	2018	AMITRIPTYLINE HCL	72	3120	24	852	28.2	84.5	43.3	130
65-74	All	2018	IMIPRAMINE HCL	16	480	4	852	4.7	18.8	30	120
75+	All	2018	AMITRIPTYLINE HCL	156	7080	52	4028	12.9	38.7	45.4	136.2
75+	All	2018	IMIPRAMINE HCL	84	2520	12	4028	3	20.9	30	210

Additional query types and enhancements will be developed in subsequent years, including the addition of incident counts and new feature for comparing results across Data Partners.

G. LESSONS LEARNED

The development and implementation of the MSCDM in less than 8 months across disparate data partners was a formidable undertaking, and the resulting success can be attributed to the expertise, commitment, flexibility, and collegiality of the Mini-Sentinel Data Partners. Several important lessons were learned along the way.

- Definitions and coding of source data are not consistent across sites. The initial focus on claims-type and administrative data minimized, in many respects, the inconsistencies and challenges, yet source data still differed in important ways. Of note, among Data Partners that provide care as well as serve as insurers, data regarding outpatient clinic visits and pharmacy dispensings do not come exclusively from claims data. Other examples include the meaning of the first-listed diagnosis on a claim, the use of local coding schemes instead of J-codes for infused medications, local standards for claims processing such as bridging short enrollment gaps.

- Communication is essential. A weekly teleconference has been the primary conduit of information between the MSOC and Data Partners, and additional information is shared via email and one-on-one teleconferences. Although manageable for the earliest phase of start-up, the approach is not sustainable over time. Establishment of a secure web site to serve as conduit for the dissemination of key documents and requests is essential and has been implemented.

- Providing a forum for the Data Partners to interact with each other allowed valuable information to be communicated directly, and for expertise and knowledge to be shared across Partners.

- Each Data Partner has working definitions of concepts like a medical encounter and a membership period, and each has business rules for handling administrative and claims data. The local rules and definitions are not consistent across Data Partners, requiring substantial effort in defining and re-defining terminology during the calls and in the draft documents so all Partners were using terms consistently.

- Data Partners must be actively engaged in establishing timelines and workflow processes. Advance planning and prioritization is necessary to ensure that resources are available when needed, and barriers are identified early. Each of the Mini-Sentinel Data Partners are engaged in other activities that compete for their time and resources, making planning and scheduling of tasks a primary activity of the MSOC.

- The recurring one-year nature of the Mini-Sentinel base contract limits long-term activities that may be necessary for efficient development of new data areas. Some data projects will by necessity require over 12 months of planning and development. The one-year contracting limit also makes it difficult for the data partners to effectively plan for long-term resources related to the program.

IV. FUTURE WORK

Several additions and enhancements to the MSCDM are planned. First, the MSCDM will be expanded to include specifications for selected clinical and laboratory data. Second, with the FDA, the Mini-Sentinel Operations Center will identify relevant national data standards and controlled terminologies. A formal evaluation of the appropriateness of the identified standards and terminologies will be undertaken and a strategy for incorporating appropriate standards in the MSCDM will be developed. Third, the MSOC will develop and test new capabilities for extracting information from EHRs and incorporating it into the MSCDM. Adapting an existing open source EHR surveillance platform will facilitate vital sign and laboratory data acquisition. Finally, the MSOC will develop additional standardized tools for commonly performed, routine operations such as incident cohort identification, rate of outcomes among a specified cohort, data checking, and patient natural histories.

As these planned additions and enhancements proceed, we will work to make the data model more robust from the standpoint of computational efficiency. At most Data Partner sites, the MSDD does not reside in a standalone environment so parallel research projects and operational projects may negatively affect the speed of MSCDM querying and updating.

Identifying the needs and expansion priorities for the Mini-Sentinel Distributed Database will continue to be a high priority for the Mini-Sentinel Operations Center and Data Core. In general, data needs and priorities will be driven by specific needs as identified by FDA and informed through discussions with the Mini-Sentinel Data Partners. These might include new Data Partners, data sources available through linkage, and inclusion of additional data tables and data elements in the MSCDM. Findings from other Mini-Sentinel activities (e.g., health outcomes of interest reports), and those of projects such as OMOP, will help direct future data collection efforts. Independent of those activities and any specific needs identified by FDA, it is expected expanding the MSCDM to incorporate more clinical data will have the biggest impact of the ability of the system to effectively conduct medical product surveillance. For instance, pathology reports, additional laboratory values, and other clinical measures that help to better classify risk and define covariates will likely yield substantial benefits. Additionally, in discussions with FDA, Data Partners, and others, we will identify new data resources to expand the underlying population and/or enhance the clinical data available to existing data partners.

Incorporation of new data resources may necessitate development of new capacities to extract the data, link data resources, and evaluate the data within a distributed environment. These new resources could include:

- Enhanced natural language processing to extract clinical data from medical text such as radiology reports or the medical record.
- Improved methods for extracting EHR data and mapping it to existing or new MSCDM data tables.
- Linkage of Mini-Sentinel Data Partners to external data sources such as immunization registries for incorporation into the MSDD.
- Distribute analytic techniques for horizontally distributed data (such as currently exists within the MSDD) that allow fully distributed regression analysis within the Mini-Sentinel environment.
- Investigation of vertically distributed analytic techniques that would allow inclusion of Data Partners that have access to clinical data for Mini-Sentinel Data Partner members. This entails developing an approach to have access to data held by different institutions for the same individual. This would greatly expand the clinical data available for members included in the MSDD, but faces substantial technical barriers.

V. REFERENCES

1. Maro JC, Platt R, Holmes JH, et al. Design of a National Distributed Health Data Network. *Ann Intern Med*. 2009; 151: 341-344.

2. Brown JS, Lane K, Moore K, et al. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative; U.S. Food and Drug Administration: FDA-2009-N-0192-0005. 2009. Available at: http://www.regulations.gov/#!documentDetail;D=FDA-2009-N-0192-0005. Accessed 3/16/11.

3. Velentgas P, Bohn R, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiology and Drug Safety*. 2008; 17: 1226-1234.

4. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care*. 2010; 48: S45-51.

5. Brown J, Holmes J, Maro J, et al. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. Effective Health Care Research Report No. 13. (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.) Rockville, MD: Agency for Healthcare Research and Quality, July 2009. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=150. Accessed 3/16/11.

6. Brown J, Holmes J, Syat B, et al. Proof-of-principle evaluation of a distributed research network. Effective Health Care Research Report No. 26. (Prepared by the DEcIDE Centers at the HMO Research Network and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.) Rockville, MD: Agency for Healthcare Research and Quality, June 2010. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=464. Accessed 3/16/11.

7. Brown J, Syat B, Lane K, et al. Blueprint for a distributed research network to conduct population studies and safety surveillance. Effective Health Care Research Report No. 27. (Prepared by the DEcIDE Centers at the HMO Research Network and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.) Rockville, MD: Agency for Healthcare Research and Quality, June 2010. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=465. Accessed 3/16/11.

8. FDA Sentinel Initiative report. Available at: http://www.regulations.gov/#!docketDetail;dct=FR+PR+N+O+SR;rpp=10;po=0;D=FDA-2009-N-0192. Accessed 3/16/11.

9. HMO Research Network (HMORN) Website. Available at: http://www.hmoresearchnetwork.org. Accessed 3/16/11.

10. Vaccine Safety Datalink Project Website. Available at: http://www.cdc.gov/vaccinesafety/Activities/VSD.html. Accessed 3/16/11.

11. Lieu TA, Kulldorff M, Davis RL, et al. Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care*. 2007; 45(10 Supl 2): S89-95.

12. Lee GM, Yih K, Brand W, Brown JS, Fahey K, Rett M, Roddy J, Rosofsky R, Van Dyke J, Weintraub E, Zhu S, Lieu T, Platt R, Salmon D. H1N1 Influenza Vaccine Safety in the Post-licensure Rapid Immunization Safety Monitoring (PRISM) Network. Presented at the 48th Annual Meeting of the Infectious Diseases Society of America. Vancouver, British Columbia, Canada, October 2010.

13. Informatics for Integrating Biology and the Bedside (i2b2). Available at: https://www.i2b2.org. Accessed 3/16/11.

14. Electronic Primary Care Research Network (ePCRN). Available at: http://www.epcrn.bham.ac.uk. Accessed 3/16/11.

15. Moore K, Cheetham TC, Dublin S, Hecht J, Robsinson S, Butani A, Hitz P, Krajenta R, Miroshnik I, Ren J, Saylor G, Schmidt M, Liu J, Yang X, Zhao Y, Brown JS. Virtual Data Warehouse Pharmacy File: Strengths, Weaknesses, and Recommendations. Presented at the 2009 HMO Research Network Annual Meeting, Danville, PA, April. 2009.

VI. APPENDICES

A. APPENDIX A: EHR DISCUSSION OUTLINE

EHR INVENTORY – PHASE 1 DISCUSSION OUTLINE

Revised and Final - 9/9/10

Instructions: The Mini-Sentinel Operations Center (MSOC) seeks to develop an initial profile of the Data Partners' ambulatory and hospital electronic health record (EHR) systems in order to understand the breadth of information and some of the limitations of the data available for queries on drug and device safety. A more detailed data inventory will be developed in the coming year. At this time, we seek your insights about the types of information that could be available for the inventory.

For purposes of this project, please use the following definition of "EHR Data Warehouse".

> *"EHR Data Warehouse": refers to the administrative and clinical data set used for day-to-day health care operations in the hospital or ambulatory care setting. It does not include health insurance claims data or pharmacy benefit management (PBM) data. The EHR Data Warehouse may receive data from multiple sources and subsystems.*

Please note that the MSOC is more interested in data that is stored as discrete information in a database rather than data that needs to be manually abstracted from clinical documents. This distinction as well as the difference between the nature and availability of data in the inpatient and ambulatory settings is a recurring theme in many of the questions that follow.

MSOC has asked Outcome Sciences, Inc. (DBA Outcome) to lead Phase 1 of the project and to deliver an initial inventory by September 22, 2010. **Therefore, please send your responses to these questions by September 16 to XXXX at XXXX@outcome.com, 617-XXX-XXXX.** If you would like further clarification of any terms or have any other questions or suggestions, please contact XXX.

If you are not able to complete all questions due to time constraints, please fill in what you can and provide an explanatory comment, as it will be better to include even partial information at this time.

1. Please give us a brief overview of your inpatient and ambulatory EHR systems with a focus on the nature of the patients whose data are available through these systems (are they limited to your covered population?) and the breadth, longitudinal scope and possible holes in the data you have on these patients**.** As noted above, we are more interested in data that is stored as discrete information in a database rather than data that needs to be manually abstracted from clinical documents. To the extent that some important clinical data such as Pulmonary Function Tests, and Echo reports among many others may be stored in source systems that are distinct from the EHR, we are interested in these types of data as well.

2. Are the systems compatible across settings (i.e., inpatient, outpatient, other) and sites (clinics, owned hospitals, contracted hospitals)? Can the systems be linked across settings and sites? Can you follow a patient across settings and sites?
   a. If they can be linked
      i. What proportion of patients can be linked?

ii. Over what time period?

iii. Could an individual's care be tracked from childhood through adulthood in the EHR data warehouse?

b. If they can't be linked, what are the barriers to linkage? (Different staff, hardware issues, software issues, lack of electronic records, etc.)

3. Tell us about the following clinical data subsystems that may be present in your information system environment. Include approximate dates from which data might be available and comments on data quality and completeness, and the scope of patients on whom the data are available (inpatients, ambulatory patients, and degree of overlap with "covered" population). You will be asked whether or not certain data elements are stored as discrete fields. If the answer is yes, please include the dates these data would be available and provide any additional comments. If the answer is no, please comment whether or not these elements could be manually abstracted from your system.

- Cardiology system
  - Echocardiogram: Is ejection fraction stored as a discrete field?
    - Yes or no: ___________
    - Dates available: _________________
    - Comment: _____________________________
  - Catheterization report: Is the degree of stenosis in each vessel stored as a discrete field?
    - Yes or no: ___________
    - Dates available: _________________
    - Comment: _____________________________
  - Stress test results: Is the presence, degree and location of perfusion defects stored as discrete fields?
    - Yes or no: ___________
    - Dates available: _________________
    - Comment: _____________________________
- Radiology system
  - Are concluding remarks about the presence or absence of findings stored as discrete information?
    - Yes or no: ___________
    - Dates available: _________________
    - Comment: _____________________________
- Pharmacy system
  - In the inpatient setting, do you have discrete information on medications dispensed, including the medication name, the dosage and the time administered?
    - Yes or no: ___________
    - Dates available: _________________
    - Comment: _____________________________
  - In the inpatient and ambulatory settings, how do you capture medication allergies and adverse reactions?
    - Inpatient approach: ______________________________
    - Ambulatory approach: ____________________________
  - Do you capture information on reasons for discontinuation of medications other than allergies/adverse reactions (e.g. medication was ineffective)

    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
  - Do you capture date of discontinuation of medication orders in the inpatient setting?
    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
  - Do you capture date of discontinuation of medication orders in the ambulatory setting?
    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
- Pulmonary function testing
  - Is the FEV1 and FVC available as discrete fields?
    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
- Oncology system / chemotherapy system
  - Does your system record specific details (timing, dosage, administration interval) of chemotherapy?
    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
- Pathology system
  - Are concluding remarks about the presence or absence of findings stored as discrete information?
    - Yes or no: ____________
    - Dates available: ___________________
    - Comment: ________________________________
- What other clinical data subsystems are available? (e.g., Endoscopy, EMG results, EEG, Vascular studies)

4. Laboratory Tests

   a. Are **inpatient** laboratory results included in your EHR (or in an accessible laboratory source system) as discrete data?
   Yes or no: ___________________. If yes, since when? __________________

   b. Are **ambulatory** laboratory results included in your EHR (or in an accessible laboratory source system) as discrete data?
   Yes or no: ___________________. If yes, since when? __________________

   c. Do you use LOINC codes, CPT codes or other controlled terminology for laboratory testing, or are your lab test names encoded using a local or nonstandard schema?
   Yes or no: ________________. Comment: ____

d. Do you include the normal ranges for each lab test within your EHR?
Yes or no: ________________. If so, how are changes in normal ranges for test results reflected over time? Comment: _____________________________________

e. Of the laboratory blood tests listed below, which, if any, are **NOT** included in your EHR?
____

- Alkaline Phosphatase (ALP)
- Alanine Aminotransferase (SGPT, or ALT)
- Aspartate Aminotransferase (SGOT, or AST)
- Gamma-glutamyl Transpeptidase (GGT, GGTP)
- Bilirubin
- Lactate Dehydrogenase (LDH)
- Brain, B, or Beta-type Natriuretic Peptide (BNP)
- Urea Nitrogen
- Creatinine
- Estimated Glomerular Filtration Rate
- Glucose
- Glycosylated hemoglobin (HbA1c)
- High Density Lipoprotein
- Low Density Lipoprotein
- Triglycerides
- Total Cholesterol
- Hemoglobin
- Hematocrit
- Prothrombin Time
- Platelets
- White Blood Cell Counts
- PTT
- International Normalized Ratio (INR)
- Potassium
- Sodium
- Thyroid Stimulating Hormone (TSH)
- Creatine Kinase (CK, CPK)
- C-reactive Protein (CRP)
- Troponins

5. Indicate which of the following vital signs your EHR system captures as discrete data in the inpatient and ambulatory settings and comment as needed.

Vital Sign	Inpatient	Ambulatory
Systolic Blood Pressure		
Diastolic Blood Pressure		
Height		
Weight		
Body Mass Index (BMI)		
Heart Rate		
Pulse Ox		
Temperature		

6. Indicate which of the following social history variables are captured in your EHR as discrete data in the inpatient and ambulatory settings and comment as needed.

Social History Variable	Inpatient	Ambulatory
Tobacco Status (e.g. Smoker, non-smoker, former smoker)		
Tobacco type (e.g. cigarette, cigar)		
Alcohol use		

7. Does your EHR capture information on OTC medications and nutritional supplements in a standardized way?

8. Number of Patients.

   a. Roughly, how many unique patients have clinical data in the **ambulatory** setting that includes at least one recording of a set of vital signs in calendar year 2009? How many patients have at least one blood laboratory test result in calendar year 2009? How many of these patients overlap the administrative data already stored in the Common Data Model?

   b. Roughly, how many unique patients have clinical data in the **inpatient** setting that includes discrete pharmacy administration records and laboratory results?

9. Do you capture information on device use in your patient population?

   - If so, please describe what information, what devices (i.e., categories, specific types [pacemakers, stents, valves. joint replacements,), where, and how device information is captured.
   - Is device information incorporated into the EHR or linked to the EHR?
   - Is device information captured in registries in your patient population?
     - If so, is the registry data incorporated into the EHR or linked to the EHR?
   - Please comment on the completeness and quality of the device data in your EHR.

10. Are there issues in distinguishing maternal and infant test results in the period shortly after delivery where the infant may have the same member ID as the mother?

11. With a special focus on problem list diagnoses, medications, vital signs and other clinical findings and procedures, to what extent are standardized clinical vocabularies incorporated into the way data is captured and stored in the EHR?

12. Have you used any of your EHR data in research studies?

    a. Have you used laboratory or medication data for research?
    b. Have you used laboratory or medication data for active surveillance?
    c. What is the earliest date from which data from the EHR data warehouse might be available for Mini-Sentinel activities?

13. Have you undertaken any validation studies using EHR data? (Validation studies might be done for routine clinical operations or for clinical research.)

14. Please provide any additional comments on the quality and completeness of the data in the EHR data warehouse.

Contacts:
Please provide the name, email address, and phone number(s) of at least one person in your organization whom the MSOC may contact regarding this inventory project.

B. APPENDIX B: MINI-SENTINEL COMMON DATA MODEL INITIAL ETL REPORT TEMPLATE

MINI-SENTINEL COMMON DATA MODEL VERSION 1.0
PRELIMINARY ETL REPORT

Instructions: This document gathers information on the initial Extract-Transform-Load (ETL) process for the Mini-Sentinel Common Data Model V1.0. The information provided will be used for high-level assessment of the transformation and will be kept confidential. **Section B** gathers high-level information from the MSCDM tables. It should be used as a guideline for internal data quality checks. Please ensure that transformations do not lead to obvious errors before sharing the data with the Mini Sentinel Coordinating Center for data checking activities. **Section C** gathers detailed site-specific information on how each table was created. An example of a completed Section C table is included as an Appendix. Completion of Section B requires access to the transformed tables and simple programming procedures. Please direct any questions and return the completed documents to Nicolas Beaulieu (Nicolas_Beaulieu@HarvardPilgrim.org) at the Mini-Sentinel Operations Center.

1. Data Partner Information

Data Partner Name:
Date Completed:
Contact Name:
Email Address:
Phone Number:

2. SAS Technical Environment

1. Please execute these statements in the environment where you will be writing and reading the CDM:

```
%put; %put ******* We use SAS version: &sysver. - &sysvlong.; %put;
Proc setinit; run;
```

And place the results found in the LOG into this box:

2. Please state the operating system name and version under which you run SAS _________

3. Preliminary Assessment of Key Variables

Please provide the following information based on the transformed MSCDM tables:

Enrollment

1. Number of unique members placed into the Enrollment table: _____________
2. Number of unique members enrolled in most recent month of the Encounter table: _____________
   Please also specify the calendar month and year: _____________
3. Number of records placed into the Enrollment table: _______________
4. Number of records with only Medical coverage: __________
5. Number of records with only Drug coverage: __________
6. Number of records with both Medical and Drug coverage: __________
7. Minimum, median, mean, and maximum values of length of enrollment:
   Minimum: ___________
   Median: ___________
   Mean: _________
   Maximum: ___________

Demographic

8. Number of unique members in the Demographic table: __________

9. Please fill the following tables to document distribution by Sex and Race:

Sex		
Value	**Count**	**Percent**
F		
M		
U		

Race		
Value	**Count**	**Percent**
0		
1		
2		
3		
4		
5		
6		

Dispensing

10. Number of unique members placed into the Dispensing table: ___________

11. Annual mean number of dispensings per patient:

2000: _____
2001: _____
2002: _____
2003: _____
2004: _____
2005: _____
2006: _____
2007: _____
2008: _____
2009: _____
2010: __________

12. Minimum, median, mean, and maximum values RXSup (days supplied):

Minimum: ___________
Median: ___________
Mean: _________
Maximum: ___________

13. Minimum, median, mean, and maximum values RXAmt (Amount supplied):

Minimum: ___________
Median: ___________
Mean: _________

Maximum: ___________

Encounter

14. Number of unique members placed into the Encounter table: _______

15. Number of unique members placed into the Diagnosis table: _______

16. Number of unique members found in the Procedure table: _______

17. Number of records in the Encounter by EncType (Encounter Type):

    Inpatient (IP): _______
    Outpatient/Ambulatory Visit (AV): _______
    Emergency Department (ED): _______

18. Number of records in the Diagnosis by EncType (Encounter Type):

    Inpatient (IP): _______
    Outpatient/Ambulatory Visit (AV): _______
    Emergency Department (ED): _______

19. Number of records in the Procedure by EncType (Encounter Type):

    Inpatient (IP): _______
    Outpatient/Ambulatory Visit (AV): _______
    Emergency Department (ED): _______

20. Number of unique EncounterIDs placed into the Encounter table: ________

21. Number of unique EncounterIDs placed into the Diagnosis table: ________

22. Number of unique EncounterIDs placed into the Procedure table ________

Death and Cause of Death

23. Number of unique members placed into the Death Table: ______

24. Number of records in the Death table: _____

25. Number of unique members placed into the Cause of Death table: ______

26. Number of records in the Cause of Death Table: ______

27. Percent of unique members found in both the enrollment table and each of the other tables (based on answers to Questions 8, 10, 14, 15, 16, 20 and 23):

Table Name	% Unique Members also in Enrollment Table
2. Demographic	
3. Dispensing	
4.1 Encounter	
4.2 Diagnosis	
4.3 Procedure	
5.1 Death	
5.2 Cause of Death	

4. Report on Mini-Sentinel Common Data Model ETL

For each of the MSCDM tables, please provide details regarding how each variable was created and document any issues or recommendations regarding use. For the overall table comment section, please provide a description of (a) the definition of a unique record in the table, and (b) overall suggestions or comments about creation and use of the table for Mini-Sentinel. The Appendix provides an example of a completed table.

Table 1. Enrollment

Table Name: ______		
	Variable Name	**Site Comments (Required for All)**
1	PatID	
2	Enr_Start	
3	Enr_End	
4	MedCov	
5	DrugCov	
Site Comments for Entire Table (Required)		

Table 2. Demographic

Table Name: ______		
	Variable Name	Site Comments (Required for All)
1	PatID	
2	Birth_Date	
3	Sex	
4	Race	
Site Comments for Entire Table (Required)		

Table 3. Dispensing

Table Name: ______		
	Variable Name	Site Comments (Required for All)
1	PatID	
2	RxDate	
3	NDC	
4	RxSup	
5	RxAmt	
Site Comments for Entire Table (Required)		

Table 4.1. Encounter

Table Name: ______		
	Variable Name	**Site Comments (Required for All)**
1	PatID	
2	EncounterID	
3	ADate	
4	DDate	
5	Provider	
6	Facility_Location	
7	EncType	
8	Facility_Code	
9	Discharge_Disposition	
10	Discharge_Status	
11	DRG	
12	DRG_Type	
13	Admitting_Source	
Site Comments for Entire Table (Required)		

Table 4.2. Diagnosis

Table Name: ______		
	Variable Name	**Site Comments (Required for All)**
1	PatID	
2	EncounterID	
3	ADate	
4	Provider	
5	EncType	
6	DX	
7	Dx_Codetype	
8	OrigDX	
9	PDX	
Site Comments for Entire Table (Required)		

Table 4.3. Procedure

Table Name: ______		
	Variable Name	Site Comments (Required for All)
1	PatID	
2	EncounterID	
3	ADate	
4	Provider	
5	EncType	
6	PX	
7	PX_Codetype	
8	OrigPX	
Site Comments for Entire Table (Required)		

Table 5.1. Death

Table Name: ______		
	Variable Name	Site Comments (Required for All)
1	PatID	
2	DeathDt	
3	DtImpute	
4	Source	
5	Confidence	
Site Comments for Entire Table (Required)		

Table 5.2. Cause of Death

Table Name: ______		
	Variable Name	**Site Comments (Required for All)**
1	PatID	
2	COD	
3	CodeType	
4	CauseType	
5	Source	
6	Confidence	
Site Comments for Entire Table (Required)		

5. Example Report (Enrollment)

Table name: Enroll (Example)		
	Variable Name	**Site Comments (Required for All)**
1	PatID	Site defined: actual local Patient ID is PHI – scrambled for Mini-Sentinel purposes.
2	Enr_Start	Actual start date of enrollment period.
3	Enr_End	Actual end date of enrollment period. If actual day missing then imputed to last day of month.
4	MedCov	Site defined: aggregated from local “plan type” information.
5	DrugCov	Site defined: aggregated from local “plan type” information.
Site Comments for Entire Table (Required)		
Each record represents a unique combination of Patient ID, beginning and end dates of enrollment, Medical Coverage, and Drug Coverage; any change in any of these variables generated a new record. 1- Enrollment gaps of less than 30 days have been bridged; our health plan does this as standard practice.. 2- Records with DrugCov=”Y” and MedCov=”N” are Medicare PDP patients. 3 - The drop in members in 2006 is real, we left the Rhode Island insurance market that year. 4 - All of our non-Medicare plans have medical coverage, A few do not have drug coverage.		

C. APPENDIX C. PRELIMINARY ETL REPORT SUMMARY RESULTS*

Table 1. Summary ETL Report

	DP1	DP2	DP3	DP4	DP5	DP6	DP7	DP8	DP9	DP10	DP11	DP12	DP13
ENROLLMENT													
Unique members													
Current members/All members	55.1%	43.3%	0.0%	18.3%	32.5%	27.0%	32.7%	17.8%	22.3%	43.2%	21.8%	51.8%	38.6%
Enrollment periods per person	2.9	1.7	4.0	1.9	1.6	0.2	1.6	4.8	2.2	3.1	2.1	1.0	1.2
Records with only Medical coverage	54.9%	5.3%	44.2%	0.0%	11.4%	100.0%	6.0%	23.1%	5.8%	24.4%	100.0%	1.5%	9.2%
Records with only Drug coverage	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	52.5%	43.6%	0.3%
Records with both Medical and Drug coverage	45.1%	94.7%	55.8%	0.0%	88.6%	0.0%	94.0%	76.9%	47.1%	75.6%	52.5%	54.9%	90.4%
Length of enrollment: (differing units)													
Minimum	0.0	0.0	0.1	---	-0.1	0.0	0.0	0.1	0.0	0.1	0.1	1.0	1.0
Median	3.1	2.5	1.7	3.6	1.6	1.9	1.7	2.0	1.3	2.3	1.7	580.0	578.0
Mean	5.6	4.1	3.6	5.7	2.7	3.5	3.3	4.9	3.7	4.2	4.2	542.0	726.0
Maximum	18.4	32.1	22.4	1,006.6	13.4	37.0	16.5	59.5	52.6	15.0	50.4	914.0	2,223.0
DEMOGRAPHIC													
Unique members													
Sex, N (%)													
Ambiguous	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Female	51.9%	50.8%	52.4%	53.6%	50.9%	53.4%	51.9%	52.1%	49.9%	50.4%	51.3%	56.0%	50.7%
Male	48.1%	49.2%	47.5%	46.4%	48.9%	46.6%	48.1%	47.9%	49.9%	49.6%	48.6%	44.0%	49.3%
Unknown	0.0%	0.0%	0.2%	0.0%	0.3%	0.0%	0.0%	0.0%	0.2%	0.0%	0.1%	0.0%	0.1%
Race, N (%)													
0 = Unknown	100.0%	93.0%	98.9%	16.5%	86.2%	100.0%	81.8%	89.1%	84.7%	67.3%	79.9%	36.7%	---
1 = American Indian or Alaska Native	0.0%	0.0%	0.0%	0.4%	0.1%	0.0%	0.1%	0.0%	0.3%	0.2%	0.2%	0.3%	---
2 = Asian	0.0%	0.0%	0.1%	3.2%	0.8%	0.0%	0.7%	0.5%	5.6%	5.4%	0.8%	0.9%	---
3 = Black or African American	0.0%	0.0%	0.0%	30.7%	1.6%	0.0%	1.2%	4.6%	0.4%	3.4%	0.7%	6.1%	---
4 = Native Hawaiian or Other Pacific Islander	0.0%	0.0%	0.0%	49.3%	0.0%	0.0%	0.1%	0.0%	3.8%	0.2%	0.1%	0.9%	---
5 = White	0.0%	7.0%	0.9%	0.0%	11.3%	0.0%	16.2%	5.7%	5.3%	23.5%	18.3%	55.0%	---
DISPENSING													
Members with 1+ dispensings	68.2%	77.4%	93.0%	54.6%	59.3%	67.8%	72.3%	59.7%	56.7%	81.5%	68.0%	83.0%	71.0%
Days supply (RxSup)													
Minimum	0	1	1	0	-30	0	1	0	0	0	0	1	0
Median	30	30	30	30	30	30	60	30	30	30	30	30	30
Mean	33	31	38	34	33	28	43	31	37	50	180	32	28
Amount supply (RxAmt)													
Minimum	0	1	0	0	-89	0	0	0	-300	0	0	1	0
Median	30	30	30	35	30	30	60	30	40	60	56	30	30
Mean	57	70	63	63	64	56	78	61	71	92	76	49	50

Table 1. Summary ETL Report (Cont.)

	DP1	DP2	DP3	DP4	DP5	DP6	DP7	DP8	DP9	DP10	DP11	DP12	DP13
ENCOUNTER													
Percentage of all members in encounter file	113.2%	89.1%	88.1%	242.7%	71.4%	78.9%	75.8%	69.5%	67.3%	95.3%	66.1%	46.0%	76.0%
Percentage of all members in diagnosis file	113.2%	88.7%	86.1%	201.8%	71.4%	78.3%	74.8%	71.5%	59.0%	93.4%	60.6%	46.0%	76.0%
Percentage of all members in procedure file	108.0%	88.7%	86.9%	309.6%	71.4%	78.8%	74.2%	75.6%	65.5%	84.1%	69.4%	46.0%	76.0%
Distribution of encounter type in encounter file													
Inpatient (IP)	2.5%	1.6%	1.5%	5.5%	1.1%	2.3%	2.7%	4.2%	2.2%	1.7%	1.4%	14.7%	1.5%
Outpatient/Ambulatory Visit (AV)	92.5%	97.2%	72.9%	89.1%	97.0%	93.9%	93.7%	91.5%	97.2%	93.9%	94.5%	77.4%	93.9%
Emergency Department (ED)	5.0%	1.2%	25.6%	5.4%	1.9%	3.8%	3.7%	4.3%	0.6%	4.4%	4.1%	8.0%	4.6%
Total excluding other encounter types (in % of Unique EncounterIDs)	85.9%	96.4%	103.5%	102.5%	104.4%	80.6%	67.2%	100.0%	99.6%	88.0%	54.4%	77.5%	82.8%
Distribution of encounter type in diagnosis file													
Inpatient (IP)	11.2%	8.8%	6.8%	11.6%	5.4%	6.1%	8.4%	7.6%	10.5%	4.8%	5.9%	25.3%	6.9%
Outpatient/Ambulatory Visit (AV)	84.3%	88.6%	89.3%	82.5%	92.1%	84.8%	86.3%	85.8%	89.5%	91.3%	88.9%	65.9%	88.3%
Emergency Department (ED)	4.5%	2.6%	3.9%	6.0%	2.5%	9.1%	5.3%	6.6%	0.1%	3.9%	5.2%	8.9%	4.8%
Diagnoses per encounter by encounter Type:													
Inpatient (IP)	9.08	10.77	5.82	4.77	9.49	5.20	6.43	4.36	4.84	6.41	6.42	5.13	7.26
Outpatient/Ambulatory Visit (AV)	1.84	1.83	1.54	2.09	1.79	1.80	1.88	2.27	0.94	2.14	1.44	2.54	1.49
Emergency Department (ED)	1.82	4.30	0.19	2.52	2.47	4.84	2.98	3.73	0.11	1.91	1.95	3.32	1.65
Distribution of encounter type in procedure file													
Inpatient (IP)	9.4%	10.1%	8.2%	4.4%	6.6%	8.3%	17.3%	6.7%	2.2%	7.2%	9.5%	34.5%	11.0%
Outpatient/Ambulatory Visit (AV)	88.5%	85.5%	88.8%	91.4%	87.4%	81.9%	70.5%	85.5%	97.7%	83.2%	81.8%	51.4%	79.9%
Emergency Department (ED)	2.1%	4.4%	3.1%	4.2%	6.0%	9.8%	12.2%	7.8%	0.1%	9.6%	8.6%	14.2%	9.1%
Procedures per encounter by encounter Type:													
Inpatient (IP)	3.71	18.85	6.42	2.51	17.66	9.13	17.19	4.30	1.76	2.70	12.00	13.94	19.07
Outpatient/Ambulatory Visit (AV)	0.94	2.70	1.41	3.20	2.62	2.24	1.99	2.54	1.78	0.55	1.52	3.94	2.21
Emergency Department (ED)	0.42	11.14	0.14	2.45	9.24	6.69	8.83	4.97	0.32	1.34	3.69	10.55	5.10
DEATH AND CAUSE OF DEATH													
Unique members found in both the Enrollment table and table for													
Demographic	86.2%	100.0%	69.6%	---	54.6%	98.8%	71.6%	99.0%	72.5%	54.8%	81.2%	100.0%	100.0%
Dispensing	99.5%	99.8%	80.9%	---	99.8%	98.8%	93.8%	99.1%	86.9%	95.6%	93.1%	83.0%	100.0%
Encounter	88.3%	99.8%	87.5%	---	99.9%	99.8%	96.5%	100.0%	79.0%	90.4%	90.6%	46.1%	100.0%
Diagnosis	88.3%	99.9%	88.1%	---	99.9%	99.8%	96.8%	100.0%	81.5%	91.2%	92.9%	46.1%	100.0%
Procedure	92.6%	99.9%	88.0%	---	99.9%	99.8%	96.7%	100.0%	78.8%	92.2%	91.8%	46.1%	100.0%
Death	48.9%	---	58.8%	---	55.2%	0.0%	83.5%	100.0%	91.2%	43.4%	88.8%	8.7%	100.0%
Cause of Death	100.0%	---	69.8%	---	55.2%	0.0%	100.0%	100.0%	0.0%	47.7%	89.5%	---	---

*Note: This table represents data reported by the Data Partners as part the initial ETL Report requirement, and was generated before any data checking or characterization was completed. The information was used to identify potential problems with the ETL process and the MSCDM.

D. APPENDIX D: EXAMPLE OF DATA CHECKING REPORT QUESTIONS IDENTIFIED IN YEAR 1

Table 1. Enrollment Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
ENRL3	MedCov (Dataset: Enr_L3_enrmd_y)	All enrollment in 2010 is Med coverage only. There are no Drug only or Med and Drug Coverage. Please confirm that this is correct.	Updating drug coverage data involves a manual process that we perform on an as-needed basis due to resource constraints. Historically 'as-needed' has been about every 9 months or so.
ENRL3	DrugCov flag = U (Dataset: Enr_l3_meddrugcov)	More than 14% of all records have MedCov=Y and DrugCov=U. What are those records? Please explain.	This enrollment file contains periods that start before Year1. Drug coverage information is unknown before Year2 but is known afterwards. The Year3 and Year4 unknowns are 14% of total.
ENRL3	MedCov (Dataset: Enr_l3_medcov)	There are no records with Drug Coverage only (i.e., MedCov=N). Please confirm that this is the case.	This is accurate. Our company does not offer drug coverage in the absence of medical coverage
ENRL3	MedCov & Drug Cov distribution (Dataset: Enr_l3_meddrugcov)	Very few (0.3%) enrollment records have the MedCov=No and DrugCov=Yes. Are these errors or expected from your population?	Yes, this is expected and will increase as we have included adding RX-Only members in extraction 2.
ENRL3	Enrollment Counts (Dataset: Enr_l3_stats_enrd) (Dataset: Enr_l3_stats_enrm)	The max number of months of enrollment found in the data is XXX for drug and XXX for medical. Is that explained by question ENR2.1 above with multiple records duplicated imply that some patients are over-represented (i.e. one PatID with 2 identical records?	That is correct.
ENRL3	Membership Rate	There seems to be a decreasing number of members in your data starting somewhere around in mid 2007 (Dataset: Enr_l3_enrmd_ym). Is this a known trend in your data (i.e. the plans you have used for MSCDM) or does this reflect some change in the way the enrollment periods are coded? It seems to be confirmed by the counts of records in the dispensing and utilization tables.	The peak in late 2007 is consistent with our total membership trend.
ENRL3	Membership Rate	There seems to be a slight decrease in number of members in your data starting somewhere around in mid XXXX (Dataset: Enr_l3_enrmd_ym). Is this a known trend in your data (i.e. the plans you have used for MSCDM) or does this reflect some change in the way the enrollment periods are coded? It seems to be somewhat confirmed by the counts of records in the dispensing and to some extent in the utilization tables.	This is caused by the dramatic drop of enrollment during that year.

Table 2. Demographics Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
DEM1.9	Birth_Date: Between 1/1/1885 - Current Date	A small percentage (64 records) of patients have birthdates outside this range. Please also see remark DEML3 below on Age.	These 64 records all seem to be data entry issues.
DEM2.1	Definition of the table: 1 record per person	66K records have duplicate information, i.e. identical information (Dataset: Dem_l2_def). Please explain.	This is due to internal system issues and some of the patIds being the same. This is related to the duplicate patients item that was previously addressed. These will be dropped in the next round
DEM1.9	Birth_Date: Between 1/1/1885 - Current Date	655 records with birthdates outside of this range. Please also see remark DEML3 below on Age.	There only 6 people (0.00022% of total) This is a known issue and it is how DOB comes from the source data base
DEML3	Age	Age: Min=-2 and Max=128. Certainly outliers but problems with Date of Birth? Please explain.	Negative age is due to newborn's DOB being after enrollment start date, or know source issues with mapping multiple member ids to the same person. There are only 2 persons with age over 120 years old, and it is how it came from the source data. Very low %.
DEM2.1b	PatId	There are about 70K more records than unique PatIDs. Please explain.	This is due to internal system issues and some of the patIds being the same. This is related to the duplicate patients item that was previously addressed.
DEML3	Age	Age: Min=-2290 and Max=232. Certainly outliers but problems with Date of Birth? Please explain.	These 64 records all seem to be data entry issues.

Table 3. Dispensing Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
DIS1.10	NDC: Length of NDC= 11	There are some NDC values = 0. Since we don't have frequencies we don't know if that is a issue or not. Please confirm.	There are only 13 records that have this value. Noted to be cleaned up during next extraction. Data probably came in as a blank, not a local code, does not represent a dispensing.
DIS1.10	NDC: should be 11 characters	Almost 1% of records are less than 11 characters	This is a known issue. A small percentage of records contains homegrown, invalid or missing NDC codes. 1) Our internal pharmacy system is keyed off of an internal drug numbering system, not NDC. For very old prescriptions, we were unable to crosswalk a very small number of records to NDC. In these instances, we create an NDC value of "AAAxxxxxAAA" where the "x" indicate the internal drug number. This allows us to preserve the occurrence of prescriptions and allows a project to convert these manually if they deem it worth the resources to do so. 2) Our pharmacy system occasionally contains non-drug based transactions for pharmacy related items such as needles for insulin. These items do not have NDC values.
DIS2.1	Definition of the table (Dataset: Dis_l2_2, Dis_n_all): 1 record per NDC per day	>21K records have duplicate information for PatID, RxDate, and NDC Please explain.	Thanks for identifying this issue. We have researched it and will correct it in our next update. Our local files have disp date as rxdate. For XXXX forward we also have sold date which we were using as rxdate for the CDM. There are no dups by pat_id, disp date, ndc - But there are dups by pat_id, sold_dt, ndc. We had not thought of this situation. Per MSCDM specification, the RxDate should be the closest one to fill/dispense. If some date need to use Sold Date for the RxDate date, please remove duplicates. OK to be action item for next refresh of the data.
DISL3	Max Rx/Yr (Dataset: Dis_l3_rxptyr)	Max number > 250 between 2000-2010. Outliers? Known issue? Please only include one record per dispensing and make sure reversals and adjustments are carefully removed. Please explain.	There are no reversals or adjustments in the table. The 30 members that have over 250 fills in a year are correct. (Some members have multiple years of over 250 fills.) High number of fills occurred when a patient had 1) many unique drug types, 2) 30 day benefits, 3) home iv medical products and drugs (dispensed on weekly basis). These patients appeared to have aids, mental health problems, or require home iv.

Table 3. Dispensing Table (Cont.)

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
DISL3	RxSup (Dataset: Dis_L3_rxsup)	A small percentage (0.22%) of all RXs with supply between >100. Please confirm that this is what you would expect in your data.	This represents our source data. It appears that ointments, creams, solutions have a 'default' value of 120. Some titrating of doses up or down seems to be happening. Also, a person may have reason to pick up two 60 day fills in the same day (if they have extended travel plans, for example.) Some entries represent a course of therapy. We also suspect that there were some data entry errors. Perhaps, in some cases, the rxamt and rxsup were swapped.
DIS1.15	RxSup: Non-negative numbers	0.04% of records are negative. Adjustments? Reversals?	This is an error - our specification does not allow non-positive amounts. We will look into.
DIS1.18	RxAmt: Non-negative numbers	0.04% of records are negative. Adjustments? Reversals?	This is an error - our specification does not allow non-positive amounts. We will look into.
DISL3	RxAmt (Dataset: Dis_l3_rxamt)	• 0.04% of records are missing • A number of records have decimal values. Are these mL, calculated values, etc?	This is known. Yes, fractions are calculated values.
DIS1.15	RxSup: non-negative	• More than 6% of all RXs have an supply = zero. That seems to be high. Please explain what those are. • 0.04% of all RXs with supply > 100. Assuming these are different units such as number of milliliters, etc. Please confirm.	The 6% are related to reversals that did not have a corresponding rebill claim. This is correct. Previously verified.
DIS2.1	Definition of the table (Dataset: Dis_l2_2): 1 record per NDC per day	1,434 records have duplicate information for PatID, RxDate, and NDC Please explain.	This is due to internal system issues and some of the patIds being the same. This is related to the duplicate patients item that was previously addressed.
DISL3	Max Rx/Yr (Dataset: Dis_l3_rxptyr)	Max number around 600-700 between 2004-2009. Outliers? Known issue? Please confirm.	Being investigated.

Table 4. Encounter Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
ENC1.20	Provider Flag: Special characters	0.7% of values have special characters.	These are all ~~~~~~ or ~~~~~~~~~.
ENC1.22	Facility Location: 3 characters in length	All of values are missing/blank. Please confirm this is accurate	This field is not native to our source data.
ENC1.47	Admitting Source: must be selected values	All of values are missing/blank. Please confirm this is accurate	This field is not native to our source data.
ENC2.1	Definition of the table : 1 record per encounter; 1 EncounterID per PatID, ADate, Provider and EncType combination	EncounterID is not unique (i.e. per row), there are >3 Million duplicates (Datasets: Enc_l3_n_encid, Enc_l3_n_records, Enc_n_all). Please explain.	Internal data allow multiple encounters by PatID, ADATE, Provider, and EncType which are differentiated by Admit Time. We need to reconfigure the MSCDM EncounterID to be unique for each encounter, but we will still have multiple PatID-ADate-Provider-EncType combinations for a single EncounterID.
ENC2.1	Definition of the table : 1 record per encounter; 1 EncounterID per PatID, ADate, Provider and EncType combination	EncounterID is not unique (i.e. per row), there are 56,600,056 unique EncounterIDs (Dataset: Enc_l3_n_encid) and about 56,602,419 records (Dataset: Enc_l3_n_records).	Thank you for bringing this to our attention. Currently unclear of reason - could be due to update processing for our internal data source.
ENC2.1	Definition of the table (Dataset: Enc_l2_2): 1 record per encounter; 1 EncounterID per PatID, ADate, Provider and EncType combination	EncounterID is not unique (i.e. per row), there are 557M unique EncounterIDs (Dataset: Enc_l3_n_encid) and about 581M records (Dataset: Enc_l3_n_records). Please explain.	EncounterID/PatID is unique composite primary key. This has been elaborated in CDM for Extraction 2 which will be changed to unique table key for EncounterID only.
ENC1.37	Discharge Status: must be 2 characters	• All values missing/blank. Please confirm this is accurate. • Value of ",O" shows up as well, is this a typo?	This is not correct. There are valid values for IP encounter type in this field.
ENC1.40	DRG: must be 3 characters (Dataset: Enc_l3_drg_drg_type)	All values missing/blank. Please confirm this is accurate	This is correct. We do not have this information
ENC1.41	DRG: Only numeric digits	All values missing/blank. Please confirm this is accurate	This is correct. We do not have this information
ENCL3	Encounter/Patient per year	Max number around 900-1,100 between 2004-2009 (Dataset: Enc_l3_stats_y). Outliers? Known issue? Please confirm.	Being investigated.

Table 5. Diagnosis Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
DIA1.26	Dx: no special characters	1 record with space, 17 records with apostrophes, and 2 records "ZZZ.XX"	Cleaning routine will be added to 2nd extraction.
DIAL3	Dx/Encounter consistency	Even though there seems to have a decreasing trend in the NUMBER of encounters per year between 2004-2009 (Dataset: Enc_l3_adate_y) we noticed an increase in the NUMBER of total diagnoses per year during the same period (Dataset: Dia_l3_adate_y).	Being investigated.
DIA2.4	Definition of the table: one record per DX per unique PatID, ADate, Provider and EncType (Dia_n_all)	There are >95K duplicates (Datasets: Dia_l3_n_encid, Dia_l3_n_records, Dia_n_all). Please explain.	
DIAL3	Primary diagnosis flag (PDX) for IP encounters only (Dataset: Dia_l3_pdx_et).	Primary diagnosis flag is a concept associated with inpatient encounter (IP) types only. This flag is populated for all records/diagnosis in your data; should be missing for non-IP encounter types (per MSCDM v1.0 specifications). Please explain.	
DIAL3	PDx	No values of X (unable to classify) were recorded. Just wanted to confirm that this was accurate (Dataset: Dia_l3_pdx_et).	That is correct. This has been adjusted for Extraction 2 to classify all professional claims with a value of "X".

Table 6. Procedure Table

Error #	MSCDM Item	MSOC Comment (obfuscated if necessary)	Data Partner Response (obfuscated if necessary)
PRO1.24	Px: no special characters	A number of records contain special characters such as *, ' , $, +, /, and spaces, as well as faulty values like "ZZZ" and other letters not at the beginning of the code	Cleaning routine will be added to 2nd extraction. This is a very small percent.
PRO1.31	OrigPx: between 2-25 characters	All values missing/blank	This is correct. We do not have this information
PRO1.	PxCodes	ICD-9-CM procedure codes: just to confirm that you are aware that MSCDM v1.0 asks Data partners to keep decimal point.	Yes, this has been clarified in the CDM specification.
PROL3	Distribution of Visit Type	There are a large proportion of ambulatory visits (AV) with Revenue codes: what are those visits exactly? Professional services rendered during hospitalizations?	Being investigated.