Mini-Sentinel

MINI-SENTINEL COORDINATING CENTER DATA CORE

YEAR TWO COMMON DATA MODEL (CDM) REPORT

REPORT OF DATA CORE ACTIVITIES, OCTOBER 2010 - SEPTEMBER 2011

Prepared by: Mini-Sentinel Data Core and Data Partners

June 2012

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.

TABLE OF CONTENTS

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 Section 905, of the Food and Drugs Administration Amendment Act (FDAAA) of 2007 to create an active surveillance system using electronic health data for 100 million people by 2012.

The Mini-Sentinel program currently focuses on three major activities:

- Assessments - Medical product exposures, health outcomes, and links between them
- Methods - Techniques for identifying, validating, and linking medical product exposures and health outcomes
- Data - Mini-Sentinel Distributed Dataset and tools used to access the data

Collaborating Institutions enable access to data environments and provide other resources to support meeting the 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. For additional information, please see www.mini-sentinel.org.

B. MINI-SENTINEL SCIENTIFIC OPERATIONS CENTER

The Mini-Sentinel Operations Center (MSOC) leads Mini-Sentinel's scientific and management operations, via the Scientific and Management Operations Centers. The Scientific Operations Center oversees the data infrastructure and overall operation of the program. It oversees and 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. (See **Figure 1**.)

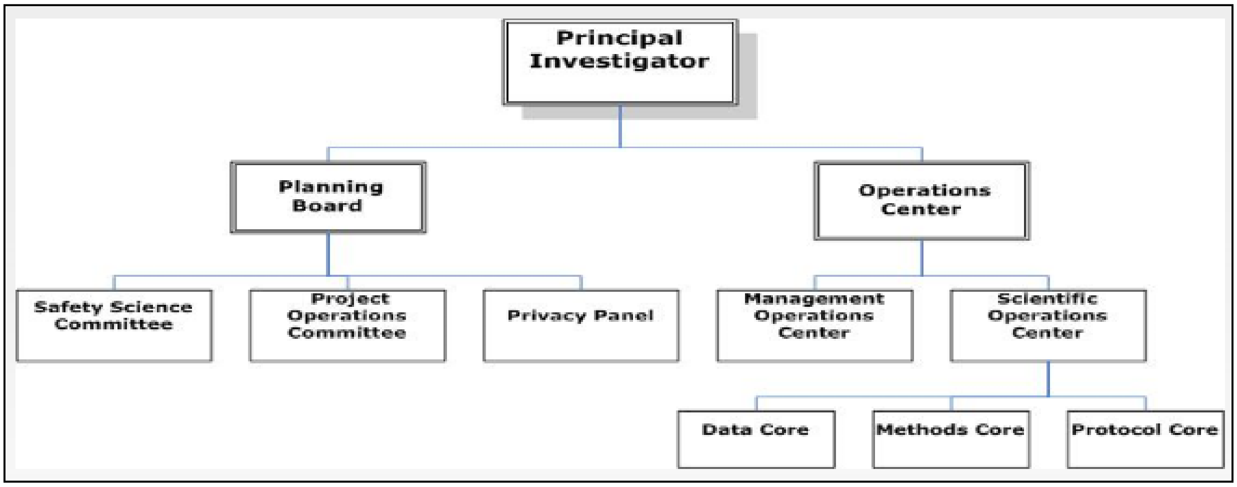
The Data Infrastructure Division oversees data development and data source documentation, as well as evaluation implementation activities of Mini-Sentinel. Individuals working within this Division possess expertise in database design, implementation, and analysis. Data Infrastructure Division staff are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

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

- Assess data quality and characteristics
- Develop reusable analytic tools (e.g. Modular Programs)
- 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 or support ad hoc 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 DATA CORE

1. Overview

The Mini-Sentinel Data Core directs the 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 maintenance of the Mini-Sentinel Distributed Database, the data held 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, FDA, 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 strategic planning of data development
- Establish ad hoc data workgroups to investigate specific topics of interest

- 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 of the Data Core

- 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 with health plan administrative claims data in the MSCDM format include Aetna, 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). 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 are ongoing and will continue to 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.[1] A centralized system stores all patients' administrative 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 Partners. 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 involves 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.

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]The Mini-Sentinel Common Data Model standardizes administrative and clinical information across Data Partners. Data Partners execute standardized programs provided by the Operations Center or project workgroups and typically share the output of these programs in summary form with the Operations Center and project workgroups. 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)[i]. 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. In addition, the distributed model 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 the MSOC or FDA.

A mixed model is used on a case-by-case basis when evaluations require person-level intermediate analytic datasets, for example, when performing multivariate analyses.[1,3] A mixed model uses a distributed approach for analyses 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. Only the minimum necessary data are transferred, which typically include 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.

II. OVERVIEW OF COMMON DATA MODEL

The MSCDM v2.1 includes 10 tables that represent specific data domains.[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. The data tables keep similar clinical concepts together and whenever possible keep the source "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 affecting 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 v2.1.

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., Social Security Number), but rather a health-plan generated, alpha-numeric string that is unique to each person in the data files. Each health plan maintains a link between the unique person identifier and the true identifier, which is retained by the Data Partner. The person identifier is unique within a health plan and is not shared outside the health plan with either the MSOC or the FDA.

Each table is briefly described below.

[i] http://www.hhs.gov/ocr/privacy/

[ii] MSCDM v2.1 is the current version. As the MSCDM is revised, newer versions will replace the older documents. MSCDM v2.1 is available at http://minisentinel.org/data_activities/details.aspx?ID=105.

Enrollment. The ability to ascertain who is eligible to receive specific kinds of care at any particular time is required for most Mini-Sentinel investigations. In many medical product safety evaluations, it is important to know the period of time during which an event of interest would be observed if it occurred. 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 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 use the enrollment table to verify the specific dates during which medical utilization identified in other tables (e.g., exposed to a specific medication) are eligible to contribute to an evaluation. The table structure is a simplification of the HMO Research Network's Virtual Data Warehouse (VDW).[8] enrollment table structure and similar in structure to the other common data models evaluated.

Demographic. The demographic table includes the 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 dispensing 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 to 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), and the days supply and amount dispensed as listed on the dispensing record. 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 900 days supplied.

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 (which is rarely the case). An analysis of pharmacy dispensing data for 11 HMORN health plans found that OTC medications accounts for 2% to 9% of all outpatient dispensings between 2000 and 2007, although this rate of capture is likely to be a small portion of all OTC use.[9] Infused medications, vaccinations, and other 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.[9] 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 an 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 sends the patient to the emergency department and the patient is later admitted to a hospital, the encounter table contains three records. Additional information in 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 the associated diagnoses are recorded in the diagnosis table. The diagnosis table includes one row for each unique diagnosis recorded during an encounter. The table also includes a flag for whether the diagnosis was recorded in the primary diagnosis field for the encounter (applies only to care in the inpatient setting), an indicator for the care setting in which the diagnosis was recorded, and an indicator for the type of diagnosis code. This "long and thin" table structure 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 those with 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 the associated procedures are recorded in the procedure table. The procedure table includes one row for each 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). Currently many coding standards are used to record procedures, including ICD-9 CM procedure codes, CPT-4 codes, and HCPCS codes; the table allows capture of any existing or future coding standards. This "long and thin" table structure 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 received specific vaccinations.

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 two-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 codes reflecting the underlying condition, along with coding dictionary used, type of contribution to the death, and the source of the information.

Laboratory. The laboratory table, which was added during this contract year, represents results and information from selected laboratory tests captured by select Data Partners. Because laboratory results can have different interpretations based on type of test or how the test is administered, the model also includes variables for test subcategory, specimen source, patient location, result location, and result unit.

HealthCore, Kaiser, and selected HMORN sites have implemented the following laboratory: alkaline phosphatase (ALP), alanine aminotransferase (SGPT), total bilirubin, glucose, glycosylated hemoglobin (HbA1c), creatinine, hemoglobin, International Normalized Ratio (INR), fibrin d-dimer, absolute neutrophil count (ANC), and lipase. Additional laboratory tests may be added in subsequent years.

Vital Signs. Two Data Partner organizations are currently contributing information on height, weight, systolic and diastolic blood pressure, and tobacco-use status for this table. This table was added during this contract year.

Detailed information on the addition of the Laboratory and Vital Signs tables are included in the following section.

III. EXPANSION OF MINI-SENTINEL COMMON DATA MODEL

A. CLINICAL DATA ELEMENTS

1. Overview

In Year Two, the Mini-Sentinel Common Data Model (MSCDM) was expanded to include Clinical Data Elements consisting of vital signs and selected laboratory test results. . Because vital signs are collected and stored as part of clinical encounters, only the six Kaiser Permanente sites and three of the HMORN sites with direct access to Electronic Health Record data could provide these data. Laboratory tests are more broadly available across the Data Partners regardless of their status as payer or provider. The specific laboratory tests included in this initial extraction were selected due to their relevance to FDA investigations, the Data Partners' sense of the feasibility of extracting the laboratory data from their source systems, and the ease with which the information could be included in the MSDCM.

The **laboratory tests** included in the MSCDM during Year Two are:

- Glucose
- Hemoglobin
- Hemoglobin A1c
- Creatinine
- Alanine Aminotransferase
- Alkaline Phosphatase
- Total Bilirubin
- International Normalized Ratio
- D-dimer
- Lipase
- Absolute Neutrophil Count

The **vital signs** included in the MSCDM during Year Two are:

- Height
- Weight
- Blood Pressure
- Tobacco Status

2. Building the Clinical Components' Data Model

a. Laboratory Test Data Model

Laboratory test results are obtained either from electronic health records or from electronic laboratory reports. They are stored in the Laboratory table in the MSCDM.

Test date. Three dates are captured for each laboratory test: the test order date, the specimen draw date and time, and the date and time when the result was reported. Not all Data Partners have complete information for all dates.

Test type. For each test type, each Data Partner could have many different internal codes. Sometimes the test name itself implies a blood source of the specimen, as opposed to a test of urine or other body source. In other cases, the source of the specimen needed to be derived from a separate specimen source field. All test names for a specific test are mapped into a standard variable MS_TEST_NAME. MS_TEST_SUB_CATEGORY is a reserved space to indicate important nuances about certain tests that may not be clear from other field values.

Although a field for **Logical Observation Identifiers Names and Codes** (LOINC)[iii] codes is included in the MSCDM, these codes were not available from all Data Partners and not every lab test was uniformly associated with a LOINC code. Additionally, most tests without a LOINC code could have reasonably been associated with more than one code. LOINC codes for tests with missing values were imputed based on the most common LOINC codes among those assigned. Since imputed LOINC codes may not be accurate and are subject to change in the future, LOINC_FLAG field was created to indicate the "natural" versus "imputed" nature of each code.

The Pt_LOC field indicates the patient location when the specimen was obtained. Importantly, some Data Partners do not have access to any inpatient data. Some tests, in particular glucose, may be performed on venous blood processed in a typical laboratory setting or performed as a point-of-care test by staff in a clinical setting using a portable analyzer, usually through the use of a finger stick. The RESULT_LOC field in the MSCDM to capture the location at which each laboratory test is performed.

Test result. The same test results may be expressed quantitatively or qualitatively. For qualitative test results, it is not always clear which thresholds are used to associate the test result with a qualitative value. Furthermore, different qualitative terms can be used to represent, essentially, the same result. Some results are "semiquantitative," having a value like ">500." With quantitative tests, the results may not be directly comparable both within and across Data Partner sites because they may be associated with different units of measurement. The same test can have different reference ranges of normal depending on the characteristics of the person undergoing the test or the idiosyncrasies of the equipment or reagents used in the test. Furthermore, not all quantitative tests are associated with units or reference ranges. To account for all these contingencies, the data model includes not only a field to hold the qualitative or quantitative result but also a field to store a modifier indicating whether the result is a text value, or, if numerical, is equal to the provided value or less than or greater than the reported value. The reference range is split into two separate fields, NORMAL_LOW and NORMAL_HIGH, to express both ends of the range. Both fields are associated with a modifier field to reflect the possibility that a reference range could be greater than or less than a specified value. The units of the test result are expressed in another field and are populated with the same units as in the lab source system. In Year Two, the workgroup did not make attempts to normalize the different units associated with the same test.

b. Vital Signs Data Model

Vital signs are obtained from electronic health records. They are stored in the Vital Signs table in the MSCDM.

Mini-Sentinel adopted the vital signs format of the HMORN VDW. Vital signs data are linkable to other data in the data model by the Patient Identifier variable. A single MEASURE_DATE and MEASURE_TIME is associated with each vital sign test. Each record includes fields for all possible vital signs measured at a particular time, including Height, Weight, Tobacco use, Tobacco type, Diastolic blood pressure (BP), Systolic BP, BP type, and Position. Not all records include responses for every possible vital sign. In

[iii] LOINC is a coding system used for identifying specific medical laboratory names in electronic health records.

particular, blood pressure values are commonly recorded, but not every record that reports a blood pressure measurement also reports a smoking value, which is typically recorded on a more sporadic basis.

Since the vitals data all come from a standardized source, the reported Height values are all in inches or are already converted to inches in the source data. Weight is provided in pounds. Tobacco use is coded as current user, never, quit/former user, passive, environmental exposure, not asked, or conflicting. Tobacco type can be Cigarettes only, Other tobacco only, Cigarettes and other tobacco, or None. The BP Type can be Rooming (measured by the triage nurse) or Orthostatic, Multiple, or Extended, which all imply physician follow-up measures. Position represents the Sitting, Standing, Supine, or Unknown position of the patient at the time the blood pressure was measured.

3. Year Two Data Checking

In addition to developing and populating the clinical additions component of the MSCDM a set of programs were designed to test the basic adherence of the data to the requirements of the model. This set of programs, or data checks, also describes the contents of the data:

- Count of total unique individuals with at least one lab test (of any type)
- Count of unique individuals by calendar year with at least one lab test (of any type)
- Count of unique individuals with specific lab test
- Count of unique individuals with specific lab test per calendar year
- Count of unique individuals with at least one lab test with coverage at some time
- Count of unique individuals with at least one lab test during period of eligibility
- Number of individuals and number of tests drawn outside enrollment period
- Counts of tests per LOINC code per calendar year
- Counts of tests per result unit per calendar year
- Counts of tests per routine/stat category per calendar year
- Counts of tests per inpatient/outpatient location status per calendar year
- Counts per test and year where “result num” is not missing
- Counts per test and year where test order date/specimen collection date and time/result date and time are available
- Descriptive statistics of test results

Results of these checks provide insight into the overall availability of lab values on patients seen within Data Partner facilities and on the large subset of patients who have labs during periods of enrollment for which the MSOC can correlate these labs with the patients’ other clinical data in the “main” data model. The results also highlight differences in the availability of inflections on lab results, such as the inpatient/outpatient status of the patient or the routine/stat status of a lab across the different Data Partners.

For vital signs data, an analogous set of checks that calculated the number of instances where blood pressures, weights, heights, and smoking are recorded. For vital signs with numerical results (SBP, DBP, weight, and height) means, standard deviations, and interquartile ranges were calculated

4. Lessons Learned

The substantial variation in representation of laboratory data in source systems posed considerable difficulty in standardizing the representation of laboratory test results in the MSCDM. Even a nominally simple test like blood glucose was challenging given the variety of glucose test types: venous, arterial, finger stick, fasting versus random, glucoses drawn in glucose tolerance tests, and glucoses associated with non-blood sources. The variety of ways the glucose test type was stored in the source systems made the task even more difficult. At times, it was necessary to infer the test type from an abbreviated test name. Other problems included different or missing test units for the same test. The D-Dimer test was particularly problematic. Both qualitative and quantitative results are provided with different associated ranges of normal for the results. The lesson learned is that the availability of lab results from source data and a well-defined target data model do not automatically make the process of data mapping simple. Substantial resources, time, and local expertise are required to make best use of available data, and each laboratory test must be investigated in detail to enable effective use in the Mini-Sentinel distributed environment.

It was not possible to determine the completeness of laboratory test information for individual health plan members. Some tests are processed by laboratories whose results are unavailable to the Data Partners. It is difficult to assess the degree to which lab results are missing, since associated billing data for laboratory studies is not consistently available or sufficiently detailed to determine that a test was performed.

Additional data checks that can provide some insight into completeness for individuals will assess the number of lab results per individual overall and within a year of the first test result in a given calendar year. These checks would provide ranges of observed frequency and timing of laboratory tests for specific groups of individuals.

B. OTHER REVISIONS TO THE MSCDM

Two other types of revisions were made to the MSCDM v1.1 in Year Two: 1) minor clarifications to the text, and 2) inclusion of all summary tables as standard elements of the Mini-Sentinel Distributed Database (MSDD). These modifications led to the creation of the MSCDM v2.0 in December 2010.

1. MSCDM Tables: Text Revisions

These revisions were necessary due to ambiguity in the descriptions of some fields (e.g., Discharge_Disposition and Discharge_Status fields in the Encounter Table should only be populated for Hospital Inpatient and Institutional encounter types). Such ambiguity created data transformation errors at some Data Partner sites. These errors were detected by the MSOC team through Data Quality Checks but caused delays in MSDD implementation at the Data Partner sites.

2. MSCDM Tables: Additions

The data dictionary for the nine Mini-Sentinel summary tables was added to the MSCDM for completeness and transparency. The summary tables are created using a distributed program that executes against the MSCDM utilization, enrollment, and demographic tables.

3. Discussions for Future Enhancements to the MSCDM

Experience in responding to FDA queries and needs identified items for addition to the MSCDM. The highest priority enhancements related to the enrollment table.

These fields were proposed for the enrollment tables:

a. Exclusion flag that indicates whether members' data is available for chart abstraction: some plans restrict use of administrative data for such activities and the ability to exclude them would improve data extraction efficiency.
b. Plan indicator: e.g., PPO, POS. These can be used as confounders in various analyses.
c. Primary/secondary insurance indicator: This could inform the MS investigator on data completeness; the assumption is that Data Partners only include data for members for whom they are fully financially responsible and hence all medical care information is captured in the MSDD and available for Mini-Sentinel activities.
d. Most recent patient zip code (3 digits): Adding this to the enrollment table (rather than the demographic table) allows for patients who move during the enrollment period, without adding complexity to the demographic data.

Other topics that deserve attention include, 1) adding a table containing detailed information on provider of services (e.g., specialty), 2) investigation of validity and meaning of the Primary Diagnosis flag at various Data Partners, and 3) assessment of the impact of primary/secondary insurance indicators on data completeness, in particular regarding the 65 and over age population.

C. LESSONS LEARNED AND SUGGESTIONS FOR FUTURE WORK

There is an ongoing challenge in achieving and maintaining an appropriate balance between creation of a broadly useful data resource and ensuring that goals are reasonably achievable. Firstly, discussions regarding expansion of the MSCDM are most productive when focused on specific data elements in specific delivery settings. A general discussion of expansion of laboratory data, for example, ignores the fact that data streams for inpatient labs are separate from those for outpatient labs. The context in which the data will be used informs the content, so presenting the Data Partners with very specific and structured questions regarding expansion topics allows them to respond with more thorough and helpful information. Secondly, exploration of the feasibility of accessing specific data streams can be a time-intensive undertaking for the Data Partners. In the future, prioritization of potential expansion targets will be considered essential.

Inclusion of additional clinical data elements is most efficient when there is a focus on specific cohorts for whom the specific data will be relevant. Finally, the selection of target laboratory tests to add each year led to selection of feasible but not strategically important data types. In the future, it will be helpful to identify a three-year plan for MSCDM expansion. This plan will reflect the FDA's strategic priorities for MSCDM expansion and identify the steps needed to ensure that the FDA's priorities will be met.

IV. MINI-SENTINEL DISTRIBUTED DATABASE

A. DATA QUALITY ASSURANCE AND CHARACTERIZATION

1. Overview

All data transformed by the Data Partners into the MSCDM were checked through the use of standard programs/data characterization code developed by the Mini-Sentinel Operations Center and refined through feedback from the Data Partners. The Data Partners each ran the data characterization programs on their local implementation of the MSCDM after each Extract-Transform-Load (ETL). The ETL process is described in detail in our Year 1 report (http://minisentinel.org/data_activities/details.aspx?ID=128). The data review procedure included: 1) steps culminating in detailed documentation of the data available at each Data Partner and 2) 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) Implementation and Reporting of ETL
    a. Data Partner execution of data characterization code provided by MSOC
    b. MSOC review of data characterization output consisting of datasets and log files, revise ETL as necessary, re-run data characterization code
    c. MSOC review of data characterization output, within and across sites and within and across ETLs
    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 their timeline (including revised ETL)
2) Acceptance of the ETL

2. Data Characterization Specifications

The Mini-Sentinel program relies on the comprehensiveness and quality of the data available in the Mini-Sentinel Distributed Database (MSDD). The MSOC works closely with each Data 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 and report address areas such as missing data, invalid values, invalid date ranges, and internal inconsistencies. Issues identified in the report are discussed with Data Partners and resolved on a case-by-case basis. The design and the scope of the data characterization programs take into account:

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

- It is vital that the tables created match the defined Mini-Sentinel requirements.

The data characterization programs are run on each ETL of MSCDM data. The data quality activities are organized into three levels of data characterization, based on the type of checks being performed. A description of the data characterization approach and the findings accompanies this report and can be found under the Data tab of 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. **Table 1** 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	DIS1.1.1
	PatID	Must be non-missing	DIS1.1.2
2	RxDate	Must be a SAS date value of numeric data type	DIS1.2.1
	RxDate	Must be of SAS length 4	DIS1.2.2
	RxDate	Must be non-missing	DIS1.2.3
3	NDC	Must be character data type	DIS1.3.1
	NDC	Must be exactly 11 characters in length	DIS1.3.2
	NDC	Must be non-missing	DIS1.3.3
	NDC	Must only contain digits from 0-9 (i.e., no space or other characters)	DIS1.3.4
4	RxSup	Must be a SAS date value of numeric data type	DIS1.4.1
	RxSup	Must be of SAS length 4	DIS1.4.2
	RxSup	Must be non-negative	DIS1.4.3
5	RxAmt	Must be a SAS date value of numeric data type	DIS1.5.1
	RxAmt	Must be of SAS length 4	DIS1.5.2
	RxAmt	Must be non-negative	DIS1.5.3

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, the unique person identifier can 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 should occur only once. Further, the person identifier 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, corresponding demographic information exists. **Table 2** 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		The combination of PatID, Enr_Start, Enr_End, and DrugCov must occur only once in the table	ENR2.0.0
2	PatID	Must have a corresponding value in the Demographic table	ENR_DEM2.1.1
3	Enr_Start	Must be earlier than or equal to Enr_End	ENR2.2.1
	Enr_Start	In combination with PatID, MedCov, and DrugCov, must occur only once in the file	ENR2.2.3
4	Enr_End	In combination with PatID, MedCov, and DrugCov, must occur only once in the file (implemented in Year Two)	ENR2.3.4

After each ETL, Level 1 and 2 data characterization reports are sent to MSOC for review. MSOC staff 1) manually inspects the level 1 and level 2 reports, 2) identify data anomalies and reported them back to Data Partners, and 3) discuss the potential for developing ranges of acceptable error threshold rates. All anomalies are reported to the Data Partners to determine whether the issue can be fixed or is part of the underlying data. If necessary, a plan for remedying the anomalies is developed—this typically entails a correction in the subsequent data extract—or the anomaly is documented so it does not 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. For example, trends in the number of outpatient dispensings per person or the rate of hospitalizations commonly following similar patterns across partners, and any obvious divergence from the general trend requires investigation. Periods of sharp increases or decreases 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 both within an ETL and across Data Partners’ ETLs. Examples of trends within a single ETL include:

- Outpatient pharmacy dispensing per member per month
- Hospital admissions per member per month
- Total dispensing per month
- Total encounters by encounter type per month

Examples of trends across ETLs, include number of members and number of records—both of which are expected to always increase with each ETL and with the addition of new data. 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.

As an example, several Level 3 data characterizations for the dispensing table are:

- 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 all data anomalies, 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. Reporting

Results of the data characterization activities are shared with the Data Partners. Two companion documents—the *Data Quality and Characterization Procedures and Findings Report* and the *Mini-Sentinel Distributed Database Year Two Summary Report*—provide details of the data checking and characterization activities and results. These reports accompany this report and can be found on the Mini-Sentinel public website (www.mini-sentinel.org/data_activities).

B. INCORPORATION OF NATIONAL DATA STANDARDS AND CONTROLLED TERMINOLOGIES

The MSOC is committed to adoption and use of relevant national content and vocabulary standards related to electronic health care data. The two primary activities under this task are incorporation of standards into the MSCDM and engagement standards bodies, as directed by FDA.

Incorporation of Standards into the MSCDM: Incorporation of national electronic health data standards into the MSCDM entails three key components: 1) identification of relevant standards based on the operational characteristics of the Mini-Sentinel distributed data system; 2) identification of the electronic health data standards used by the Mini-Sentinel Data Partners, and 3) incorporation of relevant and available standards into the MSCDM.

As a distributed health data network, the Mini-Sentinel approach requires all Data Partners to conform to a single data model that can accommodate longitudinal health data going back as far as 2000. The common data model enables a fully distributed analytic approach that allows a single analytic program to execute identically at each Data Partner site. The distributed analytic requirement also requires adoption of a common data model that all Data Partners can implement within their existing electronic health data systems. Currently, the Mini-Sentinel Data Partners use a limited yet comprehensive set of coding terminologies to capture medical encounter, pharmacy dispensing, demographic, laboratory results, and health plan enrollment information.

To facilitate adoption and use of the MSCDM, the MSCDM was developed as a simplified version of data models used in similar distributed networks such as the HMO Research Network. As described in the Mini-Sentinel Year 1 Common Data Model report (http://www.mini-sentinel.org/data_activities/details.aspx?ID=128) , the common data model was developed over several months of iterative discussion with the Mini-Sentinel Data Partners and informed by the Mini-Sentinel Common Data Model guiding Principles (http://www.mini-sentinel.org/work_products/Data_Activities/Mini-Sentinel_CommonDataModel_GuidingPrinciples_v1.0.pdf) . The current version of the MSCDM is available online (http://www.mini-sentinel.org/data_activities/details.aspx?ID=105). The MSCDM was designed to accommodate other coding terminologies such as ICD-10. The key data areas included in the MSCDM are listed below, with the national standards used within each data area.

Diagnoses: Diagnoses are captured using International Classification of Diseases, 9th Revision (ICD-9-CM)[iv] codes recorded during inpatient and outpatient medical encounters. Depending on the Data Partner, the diagnoses are recorded on health insurance claims submitted for reimbursement and/or in electronic health record systems for the Mini-Sentinel partners that operate as integrated delivery systems. Each of our Data Partners uses this standard terminology.

[iv] http://www.cdc.gov/nchs/icd/icd9cm.htm

Procedures: Medical procedures are captured using ICD-9 procedure codes and Healthcare Common Procedure Coding System (HCPCS)[v] codes, including Current Procedural Terminology-4 (CPT-4)[vi] codes, recorded during inpatient and outpatient medical encounters. Procedures captured using these terminologies include a wide range of medical interventions, ranging from well-child visits to immunizations, to drug infusions and inpatient surgical procedures. In addition, both CVX and MVX codes describing vaccine administration and manufactures have been adopted for vaccine-specific work involving immunization registries. Each of our Data Partners use ICD-9 procedure and HCPCS codes.

Outpatient Pharmacy Dispensings: Pharmacy dispensings are identified using National Drug Codes (NDCs) that are recorded by pharmacies at the point of distribution. Each of our Data Partners uses this standard pharmacy dispensing terminology.

Death and Cause of Death: The death and cause of death tables use ICD-9 and ICD-10[vii] diagnoses codes. These are the codes available through the source of the information, typically State death registries.

Laboratory Results: Laboratory results in the MSCDM are captured using Logical Observation Identifiers Names and Codes (LOINC) and test name. Data Partners may also use local codes and procedure codes to help identify specific lab test results. Our Data Partners use a mixture of LOINC and local codes to identify laboratory tests. To the extent possible, LOINC codes will be used to identify laboratory results.

Some commonly referenced coding terminologies such as RxNorm, CDISC, and the Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT) are not currently included in the MSCDM. Although these and several other potential relevant coding terminologies are increasingly being adopted by electronic health record systems and some health insurers, the Mini-Sentinel Data Partners do not uniformly capture information using those terminologies. The MSOC will continue to work with FDA and the Data Partners to assess inclusion of these and other standards as possible.

Engagement with National Standards Bodies**.** There are a wide range of health data standards initiatives supported by public and private partnerships in the US and abroad. These activities and the growing adoption of electronic health record systems have the potential to improve semantic and syntactic interoperability and expand the range of potential Data Partners for Mini-Sentinel. For instance, the Meaningful Use standards[viii] related to data capture and transmission promulgated by the Office of National Coordinator for Health Information Technology (ONC) have the potential to standardized data content and vocabularies, thereby enabling distributed querying of a broad range of medical practices and health facilities.

Not all health data standards are relevant to Mini-Sentinel, especially within the context of the Mini-Sentinel distributed querying approach. All uses of Mini-Sentinel are "secondary uses" of electronic health data and are therefore not directly related approaches and standards targeting point-of-care

[v] http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html

[vi] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.page?

[vii] International Classification of Diseases, 10th Revision; http://www.cdc.gov/nchs/icd/icd10.htm

[viii] http://www.healthit.gov/policy-researchers-implementers/meaningful-use

transmission of health data. So although initiatives such as health information exchanges have potential application to the MSCDM, all standards are assessed within the context of the needs of the Mini-Sentinel distributed data approach and the needs of the FDA within the system.

FDA has identified the ONC Standards & Interoperability (S&I) Framework[ix] as a key binding point for engagement related to Mini-Sentinel data standards, specifically the ONC Query health Initiative. Several members of the MSOC staff, and associated vendors, are actively engaged with the S&I Framework activities and will remain engaged with those activities. Our involvement has includes face-to-face meetings with S&I Framework staff, webinars, and participation on several working groups. These activities will continue in Year Three.

V. MINI-SENTINEL ANALYTIC TOOLS

A. MODULAR PROGRAMS

1. Overview of Modular Programs

During Year One, the Mini-Sentinel Operations Center developed four modular programs to facilitate rapid response to common queries by each Data Partner. Each program has several required input parameters (e.g., exposures 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) and includes 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.

[ix] http://www.siframework.org/

- ***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.

2. Modular Program Revisions

During Year Two, the Mini-Sentinel Operations Center: 1) added three new modular programs, and 2) enhanced all modular programs with new features and capabilities. The additions and enhancements resulted from input from FDA and Mini-Sentinel Partners, as well as experience acquired in using the programs.

a. New Modular Programs

Summary description and features for each new modular program are described below and full documentation as well as SAS code are available online. The new modular programs for Year Two are:

- **Modular Program 5 – Background Rate of Health Outcomes of Interest (HOIs) and Exposures**: Standard output provides prevalence and incidence rates of HOI use among at-risk populations. **For example**: rates of type 2 diabetes among MSDD populations broken down by various age groups, sex, and year.
- **Modular Program 6 – Drug and Procedure Use Following a Diagnosis**: Standard output provides rate of drug/procedure use among at-risk, diagnosed populations. Metrics on time to first drug/procedure use (from diagnosis index date) will be provided. Optional features include: ability to restrict to incident diagnosis and/or naïve-to-treatment (i.e., drug and/or procedure) patients, and ability to add pre-existing conditions. **For example**: rate of oral antidiabetic medication use following first diagnosis of diabetes; rate of hip replacement surgeries following a fall at home among female patients aged 65+ with osteoporosis.
- **Modular Program 7 – Most Frequently Used Codes Prior & Post Index Event**: Detailed characterization of the "Top XX" (user-defined) most frequently used diagnosis, procedure, and drug codes during a user-defined period prior to and post event index date. Event of interest can be defined using any type of code, and results are provided for both prevalent and incident users of the index event code(s). Standard output provides "Top XX" rankings using both number of users and events, and rates for both prevalent and incident use of each most frequently used codes are provided. **For example**: Top 10 drug codes used before and after a heart transplant.

b. Enhancements to Modular Programs

These enhancements were made to all Modular Programs.

i. New Features and Standard Output

   - Allows exposures of interest to be defined using procedure codes (HCPCS or ICD-9-CM), in addition to using NDC codes.

- Denominators generated as part of default output: For each stratum, the relevant at-risk population is identified and serves as denominator to be used by FDA investigator in rate calculations; "at-risk" population defined as all MSDD members with all selection criteria/parameters of interest, i.e., age, minimum length of enrollment/washout, pre-existing condition, or diagnosis of interest (as needed).
- For incident exposures or HOIs of interest: ability to capture either 1) the very first incident instance (if any) during the period of interest, 2) multiple (i.e., all) incident instances during the same period of interest, or 3) the first one ever (provided availability of data before period of interest).
- Look-back period for pre-existing conditions using a fixed number of days before index date now implemented for all modular programs (when relevant)

ii. Technical Enhancements

- Major revisions to structure of the modular program input files used to pass codes of interest and various parameters of the modular program data requests (e.g., care setting, length and type of look-back period, minimum length of treatment episode, allowed gaps in days supply). Although adding complexity to the structure of input files, these changes allow greater flexibility in defining the modular program data requests (e.g., care setting of interest for an outcome can now vary across different codes) and allow the MSOC to bundle multiple scenarios into fewer modular program runs.
- Multiple technical revisions to the SAS code make it more efficient and improve overall run time and use of disk space by the Data Partners.

c. Testing Phase

Before being used in production mode, all new modular programs go through a rigorous internal Quality Control process by developers: test cases are manually built (and documented) to stress-test the modular program SAS code to ensure it only selects desired cohorts of interest and generates the expected output. Once a modular program has passed the Quality Control process, it is then shared with at least two Data Partners for additional testing and validation. Documentation is revised to ensure compliance with specification. Any feedback or suggested modifications from the Data Partners are handled by the MSOC; then the modular program is shared with all Data Partners. The Partners then 1) run it using a test scenario to confirm it can run within their local IT environment, 2) inspect output and log files to confirm they are valid and error-free, and 3) authorize the MSOC to routinely use it with FDA data requests. MSOC accepts a Modular program for use once all Data partners have approved it.

d. New Input Forms

Due to the increased complexity of the input file structure, the MSOC development team designed a new Modular Program Query Interface (MPQI). Phase I of this project took place during Year Two and consisted of designing a new tool for FDA to provide the MSOC with the information on modular program data requests. To date, a beta version has been tested by various MSOC staff, but additional cycles with software developers are required before sharing with the FDA team. Phase II, which is planned for Year Three, aims at migrating the static, beta version of the MPQI tool to a web-based platform. This will serve two purposes: 1) easy integration of the MPQI with the new Mini-Sentinel single-sign-on web-based platform and 2) streamlining the data request process between FDA and MSOC by avoiding communication errors.

B. SUMMARY TABLES

1. Summary Table Revisions

During Year Two, two separate types of enhancements were implemented: 1) major revisions to the SAS programs generating the summary tables and 2) addition of incident tables. These are described below.

a. Prevalent Tables

A new, more efficient method to create summary tables was developed. It consists of a single SAS program with nested macros suited for all current Data Partners and the raw MSDD tables are now read only once each, recycling intermediate data files for multiple purposes. This new program was tested by three different Data Partners prior to being approved for production mode. Actual release of the new program is planned for Year Three.

In addition to major revisions to the SAS program, these minor modifications were made to the summary tables to simplify and enhance interpretation of the data:

- Addition of a new column to the enrollment summary table that contains the number of days of enrollment in a given year
- Sex category "Unknown" removed from any tables
- Addition of an "Any" care setting category for the procedure and diagnosis summary tables, aggregating the information from all other care settings (Inpatient, Outpatient, Emergency Department)
- Whenever relevant, diagnosis, procedure, or drug codes that could not be identified by the Mini-Sentinel lookup tables are bundled into a "Did not match" category instead of being excluded.
- The code now generates summary tables in text-delimited files only; proprietary formats (e.g., Microsoft® Access®) are no longer used.

b. Incident Tables

A SAS program using the same flexible approach developed for the prevalent tables was developed and tested by the Data Core for creation of summary tables for incident counts (events and members) for three different types of outcomes. The focus of the development of incident tables is on the following three definitions, as they are most useful and informative to the FDA:

i. Incident outcome by 3-digit ICD-9-CM diagnosis code
ii. Incident exposure by generic name
iii. Incident exposure by drug category

Because identifying incident diagnoses at a more granular level (e.g., 4- or 5-digit ICD-9-CM diagnosis codes) generates information that is difficult to interpret and most likely not useful to the FDA, only the 3-digit code categories are included in the incident tables.

For all three types of incident outcomes, a default set of results will be generated using the same three look-back periods of 90, 180, and 270 days, allowing for a default maximum of 45 days of enrollment.

For queries on incident outcomes by diagnosis code, the number of encounters with diagnosis of interest in the 90, 180, and 270 days following the incident outcome also will be reported. For these queries, it will be possible to further restrict the care setting of interest, although the look-back period will exclude valid outcomes in any care settings. For queries on incident exposures (generic name and drug categories), the number of dispensings and length of first treatment episode (in days) will be reported. No restriction on the length of first treatment episode will be applied and a standard 15-day gap period will be allowed.

A fourth incident table was implemented in this new program in preparation for creating a table with incident counts of health outcomes of interest. The purpose of this is to query outcomes slightly more complex than the regular one-sided queries defined above. For instance, an outcome could be defined as a list of specific ICD-9-CM diagnosis codes (e.g., a mix of 3-, 4- and 5 digit codes) or requiring two or more instances of a specific diagnosis code (e.g., two independent outpatient visits for diabetes mellitus).

C. MINI-SENTINEL DISTRIBUTED QUERY TOOL

1. Overview of Query Tool

The FDA Mini-Sentinel Distributed Query Tool and Portal allows Mini-Sentinel Operations Center staff to create and securely distribute “queries” to Data Partners and enables Data Partners to review, execute, and return the results of those queries via a secure web Portal. The distributed architecture allows Data Partners to maintain control of their data and all its uses. 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)[x] compliant TIER III data center. The Mini-Sentinel Query Tool and Portal is based on the PopMedNet^{TM} software platform. The implementation design and architecture are detailed in the Mini-Sentinel Distributed Query Tool: System Description and Technical Documentation.

The Mini-Sentinel Distributed Query Tool (**Figure 2**: screenshot of the login screen) currently allows rapid distributed querying of preprocessed summary tables. Using preprocessed summary tables speeds the querying process because it:

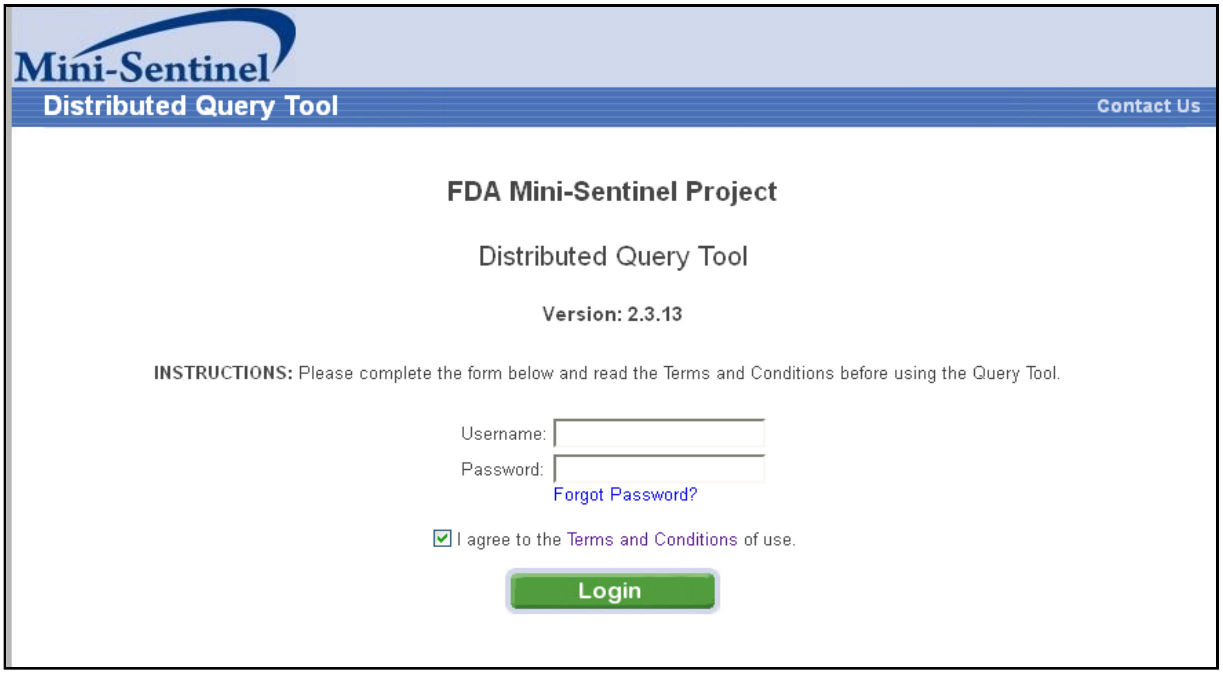
- Obviates the need to access person-level data, thereby avoiding local privacy and patient–confidential, data-release authorization procedures
- Allows use of a simple menu-driven querying tool interface
- Allows non-technical Data Partner staff to execute and return results
- Avoids the need to specify, 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 nine query types that represent prevalence counts of diagnoses, procedures, and drug exposures. For diagnoses

[x] http://csrc.nist.gov/groups/SMA/fisma/index.html

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. 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. The Mini-Sentinel Distributed Query Tool architecture is consistent with the standards promulgated by the Standards and Interoperability (S&I) Framework supported by the Office of the National Coordinator. The Mini-Sentinel staff is working actively with the S&I Framework Query Health team to communicate the lessons learned from implementation and operation of the Mini-Sentinel distributed querying system. These lessons include the need for detailed technical documentation and user training material, the need for security documentation and clearance by each Data Partner, and barriers faced related to installation of external software on local computers.

Figure 2. Distributed Query Tool Login Page



FDA Mini-Sentinel Project

Distributed Query Tool

Version: 2.3.13

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.

2. Network Setup and Training

The distributed querying network was established in partnership with the MSOC, Mini-Sentinel information technology vendors, and the Data Partners. The implementation process involved establishment of a "staging" network that allowed testing of governance, security, and querying capabilities of the software platform, development of a series of user manuals, and implementation of a production site to allow secure distribution of queries. Initial use of the system by Data Partners led to several revisions and enhancements that were implemented during Year Two.

3. Testing

All Data Partners were given login credentials to the Mini-Sentinel staging site to enable them to investigate and test the system, set permissions, and otherwise evaluate the acceptability and usability of the proposed software platform and implementation. MSOC provided Data Partners with testing scripts, role-based user manuals (e.g., *DataMart Administrator Manual*, *Investigator Manual*, *Overview and Technical Document*) and detailed setup instructions. MSOC also provided one-on-one site support for setup, site administration, and technical issues through telephone calls, webinars, and email. Technical questions about the software and security architecture were answered by the Mini-Sentinel IT vendor responsible for creating and operating the system.

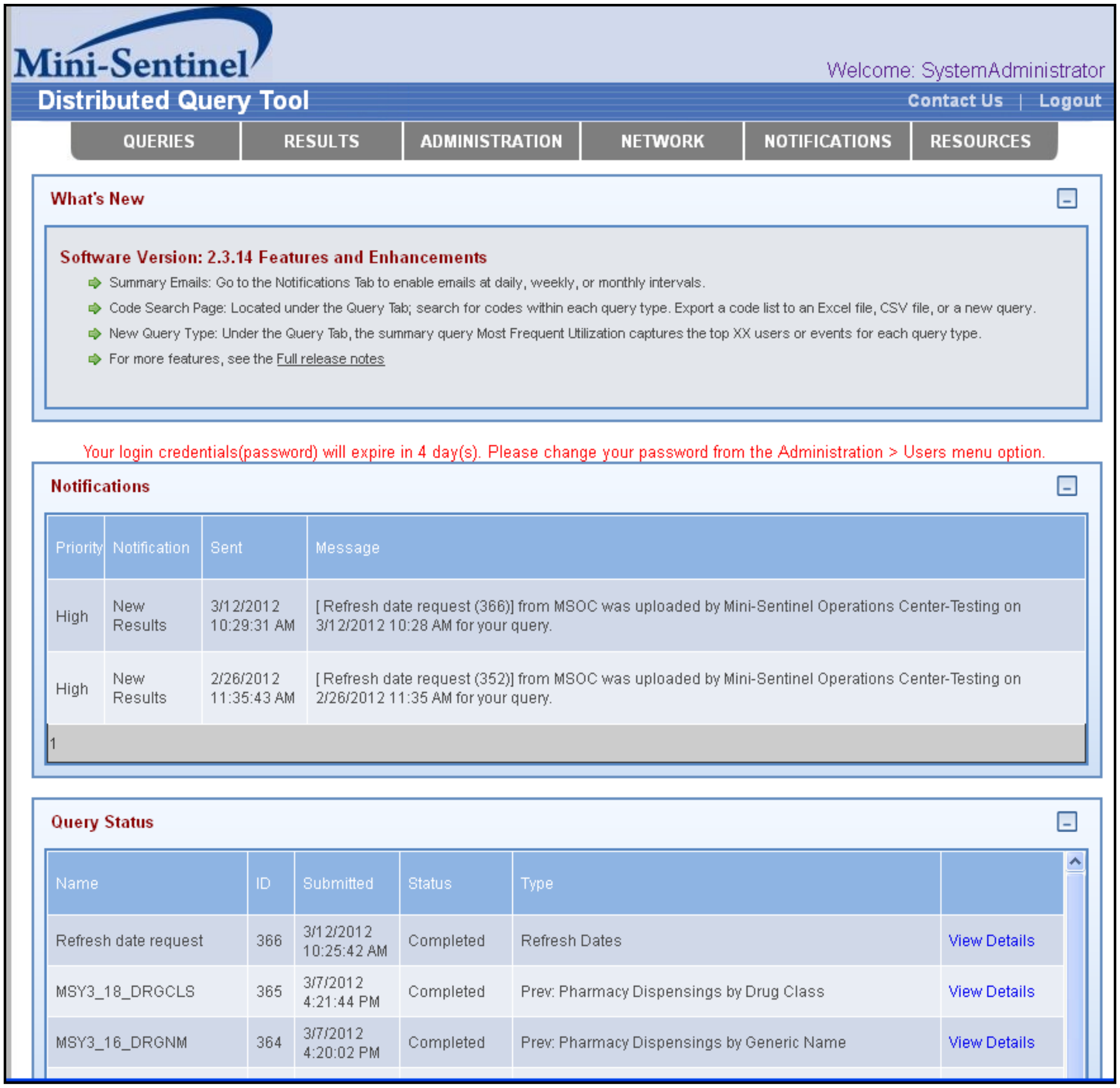
Anew system role, Group DataMart Administrator, was created based on feedback from Data Partners. The Group DataMart Administrator role has the right to review and approve query results from all organizations in the group and return aggregate or individual site results.

Formal testing of the system involved setting up all Data Partners on the staging server and using that staging network to send queries to each Data Partner (to execute against fake data) and to have each Partner respond to the queries. The MSOC reviewed sample results to confirm proper use of the query tool. Once the Data Partners and MSOC were comfortable with the functionality of the system on the staging server, all partners were transitioned to the secure production server. Once transitioned to the production server, MSOC issued test queries for each query type to ensure the system was functional and operating as expected. The MSOC and the software developer provide ongoing support as new sites and users are added, questions arise, and enhancements are requested and developed. The MSOC tests new versions of the software and creates Release Notes to inform the Data Partners of the changes. By the end of Year Two, 16 Data Partners were using the query tool.

4. Portal Enhancements

A role-based landing page (**Figure 3**) was implemented to provide each user with recent query tool activity upon login. This page provides collapsible lists of recent notifications sent and query statuses. The landing page also contains general information about recent software updates, new features, warning notes regarding password expiration, and Release Notes. A user can return to this page at any time by clicking on the banner. In addition, the architecture of the underlying querying platform was updated and revised to enable more efficient modifications and enhancements, and expansion of the functionality of the tool.

Figure 3. Distributed Query Tool Landing Page



What's New

Software Version: 2.3.14 Features and Enhancements

- Summary Emails: Go to the Notifications Tab to enable emails at daily, weekly, or monthly intervals.
- Code Search Page: Located under the Query Tab; search for codes within each query type. Export a code list to an Excel file, CSV file, or a new query.
- New Query Type: Under the Query Tab, the summary query Most Frequent Utilization captures the top XX users or events for each query type.
- For more features, see the Full release notes

Your login credentials(password) will expire in 4 day(s). Please change your password from the Administration > Users menu option.

Notifications

Priority	Notification	Sent	Message
High	New Results	3/12/2012 10:29:31 AM	[Refresh date request (366)] from MSOC was uploaded by Mini-Sentinel Operations Center-Testing on 3/12/2012 10:28 AM for your query.
High	New Results	2/26/2012 11:35:43 AM	[Refresh date request (352)] from MSOC was uploaded by Mini-Sentinel Operations Center-Testing on 2/26/2012 11:35 AM for your query.

1

Query Status

Name	ID	Submitted	Status	Type	
Refresh date request	366	3/12/2012 10:25:42 AM	Completed	Refresh Dates	View Details
MSY3_18_DRGCLS	365	3/7/2012 4:21:44 PM	Completed	Prev: Pharmacy Dispensings by Drug Class	View Details
MSY3_16_DRGNM	364	3/7/2012 4:20:02 PM	Completed	Prev: Pharmacy Dispensings by Generic Name	View Details

5. Query Enhancements

a. Query Functionality

- **Increased Metadata**. New drop down menus and text fields were added to further explain the purpose and nature of each query. The query information includes project description, relevant Mini-Sentinel activity requesting the information (e.g., Base contract, Task Order 1, workgroup), priority level (e.g., high, normal, low), and due date.
- **Incident Query Type**. New query type was developed and tested based on the same schema as the prevalence query type.

- **Most Frequent Utilization Query Type**. New query type was created to calculate the most frequently observed drug, diagnosis, and procedure codes by year, age group, and sex. This provides an assessment of the "Top 100" observed codes for each stratum.
- **Code Search Page**. Search page added to enable text string searches for relevant codes (e.g., diagnosis, procedure) for each of the nine available query types. Users can search codes and export to save as an Excel/CSV document or copy to a clipboard for use on the Summary Query page (i.e., submitting a query).

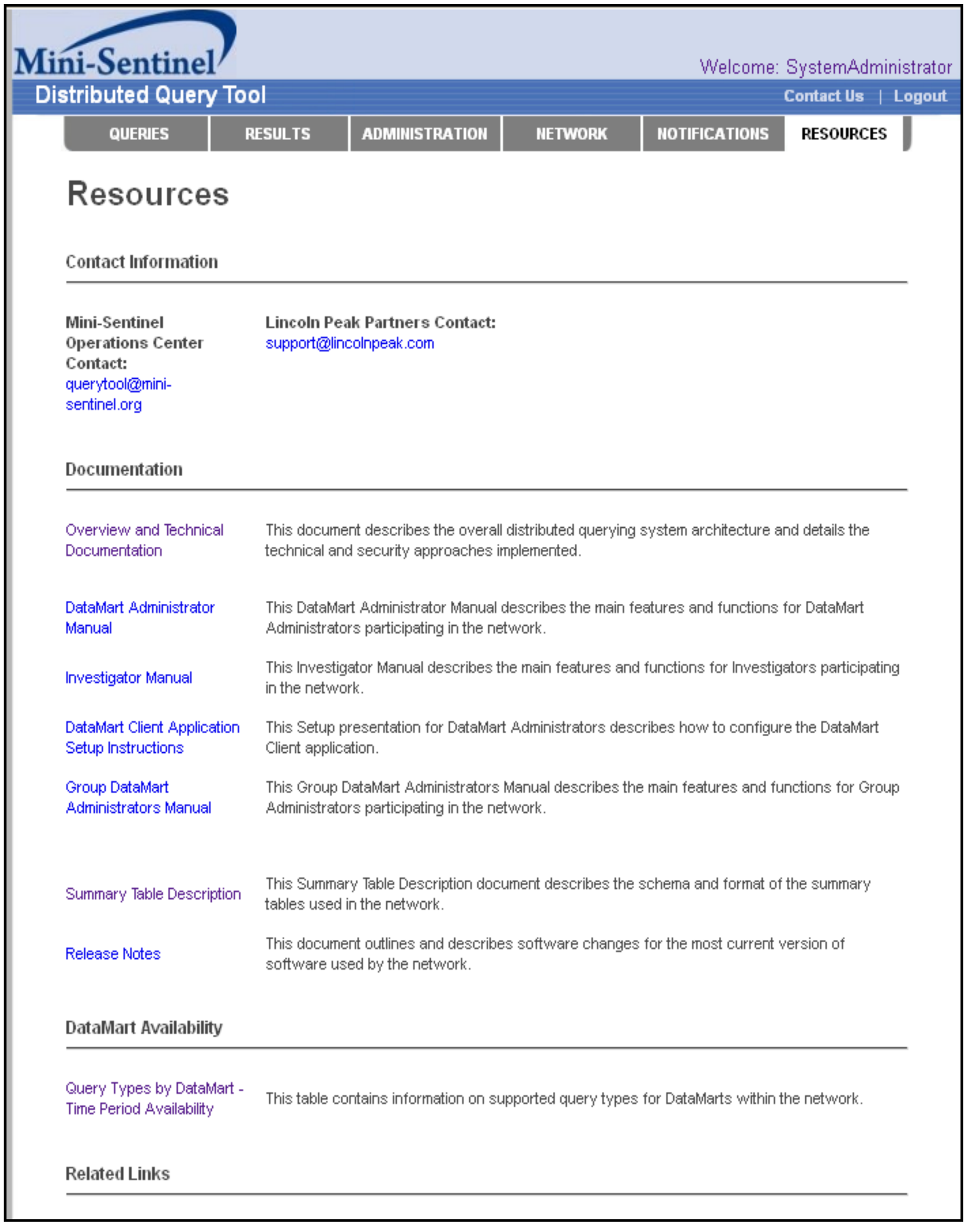
b. Result Table Enhancements

- **Rate Columns**. New prevalence rates columns were added to the results tables using the enrollment data for the calculations. Additional information for each query type is:
  - *Medical Queries*: HCPCS Procedures, ICD-9 Diagnoses and Procedures
    - Total Enrollment in Strata (Members)—Includes members with both medical and drug coverage plus those with medical coverage only
    - Prevalence Rates (Users per 1000 enrollees)
    - Event Rate (Events per 1000 enrollees)
    - Events per Member
  - *Drug Queries*: Pharmacy Dispensings by Generic Name and Drug Class
    - Total Enrollment in Strata (Members)—Includes members with both medical and drug coverage plus those with medical coverage only
    - Prevalence Rates (Users per 1000 enrollees)
    - Dispensing Rate (Dispensings per 1000 enrollees)
    - Days per Dispensing
    - Days per User

c. User Administration Functionality

- **Improved Display for Users.** Added Select All features, drop down menus, and column filter options.
- **Increased Viewable User Metadata.** All users can now view additional information regarding other users in the network. For example, a "role" column was added to the Network Users list so all users could see the role of all other users.
- **Notifications.**
  - *Summary emails*: Users now have the option to determine the frequency of notification emails sent by the portal, reducing the number of emails users receive.
  - *DataMart Client Update Email*: Notifies users when software updates are required for the DataMart Client software.
  - *Network Message*: System administrators can create network-wide messages to alert users of software updates or system down-time.
  - *Resources*: New page (**Figure 4**) allows users to access the Mini-Sentinel user guides, software documentation, and general information about the Mini-Sentinel query tool network and contact information for help.
- **Supported Queries Table**. Table of supported query types including the available periods for which data are available by partner for the Mini-Sentinel network.

Figure 4. Query Tool Portal Resources Page



Resources

Contact Information

Mini-Sentinel Operations Center Contact: querytool@mini-sentinel.org	Lincoln Peak Partners Contact: support@lincolnpeak.com

Documentation

Overview and Technical Documentation	This document describes the overall distributed querying system architecture and details the technical and security approaches implemented.
DataMart Administrator Manual	This DataMart Administrator Manual describes the main features and functions for DataMart Administrators participating in the network.
Investigator Manual	This Investigator Manual describes the main features and functions for Investigators participating in the network.
DataMart Client Application Setup Instructions	This Setup presentation for DataMart Administrators describes how to configure the DataMart Client application.
Group DataMart Administrators Manual	This Group DataMart Administrators Manual describes the main features and functions for Group Administrators participating in the network.
Summary Table Description	This Summary Table Description document describes the schema and format of the summary tables used in the network.
Release Notes	This document outlines and describes software changes for the most current version of software used by the network.

DataMart Availability

Query Types by DataMart - Time Period Availability	This table contains information on supported query types for DataMarts within the network.

Related Links

- **User Manuals**. Role-based user guides, posted in PDF format.
- **Release Notes**. Each software version is accompanied by Release Notes that highlight the most recent changes to the query tool software and outline how to download new software updates.
- **New Query Tool Roles**. Three roles give different types of users varying level of access to the query tool portal:
  - *Group DataMart Administrator*: Reviews, aggregates, and releases results for a group of organizations, i.e. Kaiser Permanente.

- *Query Administrator*: Approves outgoing queries for an organization, useful for query budgeting. This role has the same rights and views as the Enhanced Investigator role for querying and viewing results.
- *Observer*: Can oversee an organization's activity on the query tool, i.e., what queries are sent and received by the organization and the status of these queries.

6. DataMart Client Software Enhancements

Several enhancements were developed to improve functionality of the DataMart Client software that is used by the Data Partners to review and execute all queries.

- **Metadata.** Additional query metadata includes due date, project description, priority level, and relevant Mini-Sentinel activity associated with the query.
- **User Interface**. Now allows Data Partner administrators to view all the query requests by query status and filter the query list by status, sent date, or due date.
- **Export**. New option to locally export query results.
- **Messaging**. Messaging capabilities now allow Data Partners to send notes back to query requestor when queries are rejected, on hold, or submitted.
- **Automation**. The automated execution functionality now automatically masks low cell counts when the automatic-reply setting is activated.
- **Log**. New log file captures program exceptions, warnings, and debugging information.

D. MINI-SENTINEL DATA CATALOG

1. Development and Use of the Mini-Sentinel Data Catalog

During Year Two, the MSOC designed the Mini-Sentinel Data Catalog (MSDC), a software system used to track data flows within the Mini-Sentinel distributed network. The development of the MSDC was motivated by the Operations Center's need for an easily managed system that captures and organizes data related to distributed data requests. Specifically, the MSDC was designed to allow the MSOC immediate access to key data about all Mini-Sentinel data queries and Data Partner activity.

Before sending any data request to the Data Partners, the MSOC creates a workplan for the request that includes pertinent information for the Data Partners (i.e., the type of request, requesting institution, program distribution date, due date for submission of results, and location on the Secure Portal where results should be uploaded). This workplan is directly entered into a template input form in the MSDC, and the SAS package that the Data Partners will use to execute the request is attached. Each workplan entered into the MSDC is also given a unique workplan ID. The MSDC later uses the workplan information and workplan ID to determine the status of the data request.

When a Data Partner uploads results files to the Mini-Sentinel Secure Portal, an email that includes the Data Partner name, file name, file location, and date of file upload is automatically generated and sent to the MSOC. The MSDC processes this email and incorporates the information into the Data Catalog. The MSDC uses the Data Partner name and the name of the file uploaded to automatically assign the file to a workplan ID. Once this occurs, the MSOC can inspect the file and choose to either accept or reject the file based on its contents. If the file is accepted, the MSDC catalogs the date of file submission, the Data Partner that submitted the file, and the location of the file on the MSOC's shared network drive.

Through this process, the MSOC can easily track and store information about the status of each data request and the activity of each Partner.

The MSOC can also extract key metrics from the MSDC through its report feature. When requested, the MSDC produces reports that allow the MSOC to see which Data Partners have and have not submitted results for any given data request, how often each Data Partner submits results past the expected due date, the quantity and type of each request sent to Data Partners over a specified period of time, and the overall status of each project. This reporting system provides the MSOC the ability to quickly and easily retrieve up-to-date information about data requests.

2. Lessons Learned and Future Work

During Year Two, the MSOC learned that it would be beneficial to enhance adherence of all parties, including the Data Partners, to file naming conventions, as the MSDC only recognizes pre-specified file names. It would also be beneficial to enhance the capabilities of the MSDC so that multiple filetypes can be tracked (e.g. SAS log files or Excel files), in addition to the .zip files currently tracked in the system. In the future, the MSOC would like to facilitate password synchronization for the MSDC with other operations center functions, which can be accomplished most readily by hosting the MSDC on the same server as the Secure Portal. Furthermore, the MSOC would like to incorporate summary table requests and data checking output into the MSDC tracking system and to use the MSDC to generate the workplan PDFs and data packages for all requests sent to Data Partners.

E. ELECTRONIC SUPPORT FOR PUBLIC HEALTH (ESP)

1. Introduction

One of the specific aims for Year Two of the Mini-Sentinel project was to develop a generalized, secure, efficient mechanism to extract, analyze, and update electronic health record vital signs and laboratory test results to populate Mini-Sentinel. The chosen approach was to use the existing open-source EHR public health surveillance platform known as Electronic Support for Public Health (ESP, www.maehi.org/what-we-do/hie/mdphnet/esp; http://esphealth.org/redmine/) as a source-data repository and map and to transfer vital sign and laboratory result records into Mini-Sentinel as a destination repository.

2. Fake Data Loading into ESP

An existing basic infrastructure for the creation of fake data in ESP was significantly extended to allow the creation of random but realistic datasets. The make_fakes utility was upgraded to create datasets, in an EPIC ETL data load format, for the Member, Visit, Provider, Tests Resulted, and Medications load schema. The interface to the data generation process was enhanced to allow the specification of how many patients should be generated and how many lab test results, medications, and encounters should be generated for each patient. A mechanism for developing random but realistic "fake lab test results" was implemented via a driver table that contained one row for each lab test for which it was desired to generate fake data, containing the native test name and a range of possible values. For quantitative tests, the range was expressed by Normal High, Normal Low, Critical High, Critical Low. For qualitative tests, the range was just the set of allowed categorical values. Parameters were set to determine the probability of generating a random last-test result, whether within or outside the normal range specified for the selected lab test.

Datasets, created by the make_fakes utility were then loaded into ESP via the load_epic utility to populate the ESP schema tables of emr_encounter, emr_labresult, emr_patient, emr_prescription, and emr_provider.

3. Mapping Between ESP and MS Data Models

A map was developed between the ESP emr_encounter table and the Mini-Sentinel vitals schema. Another map was developed between the ESP emr_labresult table and the Mini-Sentinel labs schema. A set of specific lab tests, all used in Mini-Sentinel, was identified as the basis for transferring lab records.

4. ETL Process and Tool—from ESP to Mini-Sentinel

With test data available in the ESP schema, a full ETL transfer process and associated tool to transfer this demographic, vitals, and labs data into the Mini-Sentinel tables (according to the schema specification in a postgres database) was created. The tool presented a Mapping Catalog containing all the tests to be mapped. Mapping was maintained in the tool in a database table and an application view of the data as Mapped Codes. The mapping feature's user interface allowed the user to see what ESP heuristic the code may have been mapped to previously and then guided the user to select one of the Mini-Sentinel test names in the Mapping Catalog. Lab tests that should be ignored were added to an Ignored Code table, which was also locally maintained by the tool in a database table.

A "synchronize" button in the application dynamically built a table of all unmapped codes and showed the user a table of the set of unique native codes in the ESP lab data minus the ones that were already mapped, minus the ones that were set to be ignored.

5. Batch Transfer

Data were transferred to Mini-Sentinel via the creation of a data transfer job that could be scheduled to run later or immediately. A configured system property let us choose whether to transfer all data from ESP or to match to the unique person identifier. Jobs were also segmented by lab or vitals type. Vitals transfers were subselectable by date range, while lab transfers could be partitioned by date range and/or lab test type. The outcome of transfers was loaded into demographic, vitals, and labs data tables in the Mini-Sentinel schema.

6. Conclusion

By using the make_fakes utility in ESP with the new tool developed to perform lab mapping and transfer of vitals and labs data from ESP, it was possible to populate Mini-Sentinel data tables efficiently.

F. LESSONS LEARNED

1. SAS Program Development and Testing

A major lesson learned during Year Two is the importance of recognizing and addressing the substantial heterogeneity of the Data Partners' IT environments, software and hardware types/versions available, and relevant staff experience.

- Software, programs, tools and instructions must be adjusted to a common denominator. For example, when developing SAS programs such as modular programs or summary tables, different operating systems (e.g., Windows vs. UNIX) or different internal memory settings will require more flexible software and must be kept in mind.
- For SAS program testing. Criteria to pick testing sites should include: 1) different volume of data; 2) different software versions (e.g., SAS 9.2 vs 9.3); and 3) Different operating Systems/platforms (e.g., Windows vs UNIX).

2. Modular Programs

a. SAS Programs

Even though different modular programs answer different questions, similar data operations are executed. To speed up the audit of the programs to get final approval for production use, similar data operations should be handled using the exact same piece of code that gets recycled across multiple projects.

b. Documentation and Input Forms

Substantial training is required to understand the correct choice of modular program and specification of inputs. Since many requesters of modular program runs are occasional users, it has proven worthwhile for MSOC personnel to perform these tasks after one or more conversations with the requester. This process can be enhanced through additional training of requesters and the use of input forms.

3. Summary Tables and Distributed Query Tool Software

a. SAS Programs for Summary Tables

Since the SAS programs that create these tables use all of the data in each Data Partner's MSDD files, the structure of the programs as well as other inefficiencies can have a sizable impact on the run time at various Data Partners. During Year Two, the SAS programs were revised to make the structure simpler and more efficient. To the extent possible large, raw MSCDM tables were read only once and intermediate data were created to be recycled for different purposes. The format of the output generated (i.e., the summary tables themselves) was made consistent across tables using a unique format (e.g., tab-delimited text format) so that development of the Query Tool software is made easier/more straightforward.

b. Query Tool – DataMart Client Software

All but one Data Partner successfully installed and operated the DataMart Client to respond to queries sent via the Mini-Sentinel Distributed Query Tool. The most common barriers were related to initial setup of the software and guidance for installing updates. Most Data Partners required some form of local approval from their IT security staff to install the software, and many partners needed guidance to describe the security architecture of the system. The Mini-Sentinel Distributed Query Tool: System Description and Technical Documentation provides technical details of the system that proved useful for these technical discussions with the Data Partners. The Data Partner that has not yet received approval to install and use the software is responding to the summary table queries by executing the relevant computer code directly against the summary tables. The MSOC is actively working with this site to obtain the necessary approvals.

The lessons learned described above will be addressed during Year 3 of the project.

VI. OTHER DATA CORE ACTIVITIES

A. OPERATIONS CENTER COMMUNICATIONS

MSOC holds a weekly teleconference to maintain regular contact with and between the Data Partners. In addition to regularly scheduled meetings, MSOC is available by email, phone, and teleconference to deal with concerns and questions as they arise.

During Year Two, the MSOC also expanded its work with various workgroups. MSOC helps ensure that workgroups utilize MSDD effectively, efficiently, and properly. MSOC members are available to the workgroups during regular meetings or by email and phone as needed. In particular, MSOC reviews all workgroup plans to ensure that sensitive information is appropriately protected. MSOC also maintains a secure system used to communicate sensitive information with Mini-Sentinel Collaborators. This system has been designed to be compatible with all Mini-Sentinel Collaborators to continually facilitate data exchange.

B. STANDARD OPERATING PROCEDURE DEVELOPMENT

Given Mini-Sentinel's large number of participating organizations, its interdependency across organizations and between core work streams, and its scale of surveillance activity, formal Standard Operating Procedures (SOPs) are needed to support the program.

SOPs establish the basis for management control throughout Mini-Sentinel operations. They describe a process and provide instruction about how to perform the Procedure by detailing key steps, roles and responsibilities, and decision-making authorities. They also provide Mini-Sentinel with a basis for evaluation and improvement of processes.

Topics and processes needing SOPs were identified through workgroup discussions and reviews. The SOPs in development are listed in **Table 3**.

Table 3. Standard Operating Procedures Descriptions and Status

#	SOP	Description
1	**Standard Operating Procedure (SOP) Management**	Procedure for managing SOPs including ownership, review and approval, training, communications, and maintenance
2	**Query Request and Fulfillment**	Procedures for fulfilling queries requested from the Mini-Sentinel Operations Center, Food and Drug Administration, and Mini-Sentinel active surveillance investigators
3	**Data Update**	Procedures for managing the distribution of the Common Data Model (CDM), data update schedules, and loading data into the CDM
4	**Data Quality Checking and Profiling**	Procedures that define data quality measurement and reporting
5	**SAS Program Development**	Procedures supporting SAS code development, testing, and delivery
6	**Common Data Model Change Management**	Procedures governing revisions to the Mini-Sentinel Common Data Model

C. DATA STABILITY ASSESSMENT

Data Partners' source data is a dynamic resource that becomes more complete and accurate over the several months that elapse after care is delivered. This assessment evaluates the time by which the data is relatively complete and stable.

1. Process

With consultation with the FDA and Data Partners, the MSOC developed a specification of the data to be collected and displayed as a result of this analysis. Please see **Appendix B** for the specification and a sample page in the report. Subsequent to agreement on the specification, computer programs were developed to implement the analysis.

Programs were first developed to take a "snapshot" of the data from an individual ETL, from the most recent month, back to approximately 18 months. Usually, this was the ETL1 (i.e., the first ETL) that Data Partners developed for Mini-Sentinel. After a Data Partner created their second ETL (i.e., ETL2), they were in a position to take another snapshot, this time from their ETL2. With both set of snapshot files available, we were able compare their first ETL to the second ETL. By comparing a number of measures between the ETL1 and ETL2, we are able to assess how up to date the data was in their ETL1 and then extrapolate this assessment to future ETLs.

2. Status

Snapshot programs were developed and distributed during the interval of August 2010 through October 2011 for those Data Partners that developed their ETL1. After Data Partners developed their ETL2, they were instructed to take a snapshot of their ETL2 data. This work will continue in Year 3.

The MSOC has received data from nine Data Partners and results will be reviewed in Year Three.

D. DISSEMINATION ACTIVITY

1. Manuscripts

During Year Two, members of the Data Core produced three manuscripts detailing the goals of the Mini-Sentinel pilot project and the development and capabilities of the Mini-Sentinel Distributed Database. Two of these manuscripts, "Design considerations, architecture, and use of the Mini-Sentinel distributed data system"[10] and "The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction",[11] were published in the January 2012 *Pharmacoepidemiology & Drug Safety* supplement, and the other article, "Developing the Sentinel System – A National Resource for Evidence Development,"[12] was published in the *New England Journal of Medicine* in February 2011. Brief descriptions of these manuscripts can be found on the Mini-Sentinel website.

2. Meetings and Presentations

Table 4 includes information regarding meetings attended and presentation by MSOC Data Core members during Year Two.

Table 4. Mini-Sentinel Data Core Meetings and Presentations (Year Two)

Date	MSOC Data Core Staff	Venue	Description
10/20/10	Kim Lane, Jeff Brown, Richard Platt	AHIP's Medical Leadership Forum	Poster presentation: "FDA's Mini-Sentinel Program: A Distributed Data Network to Assess Safety of Marketed Medical Products"
11/12/10	Jeff Brown, Nicolas Beaulieu	Planning Board presentation	"Characteristics of the Mini-Sentinel Distributed Database": presented an overview of the MSDD, explained the CDM and Mini-Sentinel's data checking/characterization procedures, and discussed plans for work in Year Two. Also provided a comprehensive look at the data available and the population that comprised the MSDD at that time.
12/13/10	Jeff Brown	ONC presentation	"Distributed Research Network Technologies for Population Medicine": ONC invited Dr. Brown to present on the infrastructure and software development in the Mini-Sentinel network. He provided an overview of the Mini-Sentinel software development to date and a demonstration of the system and then discussed possible future work, needs, and opportunities.
1/13/11	Lesley Curtis, Jeff Brown, Mark Weiner	Annual Meeting at the Brookings Institute	Data Core leads attended the annual FDA meeting with Mini-Sentinel staff and collaborators to discuss completed and ongoing projects. Lesley Curtis and Mark Weiner also attended a Mini-Sentinel Planning Board meeting. Lesley Curtis gave a presentation to FDA staff and Mini-Sentinel collaborators summarizing Sentinel's accomplishments during the program's first year and a presentation on the Mini-Sentinel Distributed Database.
2/16/11	Jeff Brown	Brookings Institute Expert Workshop for the Mini-Sentinel Project	This workshop focused on topics that were specifically designed to inform about Mini-Sentinel activities directly relevant to the Mini-Sentinel Coordinating Center. The topics included appropriate analysis and interpretation of signals in the context of large sample sizes, interpreting multiple results to the same query, using distributed regression methods for signal refinement, and establishing operating characteristics of signal refinement methods.
2/24/11	Mark Weiner	OMOP Training	This training was designed for FDA and Mini-Sentinel investigators and focused on the OMOP data stimulation program, OSIM2. The program's test data environment was developed to perform statistical evaluations of the analytical methods offered to

Date	MSOC Data Core Staff	Venue	Description
			identify drug-outcome associations. The training benefited the MSOC because the OMOP simulation or test environment also conforms to the Mini-Sentinel Common Data Model and the training informed later discussions within Mini-Sentinel.
4/1/11	Lesley Curtis, Mark Weiner	ISPE Rapid Medical Product Safety WebEx Symposia	“FDA Mini-Sentinel Data Infrastructure and Use”: presented an overview of the Mini-Sentinel project and its objectives, explained the roles of the Data Partners and coordinating center, and discussed how Mini-Sentinel uses the common data model to perform different types of distributed queries. Also provided information about Data Partner response times to queries and characterized the population in the MSDD at that time.
5/23/11	Lesley Curtis, Mark Weiner	Data Core Meeting at FDA	The objective of the May 23-24, 2011 meetings was for the Data Core Leaders to meet with FDA staff during an "all hands" meeting and with the CBER and CDER staff to describe and discuss the inclusion of clinical data elements to the Mini-Sentinel Common Data Model. During the meeting, the Data Core Leaders discussed with the FDA the agency’s interests and priorities regarding the inclusion of clinical data in the Common Data Model. Lesley Curtis and Mark Wiener also presented "Building an Infrastructure for Safety Surveillance: Expanding the MS Common Data Model" to the FDA staff and investigators involved in Mini-Sentinel.
5/23/11	Jeff Brown	ISPOR Conference	"FDA’s Mini-Sentinel Program: Overview of Primary Data Resources and Distributed Data Approach": presented an update on Mini-Sentinel initiatives, future plans, and resources and tools available to health services researchers. Also participated in a panel for the conference that included Mini-Sentinel and FDA investigators.
6/3/11	Lesley Curtis, Jeff Brown, Mark Weiner	Brookings Institute Surveillance Updates	Attended Brookings Institute meeting with participants that included Mini-Sentinel collaborators and OMOP investigators. Participants discussed analytic methods development priorities for active surveillance activities. Other topics for discussion included lessons learned, current activities, anticipated needs for methods research and development, and developing an agenda to advance methods research and development.

Date	MSOC Data Core Staff	Venue	Description
6/13/11	Jeff Brown	Academy Health Annual meeting	Presented a talk titled: "Use of Administrative Claims Data for CER: FDA's Mini-Sentinel Program—A Distributed Data Network to Assess Safety of Marketed Medical Products" as part of a panel discussion of distributed networks.
6/14/11	Jeff Brown	Academy Health Annual meeting	Presented a talk titled: "Developing Better Evidence on Medical Product Safety" as part of a panel discussion of FDA Mini-Sentinel project.
8/17/11	Lesley Curtis, Mark Weiner	ISPE Conference	Presented an overview of the Mini-Sentinel Distributed Database, detailed development of the Mini-Sentinel Common Data Model and the Mini-Sentinel Query Tool, and discussed future work. Also provided characterization of the MSDD and explained Mini-Sentinel's capabilities and distributed query approach. Discussed plans for incorporating the Year Two clinical laboratory data and vitals into the CDM.
8/30/11	Jeff Brown Rich Platt	ONC Summer Concert Series on Distributed Population Queries (webinar)	Presented an overview of the Mini-Sentinel project as part of the ONC-sponsored Standards and Interoperability Framework summer concert series. Presentation includes details of the distributed querying approach and several examples.
9/15/11	Mark Weiner	FDA Webinar	"Content and Capabilities of the Mini-Sentinel Clinical Additions (laboratory results and vital signs)": presented a webinar to FDA investigators and staff regarding the work of the Clinical Additions workgroup. Detailed plans for incorporating clinical data as part of the Year Two MSCDM expansion.
9/26/11	Jeff Brown	Engelberg Center for Health Care Reform, Brookings Institute	Presentation titled: "FDA Sentinel Initiative Strategic Review: Mini-Sentinel Querying Capabilities and Lessons Learned from Recent Assessments."

E. MODULAR PROGRAMS

A total of 58 cycles of modular programs were executed to fulfill 16 data requests by FDA. FDA's Center for Drug Evaluation and Research (CDER) was responsible for all the requests. MP1 was used in five requests, MP3 in ten requests, and MP4 in one request (**Table 5**). MP2 was not used in Year Two. Each of these 14 requests involved between one and sixteen modular program executions for a total of 58 executions. The requests had varying levels of complexity, ranging from a straightforward MP1 request with one run to a rather complex request consisting of a combination of MP3 and MP4 with pre-existing conditions and incidence input files. Multiple executions are required when querying different scenarios, such as with or without pre-existing conditions input files or with or without incident input files, or when using multiple values of other parameters, such as washout period or episode gap. For example, the prasugrel request consisted of two executions of MP1 and four executions of MP3 (total of six executions) to assess overall use and use among those with certain pre-existing conditions. It used one drug and two pre-existing condition input files. In another example, the smoking cessation request required two separate runs of MP3 with six executions each (total of 12 executions). Both runs used several drug and outcome input files. With each additional run, more output files are created and must be audited and aggregated. In addition, the reports of more complex requests become larger and more challenging to produce and interpret.

Table 5. Number of Modular Programs Requested in Year 2 (September 23, 2010, to September 22, 2011)

Modular Program	Number of Requests	Number of Executions
MP1	5	22
MP3	10	35
MP4	1	1
TOTAL	**16**	**58**

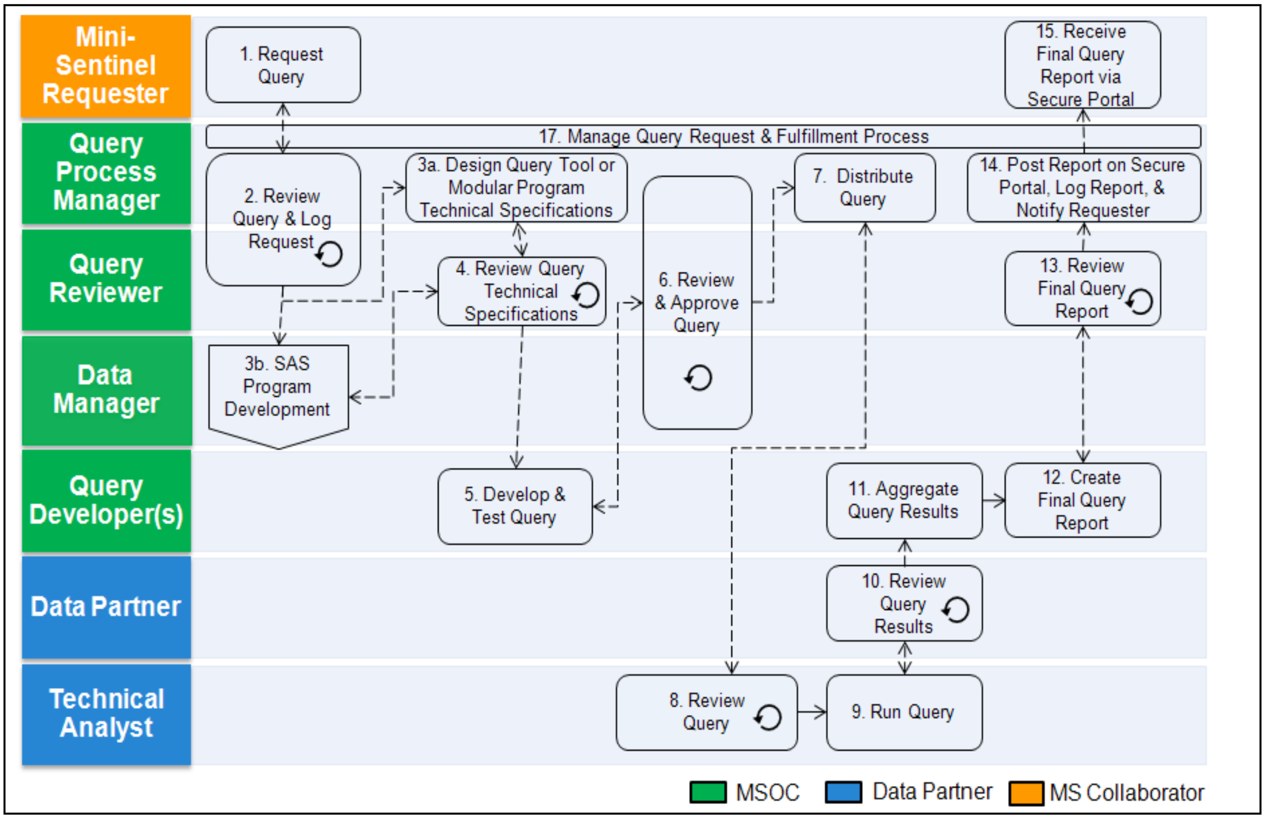
Data Partners have five business days to complete every request. However, MSOC occasionally distributed multiple requests at the same time, but staggered the due dates so as not to overwhelm Data Partners. This may account for the longer response times in some requests. The average request completion time by Data Partners during Year Two was seven days. Overall, response time by Data Partners was better than expected.

All reports were created in Microsoft Excel® and included both tables and figures along with an overview sheet describing the request specifications and contents. Most reports presented the number of users, dispensings, total days supplied, dispensings per user, days supplied per user, days supplied per dispensing, and events (for MP3) for either prevalent users, incident users, or both. Additionally, the reports showed percent contribution of each Data Partner to the total number of users, dispensing, days supplied, and events (for MP3).

The average time from data completion to report submission was 8.8 business days and the median time was 5 days. Some reports took longer due to investigation and revision of errors or unexpected data in the output at one or more of the 17 Data Partners, as well as additional consultation with FDA regarding report content and formatting.

Figure 5 shows the query request and fulfillment process for both Summary Table and Modular Program Requests.

Figure 5. Query Request and Fulfillment Process


Mini-Sentinel Requester
Query Process Manager
Query Reviewer
Data Manager
Query Developer(s)
Data Partner
Technical Analyst
1. Request Query
15. Receive Final Query Report via Secure Portal
17. Manage Query Request & Fulfillment Process
2. Review Query & Log Request
3a. Design Query Tool or Modular Program Technical Specifications
7. Distribute Query
14. Post Report on Secure Portal, Log Report, & Notify Requester
4. Review Query Technical Specifications
6. Review & Approve Query
13. Review Final Query Report
3b. SAS Program Development
5. Develop & Test Query
11. Aggregate Query Results
12. Create Final Query Report
10. Review Query Results
8. Review Query
9. Run Query
MSOC
Data Partner
MS Collaborator

F. SUMMARY TABLES AND QUERY TOOL

A total of 97 summary table queries were performed to respond to 28 requests during Year Two (**Table 6**). FDA's Center for Drug Evaluation and Research (CDER) was responsible for the majority of requests with 16, while the Center for Biologics Evaluation and Research (CBER) submitted two, the Center for Devices and Radiological Health (CDRH) submitted one, and an FDA Sentinel Core Team submitted one. The MSOC initiated eight requests for the purposes of: investigating counts as background information for a Modular Program request (4 requests); investigating counts as background information for task order activities (2); investigating counts as background information for PRISM analyses (1); and obtaining updated enrollment numbers for each Data Partner (1).

Data Partners were given 2 business days to complete each query. The average raw Data Partner response time during Year Two was 5 days. Response time was higher than expected, especially for the first few sets of queries, as Data Partners were adjusting to the Query Tool and working with the MSOC to fix any bugs that arose. Using the 90th percentile of the completion date across all Data Partners to account for outliers, the average Data Partner response time was 2 days.

Twenty of the 28 requests involved sets of summary reports that were created by the MSOC and submitted to the requester (**Table 6**). All reports were created in Microsoft Excel and included both pivot tables and figures along with an overview sheet describing the tables and figures presented in the report. Most requests involved more than one Excel file report because reports were grouped by type of query. For example, if a request involved three generic name queries and two HCPCS queries, two reports would be created—one for the generic name queries and one for the HCPCS queries. For generic name queries and drug class queries, reports displayed counts of users, prevalence rates (users per 1,000 enrollees), days per user, dispensings per user, and days per dispensing. For diagnosis and procedure queries, reports displayed counts of patients, prevalence rates (patients per 1,000 enrollees), and the number of events per patient.

Table 6. Number of Summary Table Query Requests Completed September 23, 2010, to September 22, 2011, by Requester

Center/Requester	Number of Requests (Broad Categories)	Number of Queries	Number of Requests Involving Reports to FDA
MSOC	8	35	2
CBER	2	8	2
CDER	16	49	14
CDRH	1	3	1
FDA[a] (Not Center-Specific)	1	2	1
TOTAL	**28**	**97**	**20**

[a] This request was a test query submitted by the FDA leadership team on acute myocardial infarction.

Table 7 displays the number of queries completed during Year Two broken out by requester and query type. Most queries were generic name queries (31), four-digit ICD-9-CM diagnosis code queries (28), or HCPCS queries (20). In addition, there were 7 five-digit ICD-9-CM diagnosis queries, 4 four-digit procedure code queries, 3 enrollment queries, 3 three-digit ICD-9-CM diagnosis queries, and 1 drug class query. There were no requests for three-digit procedure code queries.

Table 7. Number of Summary Table Queries Completed September 23, 2010, to September 22, 2011, by Requester and Query Type

Requester	Enrollment	Generic Name	Drug Class	3-Digit Diagnosis Code	4-Digit Diagnosis Code	5-Digit Diagnosis Code	3-Digit Procedure Code	4-Digit Procedure Code	HCPCS	TOTAL
MSOC	3	7	---	---	13	3	---	3	6	35
CBER	---	---	---	---	2	2	---	---	4	8
CDER	---	24	1	1	12	2	---	---	9	49
CDRH	---	---	---	---	1	---	---	1	1	3
FDA [a] (Not Center-Specific)	---	---	---	2	---	---	---	---	---	2
TOTAL	**3**	**31**	**1**	**3**	**28**	**7**	**---**	**4**	**20**	**97**

[a] This request was a test query submitted by the FDA Sentinel Core Team on acute myocardial infarction.

G. AD HOC REQUESTS

One request requiring ad hoc programming was made during Year Two. The purpose of this request was to characterize the Mini-Sentinel population aged 65 and over in terms of diagnoses, procedures, and medications (65+ Characterization). Results were stratified by age and gender and will be ultimately compared to the CMS population.

Two versions of the 65+ Characterization requests were distributed to Data Partners. The first version was requested on February 2, 2011 and distributed to Data Partners on March 10, 2011. This request calculated counts and rates of healthcare and drug utilization. For each calendar year (2007, 2008, and 2009), members aged 65 and over on July 1st who are continuously enrolled in a health plan (for both medical and drug benefits) for the entire year were selected. The number of enrolled patients served as the denominator for computing rates and this denominator was also stratified by sex and one of four age groups (65-69, 70-74, 75-79, and 80+). Healthcare (i.e., diagnosis and procedure) and drug utilization during the given calendar year were tabulated and stratified by sex and age groups.

The March 2010 version of the 65+ Characterization report was submitted to FDA on March 31, 2011.

After reviewing the first report, FDA revised the original specifications, program code was revised, and a new request was distributed to data partners. The following modifications were made to the code based on feedback received from FDA:

- From 4 age groups to 2 (65-74 and 75+).
- Added one calendar year (2010).
- Summarization of drug use is now done at the generic name and drug category level (as opposed to NDC level before). We provided a customized look-up table (ndc_lookup_table.sas7bdat) for each Data Partner, based on the list of NDC codes generated by the Data Partners' Quality Assessment (QA) output shared with the Mini-Sentinel Operations Center team. This list is then linked to the First Data Bank (FDB) master list.
- For each code/data type, an extra output dataset is now generated showing distribution of number of patients.

The revised package was beta tested in September 2011 and the final package distributed to Data Partners on October 18, 2011. The report was submitted to the FDA on February 8, 2012.

H. LESSONS LEARNED

As FDA and MSOC began using MS analytic tools for rapid querying of data, there became a need to log the number of requests. MSOC created the Query Tracker Spreadsheet to keep track of all current, pending and completed requests. This spreadsheet was also provided to FDA as a daily update of ongoing requests.

1. Modular Program Requests

For the first several modular program requests, many Data Partners experienced errors and warnings in their SAS logs. Most of these errors or warnings resulted from site-specific issues with CDM compliance or operating environments. Data Partners were notified of any CDM compliance issues and modular program code was revised to account for differences in operating environments.

Some Data Partners were wary of sharing the Modular Program log files with MSOC due to potential transfer of proprietary or personal health information. Modular Program code was revised to delete this information and some Data Partners manually redacted portions of the log files before submitting to MSOC.

2. Summary Table and Query Tool

When the MSOC first started creating reports summarizing summary table results, reports were created in the most current version of Microsoft Excel (2007). However, most of FDA's team was working in a slightly older version of Microsoft Excel (1997-2003). These individuals were getting errors when trying to view the pivot tables in the reports. The MSOC now knows how to create and keep each report in the correct version of Microsoft Excel (1997-2003) so that they can be viewed by all users.

With each request, the MSOC improved the reports summarizing results both in terms of the information contained in the tables and figures that are displayed and in terms of formatting. While there will always be further improvements made, the MSOC believes the content displayed in the reports is useful to the requester and is displayed in such a way that is easy to understand.

Query results from Data Partners do not display rows of data when a specific stratum has no events. For example, if there are no females ages 5-9 with exposure to the queried drug product in 2009, and data from that strata were requested, the system will not extract any information on this strata—i.e., there will be no row of data displaying 0 events and the total enrollment in that strata. Thus, when data are aggregated across Data Partners to calculate prevalence rates (the number of users/patients per 1,000 enrollees), the denominator will be underestimated due to these missing rows, and the prevalence rate will be overestimated. In Year Three, the MSOC will be updating the Query Tool to include the missing rows so that prevalence rates will be calculated correctly. However, for most reports completed during Year Two, prevalence rates were calculated incorrectly. For a couple of reports (identified by FDA), the MSOC went back and queried Data Partners' enrollment summary tables to obtain enrollment for missing strata and calculate correct prevalence rates.

3. Ad Hoc Requests

For the first several modular program requests, many Data Partners experienced errors and warnings in their SAS logs. Most of these errors or warnings resulted from site-specific issues with MSCDM compliance or operating environments. Data Partners were notified of any MSCDM compliance issues and modular program code was revised to account for differences in operating environments.

Some Data Partners were wary of sharing the Modular Program log files with MSOC due to potential transfer of proprietary or personal health information. Modular Program code was revised to delete this information and Data Partners manually redact portions of the log files before submitting to MSOC.

VII. AUTHORSHIP

This report was prepared by members of the Mini-Sentinel Data Core and Data Partners.

Data Core Leads

Lesley H. Curtis, PhD, Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, USA

Mark G. Weiner, MD, Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA.

MSOC Members

Elizabeth Balaconis, BA
Nicolas U. Beaulieu, MA
Jeffrey S. Brown, Ph
Jillian Lauer, BS
James Marshall, MPH
Megan Mazza, MPH
Richard Platt, MD, MSc
Robert Rosofsky, MA
Lisa Trebino, MPH
Tiffany S. Woodworth, MPH

Data Partners

Aetna
HealthCore, Inc.
HMO Research Network
Humana
Kaiser Permanente Center for Effectiveness and Safety Research
Vanderbilt University

VIII. REFERENCES

1. 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.

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

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. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr*. 2005:12-25.

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

10. Curtis, L, Weiner, MG, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiology & Drug Safety*. 2012; 21(S1): 23-31.

11. Platt, R, Carnahan, RM, Brown, JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiology & Drug Safety*. 2012; 21(S1): 1-8.

12. Behrman, RE, Benner, JS, Brown, JS, et al. Developing the Sentinel System — A National Resource for Evidence Development. *N Engl J Med*. 2011; 364: 498-499.

IX. APPENDIX A: GLOSSARY

- ALP: alkaline phosphatase
- ANC: absolute neutrophil count
- CBER: Center for Biologics Evaluation and Research
- CDER: Center for Drug Evaluation and Research
- CDM: Common Data Model
- CDRH: Center for Devices and Radiological Health
- CPT-4: Current Procedural Terminology-4
- DBP: Diastolic blood pressure
- EHR: Electronic Health Record
- ESP: Electronic Support for Public Health
- ETL: Extract:Transform:Load
- FDA: Food and Drug Administration
- FDB: First Data Bank
- FISMA: Federal Information Security Management Act of 2002
- HbA1C: glycosylated hemoglobin
- HCPCS: Healthcare Common Procedure Coding System
- HHS: Department of Health and Human Services
- HIPAA: Health Insurance Portability and Accountability Act
- HOI: Health Outcomes of Interest
- INR: International Normalized Ratio
- LOINC: Logical Observation Identifiers Names and Codes
- MP: Modular Program
- MPQI: Modular Program Query Interface
- MSCDM:Mini-Sentinel Common Data Model
- MSDC: Mini-Sentinel Data Catalog
- MSDD: Mini-Sentinel Distributed Database
- MSOC: Mini-Sentinel Operations Center
- NDC: National Drug Codes
- ONC: Office of National Coordinator for Health Information Technology
- PRISM: Post-Licensure Rapid Immunization Safety Monitoring
- QA: Data Partners' Quality Assessment
- S&I: ONC Standards & Interoperability
- SBP: Systolic blood pressure
- SGPT: Serum Glutamic Pyruvic Transaminase or alanine aminotransferase
- SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
- SOP: Standard Operating Procedure
- VDW: Virtual Data Warehouse

X. APPENDIX B: DATA STABILITY

A. TECHNICAL SPECIFICATIONS OF DATA STABILITY ANALYSIS

1. Row Counts Stability: For each table, compare total row counts between ETL1 and ETL2, by month, using a defined set of date variables:

- **1.1.** Enrollment table: *Enr_Start*
- **1.2.** Demographic table: *Birth_Date*
- **1.3.** Dispensing table: *RxDate*
- **1.4.** Encounter table: *ADate*
- **1.5.** Diagnosis table: *ADate*
- **1.6.** Procedure table: *ADate*
- **1.7.** Death table: *DeathDt*
- **1.8.** COD table: Link to *DeathDt* in Death table

For each of the above, the process will be as follows:

1. By calendar month as defined for each table using the variable specified above, count the number of rows in the snapshot. Call this count SnapCount.
2. By calendar month as defined for each table using the variable specified above, count the number of rows in the most current ETL. Call this count ETLCount.
3. For each calendar month, compute a proportion as SnapCount divided by ETLCount.
4. Graph the proportion for all calendar months calculated.

2. Change in Records Stability – Type A: Match the records from one snapshot to another, as another level of measuring stability. Within each table, all variables will be used as match keys.

- **2.1.** Enrollment table
- **2.2.** Demographic table
- **2.3.** Dispensing table
- **2.4.** Encounter table
- **2.5.** Diagnosis table
- **2.6.** Procedure table
- **2.7.** Death table
- **2.8.** COD table

For each of the above, the process will be as follows:

By calendar month as defined for each table using the variable specified above in section 1, eliminate all total duplicates from the most current ETL, checking values for all variables.

Count the number of resulting rows in the most current ETL. Call this count MatchETLCount.

Using the snapshot tables, eliminate all total duplicates, checking values for all variables.

Using all variables within a table as match keys, attempt a match from each row in the most recent ETL of section 0 to the snapshot row from the prior ETL in section 0. Count the number of matching rows and call this count MatchSnapCount.

For each calendar month, compute a proportion as MatchSnapCount divided by MatchETLCount.

Graph the proportion for all calendar months calculated.

3. <u>Change in Records Stability – Type B</u>: This is the inverse of section 3 above. Within each table, all variables will be used as match keys.

- **3.1.** Enrollment table
- **3.2.** Demographic table
- **3.3.** Dispensing table
- **3.4.** Encounter table
- **3.5.** Diagnosis table
- **3.6.** Procedure table
- **3.7.** Death table
- **3.8.** COD table

For each of the above, the process will be as follows:

By calendar month as defined for each table using the variable specified in section 1 above, eliminate all total duplicates from the most current ETL, checking values for all variables.

5. Using the snapshot tables, eliminate all total duplicates, checking values for all variables.
6. Count the number of resulting rows in the snapshot. Call this count MatchSnapCount.
7. Using all variables within a table as match keys, attempt a match from the snapshot rows of section 5, to the most recent ETL rows from section 6. Count the number of matching rows and call this count MatchETLCount.
8. For each calendar month, compute a proportion as MatchETLCount divided by MatchSnapCount.
9. Graph the proportion for all calendar months calculated.

4. For <u>specific</u> tables, compare counts between ETL1 and ETL2 using specific variables and values as follows:

- **4.1.** <u>Enrollment</u> table: For each calendar month, compare the following:
  - 4.1.1. Count of patients with *MedCov*="Y"
  - 4.1.2. Count of patients with *DrugCov*="Y"
  - 4.1.3. Count of patients with *MedCov*="Y" and *DrugCov*="Y"
- **4.2.** <u>Dispensings</u> table: For each calendar month, compare the following:
  - 4.2.1. Count of dispensing rows for specified drug classes, based on *NDC:* ARBs, Behavioral, and Fertility
  - 4.2.2. Count of dispensing rows for specific ranges of *RxSup* (1-30, 31-60, 61-90, 90+)
- **4.3.** <u>Encounter</u> table: For each calendar month, compare the following:
  - 4.3.1. Count of encounter rows for each *EncType*
  - 4.3.2. Count of encounter rows for each *Admitting_Source*
  - 4.3.3. Count of encounter rows, for IP, for LOS:
- **4.4.** <u>Diagnosis</u> table: For each calendar month, compare the following:
  - 4.4.1. Count of diagnose rows for each *EncType*
  - 4.4.2. Count of diagnose rows for each *Dx_CodeType*
  - 4.4.3. Count of diagnose rows for each health outcomes of interest : AMI and stroke
- **4.5.** <u>Procedure</u> table: For each calendar month, compare the following:
  - 4.5.1. Count of procedure rows for each *EncType*
  - 4.5.2. Count of procedure rows for each *Px_CodeType*
- **4.6.** <u>Death</u> table: For each calendar month, compare the following:
  - 4.6.1. Count of death rows for each *Source*

4.6.2. Count of procedure rows for each *Confidence*

For each of the above, the process will be as follows:

By calendar month as defined for each table using the variable specified in section 1 above, for each table using the variables specified in section 4 above, count the number of rows in the snapshot for each level of the specified variables. Call this count SnapCount.

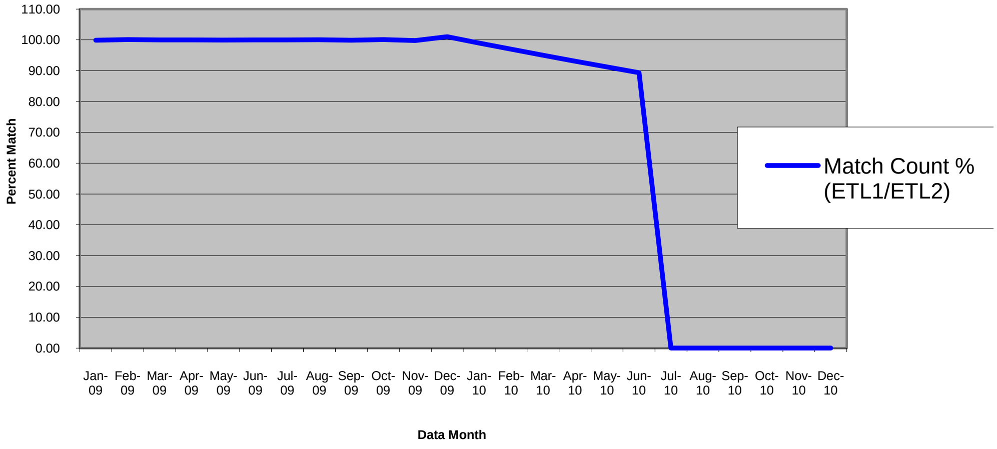
10. By calendar month as defined for each table using the variable specified in section 1 above, for each table using the variables specified in section 4 above, count the number of rows in the most current ETL for each level of the specified variables. Call this count ETLCount.
11. For each calendar month, compute a proportion as SnapCount divided by ETLCount.
12. Graph the proportion for all calendar months calculated.

B. SAMPLE PAGE OF DATA STABILITY REPORT

Calendar Month	RowCount Snapshot (ETL1)	RowCount Refresh (ETL2)	MatchCount% (ETL1/ETL2)
2009-01	100,043	100,142	99.90
2009-02	100,142	100,043	100.10
2009-03	100,043	100,024	100.02
2009-04	100,024	100,011	100.01
2009-05	100,011	100,079	99.93
2009-06	100,079	100,083	100.00
2009-07	100,083	100,072	100.01
2009-08	100,072	100,036	100.04
2009-09	100,036	100,119	99.92
2009-10	100,119	100,006	100.11
2009-11	100,006	100,202	99.80
2009-12	100,202	99,200	101.01
2010-01	97,196	98,208	98.97
2010-02	94,280	97,226	96.97
2010-03	91,452	96,254	95.01
2010-04	88,708	95,291	93.09

Calendar Month	RowCount Snapshot (ETL1)	RowCount Refresh (ETL2)	MatchCount% (ETL1/ETL2)
2010-05	86,047	94,338	91.21
2010-06	83,465	93,395	89.37
2010-07	-	92,461	0.00
2010-08	-	91,536	0.00
2010-09	-	90,621	0.00
2010-10	-	89,715	0.00
2010-11	-	88,818	0.00
2010-12	-	87,929	0.00

Data Stability
Data Partner: xxx, Table: yyy
Table Row Counts


Percent Match
Jan-09
Feb-09
Mar-09
Apr-09
May-09
Jun-09
Jul-09
Aug-09
Sep-09
Oct-09
Nov-09
Dec-09
Jan-10
Feb-10
Mar-10
Apr-10
May-10
Jun-10
Jul-10
Aug-10
Sep-10
Oct-10
Nov-10
Dec-10
Data Month
Match Count % (ETL1/ETL2)