MINI-SENTINEL COORDINATING CENTER

REPORT OF SCIENTIFIC OPERATIONS CENTER
DATA GROUP ACTIVITIES

YEAR FIVE
September 2013 – September 2014

Prepared by: Mini-Sentinel Scientific Operations Center, Mini-Sentinel Data Core, Mini-Sentinel Data Group, and Collaborating Institutions

October 2015

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

Mini-Sentinel Coordinating Center

Report of Scientific Operations Center Data Group Activities

TABLE OF CONTENTS

I. INTRODUCTION

A. OVERVIEW OF THE MINI-SENTINEL PROJECT

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 project currently focuses on three major activities:

- Assessments - Medical product exposures, health outcomes, and associations between them
- Methods - Techniques for identifying, validating, and linking medical product exposures and health outcomes
- Data Infrastructure - Mini-Sentinel Distributed Database (MSDD) and infrastructure (e.g., systems, tools, applications) used to access and use the data

Collaborating Institutions provide secure data environments, infrastructure, staff, and other resources to support Mini-Sentinel activities. 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 Operations 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 structure of the Mini-Sentinel Operations Center (MSOC) is described in the statement of Mini-Sentinel Principles and Policies and the Mini-Sentinel Coordinating Center Organizational Chart.

This report focuses on the activities and responsibilities of the Scientific Operating Center's (SOC) Data Group. The SOC Data Group is organized into four groups: Data Model and Quality Assurance, Infrastructure, Programming, and Query Fulfillment.

1. Standard Operating Procedures: Description, Revision, and Implementation

Most activities in which the Data Group is involved in are driven in accordance to a set of Standard Operating Procedures (SOPs). These SOPs establish clear and consistent processes to be used for all common activities, and ensure transparency, traceability, consistency, and quality control in various types of work products.

All SOPs used by the Data Group are posted to the Mini-Sentinel public website. Currently available SOPs include:

- Data Quality Review and Characterization: how data quality checks are determined, measured, reported, and monitored by MSOC

- Data Update: how data are loaded into the Mini-Sentinel Common Data Model (MSCDM) format by the Data Partners (DPs) and data update schedules are managed at MSOC
- SAS Program Development: how SAS programs are developed, tested, reviewed, and approved before being released for production use by MSOC.

As Data Group workflows get more complex and require more thorough quality controls, SOPs will be refined and new SOPs will be created and published. During Year Six, the Data Group will work closely with FDA and MS collaborators to enhance the set of SOPs being used in all workflows.

2. Responsibilities of the Data Model and Quality Assurance Group

The Data Model and Quality Assurance Group has primary responsibility for supporting Data Partners' development of their MSDD. This involves: developing, updating, and managing the MSCDM; managing the data refresh and approval process according to its SOPs for Data Update and Data Quality Review and Characterization; and providing standard and ad hoc data characterization reports to FDA, workgroups, and other stakeholders to help guide appropriate use of Mini-Sentinel data resources. The Data Model and Quality Assurance Group develops and implements data quality checking and characterization metrics and works with Data Partners to improve use of the MSDD.

3. Responsibilities of the Infrastructure Group

The **Infrastructure Group's** primary responsibility is to develop, manage, and maintain the analytic tools enabling rapid and efficient querying of the MSDD. The Infrastructure Group is also responsible for the development, maintenance, and enhancement of: 1) the technical infrastructure required to ensure appropriate and secure use of resources (e.g., private secure communications systems, implementation of the Mini-Sentinel Distributed Query Tool and activity tracking software, maintenance of Mini-Sentinel public website); and 2) a set of Standard Operating Procedures (SOP) that govern the work overseen by the Data Group staff.

4. Responsibilities of the Programming Group

The **Programming Group**'s overall responsibility is to manage and support all internal and external analytic programming activities. The Programming Group collaborates with the other groups in the Data Group, the MSOC Cores, FDA, and workgroups in order to ensure that: 1) programming needs are being consistently identified, specified, and met; 2) programming-related SOPs are being maintained, disseminated, and observed; 3) Mini-Sentinel programming guidelines and best-practices are being followed with regard to the development of programming code that is flexible, reusable, scalable, computationally efficient, and easily maintainable; and 4) guidance from the Data Model and Quality Assurance group regarding appropriate data usage and interpretation is available to workgroups engaged in custom programming activities. The Programming Group also develops and maintains SOPs on SAS Program Development and Quality Control of SAS Programs, as well as maintains and promotes the use of bug tracking software to document all quality control (QC) activities.

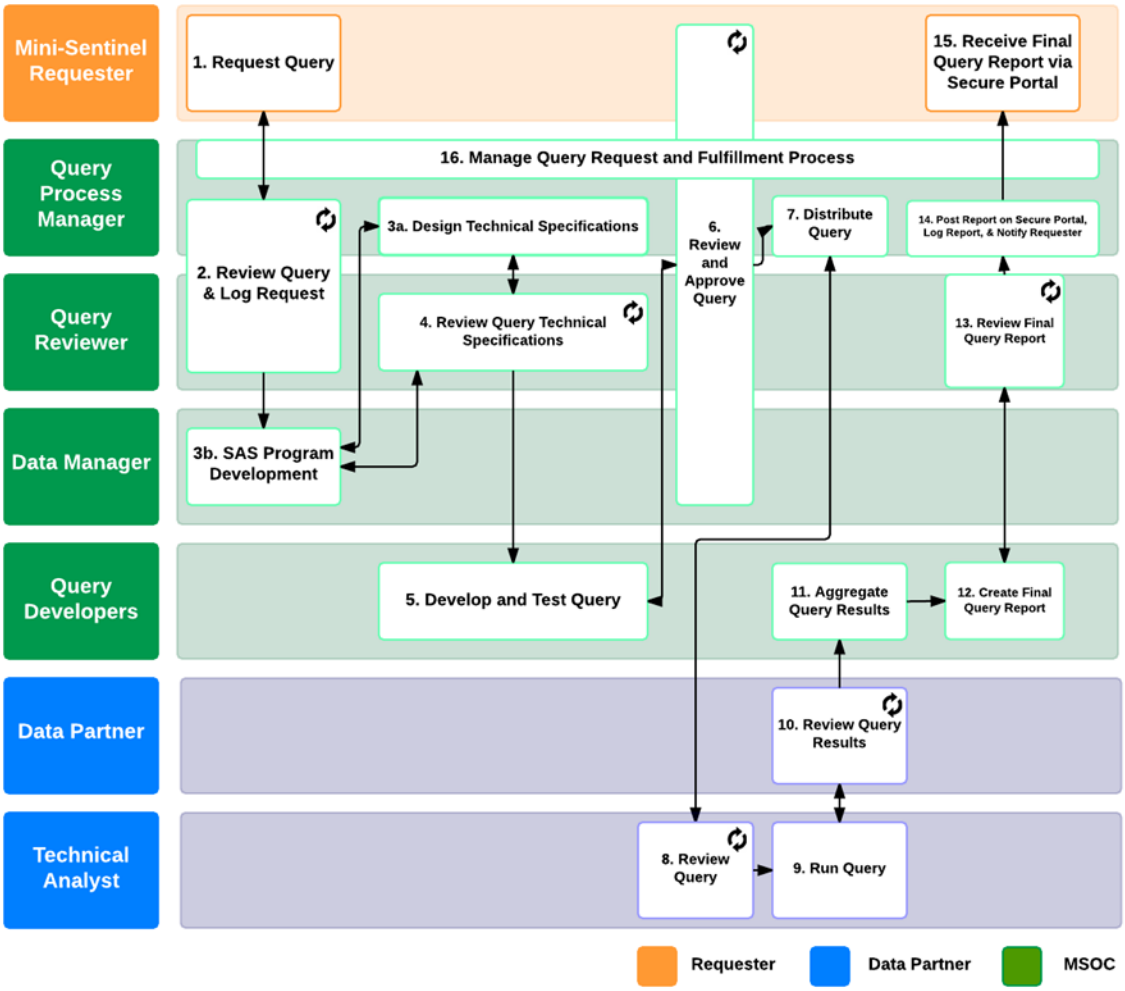
5. Responsibilities of the Query Fulfillment Group

The **Query Fulfillment Group** has primary responsibility for efficiently using the MSDD to answer data requests. Data requests can be initiated by FDA, Mini-Sentinel workgroups, or MSOC, and typically

involve the use of existing Mini-Sentinel querying tools such as modular programs and summary tables. **Figure 1** below shows the query request fulfillment process which includes: developing technical specifications; developing, testing, and distributing the query materials; aggregating the results and creating a report.

Together, the SOC's Data Model and Quality Assurance, Infrastructure, Programming, and Query Fulfillment Groups enable efficient and appropriate use of Mini-Sentinel data resources as well as provide technical and analytic support to various workgroup activities. The groups work closely on a daily basis to improve functioning of the Mini-Sentinel network, to develop new tools and streamline processes; several Scientific Operations Center analysts work across these four groups to ensure effective communication. SOC Data Group staff are members of the Mini-Sentinel Data Core and support and work closely with FDA, Data Partners, and Collaborating Institutions on all scientific Mini-Sentinel activities.

Figure 1. Mini-Sentinel Query Fulfillment Process



C. MINI-SENTINEL DATA CORE

1. Overview

The Data Core serves as an advisory committee that works with the SOC Data Group concerning development and implementation of the MSCDM, distributed data approach, and related data standards and quality measures. The Data Core recommends additional workgroups as appropriate and interacts regularly with the Methods and Protocol Cores. The Data Core is a key conduit for communication among Data and Academic Partners, project workgroups, and other parties interested in data-related aspects of Mini-Sentinel activities.

As directed by FDA or the MSOC, the Data Core Leader(s) assist the SOC with external communications, including presentation of Mini-Sentinel activities at scientific meetings and related venues. The Data Core can also recommend the formation of new workgroups.

2. Members of the Data Core

- Data Core Leaders
- Mini-Sentinel Scientific Operations Center Data Group staff
- Representatives from each Data Partner
- Representatives from FDA
- Additional analytical and technical staff as needed

3. Members' Terms and Selection

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

4. Data Partners

Of the 18 Mini-Sentinel Data Partners active during Year Five, those with health plan administrative and claims data in the MSCDM format include Aetna, HealthCore, Inc. (using Anthem data), the HMO Research Network (six sites), Humana, Kaiser Permanente Center for Effectiveness and Safety Research (KP CESR – six sites), Lovelace Clinic Foundation, OptumInsight, and Vanderbilt University (using Tennessee Medicaid data). Mini-Sentinel includes other Collaborating Institutions that have access to additional data sources of interest for medical product safety surveillance, including laboratory data, electronic health record (EHR) data, inpatient systems, and disease and device registries.

During Year Five, the SOC staff collaborated with various workgroups to onboard one additional regular Data Partner with administrative and claims data (Blue Cross Blue Shields – Massachusetts) and one with inpatient data – Hospital Corporation of America (HCA) Healthcare). Efforts to incorporate additional data areas and standard terminologies into the MSCDM/MSDD are ongoing and will continue to be the focus of activities in subsequent years.

D. DISTRIBUTED DATA APPROACH

Mini-Sentinel uses a distributed data approach in which Data Partners maintain physical and operational control over electronic data with their local, secure environments.[1-7] In this distributed data approach, each Data Partner extracts, transforms, and loads (ETL) their members' or enrollees' administrative, claims, and (in some cases) clinical data into the MSCDM format, employing identical names and formatting for each data element of the MSCDM. Data Partners execute standardized programs provided by the Scientific Operations Center groups or project workgroups and return the output of the programs to the MSOC for validation. Typically, the output is returned in summary, or aggregated, form. By allowing Data Partners to maintain control of their data and its uses, the distributed model avoids or reduces many of the data security, proprietary, legal, and privacy concerns of Data Partners, including those related to the Health Insurance Portability and Accountability Act (HIPAA).

This distributed approach also addresses the need to have local content experts maintain a close relationship with the data. For example, only a local expert can easily and effectively troubleshoot 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 one row per person with highly summarized or deidentified information such as age in an age range, number of prior hospitalizations, and total days exposed to a treatment. Although person-level data are occasionally required for some analyses, personally-identifiable protected health information are not transferred outside the individual Data Partners' environments.

II. OVERVIEW OF COMMON DATA MODEL

The Mini-Sentnel Common Data Model (MSCDM) version 4.0 was released in Year Five and is comprised of 11 data tables with person-level medical care and administrative data. This section describes the 11 data tables. Twelve summary tables, derived from these data tables are described in **Section V.B**.

Each of the 11 data tables 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 data and summary tables plus laboratory

reference guides are available in Overview and Description of the Common Data Model.[i] A common unique person identifier is included in each table to allow linkage across tables and a 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 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 true identifier is not shared outside the health plan with other Data Partners, the MSOC, the FDA, or anyone else.

Enrollment: The ability to ascertain who is enrolled in a health plan and eligible for medical and/or pharmacy benefits at any particular time is required for most Mini-Sentinel investigations. In many medical product safety evaluations, it is important to know the period 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 uses a start/stop structure and 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 to create continuous enrollment periods. 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, generates a new record even if the coverage is continuous. Disjointed periods of coverage—those that are separated by more than one day—are listed as separate records. Data Partners are not required to "bridge" gaps of more than one day in coverage; when appropriate, bridging is incorporated into analysis programs based on the specific needs of the evaluation.

Most Mini-Sentinel evaluations use the enrollment table to define periods during which we would expect to observe medical utilization in other tables (e.g., pharmacy dispensing). The table structure is a simplification of the HMO Research Network's Virtual Data Warehouse (VDW)[1,9] enrollment table structure.

In Year Five, the chart variable was added as a flag to indicate whether medical charts can be requested for the enrollee for use in validation studies. While the flag expedites identification of enrollees for which there is no contractual restriction on providing charts, it does not guarantee chart availability.

Demographic: The demographic table includes the unique person identifier, sex, birth date, race, ethnicity, zip code and zip code valid date. Only a subset of the Data Partners collects a meaningful percentage of race and ethnicity information. The demographic table includes demographic data on all individuals found in the Data Partners' database and is not restricted to members included in the enrollment table. That is, the Data Partner may provide demographic information on individuals who received care at an affiliate medical facility but is not enrolled in their health plan.

[i] MSCDM v4.0 is the version referenced in this report. The link will bring the reader to the version current at the time of reading. Information about prior versions will be available at the link.

In Year Five, two variables were added to the demographic table to allow assessment of socio-demographic factors as confounders—the 5-digit zip code of the individual's most recent primary residence and the earliest date at which the zip information is believed to be accurately captured. The zip code and zip code date are overwritten as the individual's information changes.

Dispensing: The dispensing table represents outpatient pharmacy dispensing captured by the Data Partners through pharmacy billing. Each record includes the unique person identifier, dispensed date, dispensed National Drug Code (NDC) in 11 digit format, and the days supplied and amount dispensed. Data Partners are instructed to process source transactions to remove rollback transactions and other adjustments before populating the dispensing table, although infrequently, such records may appear. This typically requires summation of dispensing information by unique person identifier, dispensing date, and dispensed NDC. No corrections are enforced at the data level for values that are out of range or implausible values (e.g., negative days, zero days, or 900 days supplied), leaving this cleaning for the analytic level.

Individual dispensings can be linked by analytic programs 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) are included if the pharmacy submits the claim to the Data Partner. Similarly, the purchase of over-the-counter medications is included if the transaction is submitted via the pharmacy to the Data Partner. At some Data Partners, infused medications, vaccinations, and other medications (e.g., injections) not dispensed through a pharmacy (e.g., provided directly by medical providers) are captured in the procedure table, as those administrations are considered "procedures" within the existing medical coding nomenclature and are captured in a separate data stream. A very small percentage (less than 0.1%) of outpatient dispensings represent NDCs for procedures. Medications dispensed in the inpatient setting are not currently available from the Data Partners and are not included in the Dispensing Table.

Encounter: Each record within the table represents a unique medical encounter defined as a unique combination of person identifier, admission/encounter date, provider, and care setting. Diagnoses and procedures recorded during encounters are recorded in the diagnosis and procedure tables. 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 and the diagnosis and procedure tables would contain all records of diagnoses and procedures. This table also includes discharge date of the hospitalization, provider code, facility code, three-digit provider zip code for the facility, Diagnosis Related Group (DRG) assigned to the admission and the DRG code version, the admitting source, the discharge status, and the discharge disposition.

Diagnosis: Most encounters are associated with at least one diagnosis; the exception is procedure-only encounters such as vaccinations. The diagnosis table is linked to the encounter table in a many-to-one 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 indicating whether the diagnosis was recorded in the primary discharge diagnosis field for the encounter

(applies only to care in the inpatient and non-acute institutional settings), an indicator for the care setting in which the diagnosis was recorded, and an indicator for the type of diagnosis code.

Procedure: Similar to diagnoses, most inpatient and ambulatory/outpatient encounters are associated with one or more procedures. The procedure table is linked to the encounter table in a many-to-one 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. Currently many coding standards are used to record procedures, including: the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) procedure codes; Current Procedural Terminology, Fourth Revision (CPT-4) codes; and Healthcare Common Procedure Coding System (HCPCS) codes. NDCs administered by a provider may also be reported in this table. The table allows capture of any existing or future coding standards.

This "long and thin" structure for diagnosis and procedure tables facilitates searching for specific diagnosis/procedure codes by allowing a single pass through the table. For example, a single pass through 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 and recorded in the Encounter table. However, many patients die outside the clinical setting. Therefore, to determine death status, many of the Data Partners link to local (state) death registries to update the death status of their members. This update is performed infrequently—about once a year for most Data Partners. As a result, a two-year lag in death data is not uncommon in the MSDD due to a data lag in the death registry itself. The table includes death date, death date imputation information, source of death data, and an indicator of how confident the Data Partner is that the member drawn from the source data represents the actual member.

Cause of Death: Since each death can be associated with one or more contributing conditions, the death table is linked to a 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. The table also includes an indicator of how confident the Data Partner is that the cause of death information is accurate based on source of information, member match, number of reporting sources used, and discrepancies among sources.

Laboratory Result: The table includes data from selected laboratory tests captured by 12 (of 18) Data Partners. Because laboratory results can have different interpretations based on type of test or method of test administration, the model also includes variables for test subcategory, specimen source, patient location, result location, and original and standardized result units. In addition, the table includes variables for Logical Observation Identifiers Names and Codes (LOINC), immediacy of the test (e.g., stat), procedure code and code type to assist with rule-outs, order date, lab date/time, result date/time, original (non-standardized result), normal ranges, abnormal result indicator, and local codes for the ordering provider department and facility.

The laboratory results table includes a list of currently known LOINC and CPT-4 codes associated with each laboratory test of interest. The LOINC list, although not necessarily exhaustive, is a helpful tool for the Data Partners as they seek to extract laboratory test results data from their source databases. CPT-4 codes are billing codes and are provided as a courtesy to Data Partners; CPT-4 codes are of very limited assistance in extracting laboratory test results correctly from source databases. The model also includes a table of standard abbreviations for common laboratory units.

The laboratory results found in the table are from blood, serum, or plasma unless noted and include: alkaline phosphatase (ALP); alanine aminotransferase (ALT); absolute neutrophil count (ANC); total bilirubin; creatine kinase total; creatine kinase MB fraction; creatine kinase MB relative index (creatine kinase MB fraction divided by creatine kinase total); creatinine; fibrin d-dimer; glucose; hemoglobin; glycosylated hemoglobin (HbA1c); influenza (throat, nasopharynx, bronchoalveolar lavage, bronchoalveolar biopsy, nasal swab, nasal wash, or sputum); international normalized ratio (INR); lipase; pregnancy (urine or serum); platelet count; troponin I; and troponin T.

Vital Signs: The vitals table includes the unique person identifier, date/time the vital signs were measured, height, weight, systolic and diastolic blood pressure, blood pressure type, position, and tobacco-use status. Nine Data Partners are currently contributing information for this table.

State Vaccine: The Mini-Sentinel Post-Licensure Rapid Immunization Safety Monitoring (PRISM) Program has created the State Vaccine Table to capture state vaccine registry information collected by the four PRISM Data Partners. The State Vaccine Table contains vaccination records received from state Immunization Information Systems (IIS) for patients identified and matched from selected Data Partners. The Data Partners and the State IIS offices manage the process for linking health plan members to the state registry and populating the Vaccine Table that resides with the Data Partners as part of the MSCDM. It contains one record per vaccination, defined as a unique combination of a unique person identifier, vaccination data, and vaccine type, provider and administration type. The table includes information on the vaccination lot number and manufacturer. Vaccines can be coded using various systems, including CPT-4 and Codes for Vaccine Administered (CVX) codes. The PRISM team manages updates and data quality checking of the State Vaccine Table.

III. EXPANSION OF THE MINI-SENTINEL COMMON DATA MODEL

A. CLINICAL DATA ELEMENTS

1. Overview

In Year Five, the Mini-Sentinel Clinical Data Elements Workgroup led and/or collaborated on laboratory result data availability, completeness, expansion, and utilization activities. Data availability and completeness activities included gaining a better understanding of the availability by surveying Data Partners and writing and implementing a data characterization program. Expansion involved harmonizing seven laboratory result test types. Data utilization involved developing a user guide, collaborating on development and implementation of Combo Tool features that involve the laboratory result table, writing and publishing the manuscript "Electronic Clinical Laboratory Test Results Data Tables: Lessons from Mini-Sentinel." [12]

2. Roles and Responsibilities

The Clinical Data Elements Workgroup lead team is comprised of the Data Core co-leads, members of the MSOC and FDA. It leads and manages all aspects of the workgroup, including:

- At least bi-weekly conference calls to address deliverables and ensure adherence to timelines
- Monthly Data Partner webinar and conference calls
- Communications with Data Partners and FDA
- Support, guidance, and assistance to Data Partners in incorporating, characterizing, and harmonizing laboratory test results
- Completing revisions to MSCDM Laboratory Results Table as a result of data characterization
- Providing reports and updates to the Data Core
- Revising programming needed to characterize and harmonize the laboratory results data
- Implementing Combo Tool enhancements for clinical laboratory test result data elements
- Writing and conducting a survey to collect details on the methods used by each Data Partner during the ETL process for source data and to collect metadata to better understand the completeness and limitations of the Laboratory Results and Vital Signs Tables
- Compare the completeness of laboratory test results and procedure claims for laboratory tests

Writing and publishing the manuscript, "Electronic Clinical Laboratory Test Results Data Tables: Lessons from Mini-Sentinel", [12] to enhance the visibility of the Mini-Sentinel Laboratory Results Table.

3. Selection of Data Elements

In Years Two and Three, the first laboratory test results were incorporated into the MSCDM. These included many types of blood chemistry, hematology, and coagulation tests; urine tests (for pregnancy); and influenza tests ([spe]cimen sources included nasal swab or wash, and oropharyngeal swab). By the end of Year Three, laboratory test results included in the MSCDM:

- glucose (random and fasting)
- hemoglobin
- HbA1c
- creatinine
- ALT
- alkaline phosphatase
- total bilirubin
- INR
- D-dimer (qualitative and quantitative)
- lipase
- absolute neutrophil count (ANC)
- troponin-T
- troponin-I
- platelets
- total CK
- CK-MB fraction
- CK-MB/CK total index
- Pregnancy (urine and blood, quantitative and qualitative)

- influenza testing

In Years Four and Five, these data have been refreshed by the 12 Data Partners that provide laboratory data. All have now updated the laboratory results data through at least 2012. Most have updated through 2013, and some have updated through the first several months of 2014.

Vital signs incorporated into the MSCDM during Years Two and Three included height, weight, systolic blood pressure, diastolic blood pressure, and tobacco status. In Years Four and Five, no additional vital sign data elements were added, but existing vital sign data elements were updated by all nine Data Partners that contribute vital sign data (six Kaiser Permanente regions and three HMO Research Network Data Partners). In Year Five, the nine Data Partners that contribute vital sign data updated these data elements through at least 2012. Most Data Partners updated through calendar year 2013, and some have updated through the first several months of 2014 data.

Further information regarding building the clinical components' data model can be found in the Mini-Sentinel Coordinating Center Data Group: Year Four Report of Activities.

4. Revisions and Implementation of the Data Model for Clinical Data

Activities during Years Two through Four focused on incorporating types of laboratory test results into the MSCDM, expanding laboratory result table capabilities and uses, and incorporating the laboratory result table into regular MSDD updates and quality checks.

In Year Five, attention continued to be focused on the quality of lab results data. Characterization and harmonization of a second group of laboratory test types was completed. The Clinical Data Elements Workgroup collaborated on development and implementation of the Combo Tool which enables use of laboratory test results in defining complex combinations of exposures and events and cross-table querying between the laboratory result table and other MSCDM tables. The Workgroup gained greater understanding of the availability of clinical data elements in the laboratory result and vital signs tables via two activities: 1) development and implementation of a metadata survey completed by Data Partners; and 2) development of a SAS program that performs detailed characterization of completeness of test result data in the laboratory result table as compared to laboratory claims for in the procedure table within and between Data Partners.

As data characterization activities revealed refinement was necessary for comprehensiveness or clarity, the laboratory result data dictionary was updated. The revised data model for the laboratory result table is included in Overview and Description of the Common Data Model.

a. Data Characterization and Harmonization

As mentioned above, characterization and harmonization of a second group of laboratory test types was conducted in Year Five to harmonize laboratory results data across Data Partners and make these data operational in routine Mini-Sentinel data activities. A systematic process (similar to the one used in Year Four) was used to determine which laboratory test types were priorities for characterization. In collaboration with the FDA, seven laboratory test types were selected for Year Five characterization. These test types included:

- hemoglobin

- total bilirubin
- alkaline phosphatase
- lipase
- platelets
- ANC
- pregnancy tests (urine and blood, qualitative and quantitative)

Following a process similar to that used in Year Four, the workgroup developed and implemented programming code to characterize and analyze these laboratory test results. The analytic code enables assessment of the laboratory test result values by other lab test characteristics (e.g., test subcategories, result units, LOINCs, patient location, specimen sources) within and across Data Partners. The Data Partners ran the characterization programs against their laboratory result table and returned the summarized results for evaluation. Evaluation proceeded on a test type by test type basis. The evaluation allowed assessment of the variability in data source and helped guide the workgroup in developing an approach for standardization within and across test types.

Once the Lead Team evaluated a test type, guidance was provided to Data Partners about how to proceed to harmonize those laboratory results data. For example, the workgroup identified a wide range of platelet result units in the data, including the subset of units shown in the **Table 1** below.

Table 1. Platelet Count Original Result Units

blank	FL	TH/UL	X10(3)
%	K/CMM	THOU/CMM	THOUSAND
/100 W	k/cmm	thou/cmm	X10(3)/MCL
/CMM	K/CU MM	thou/mm3	X10(3)/UL
CMM	K/CUMM	THOU/UL	X10(6)/MCL
10 3 L	K/MCL	THOUS/CU.MM	X10*9/L
10X3UL	K/mcL	THOUS/MCL	X10E3/UL
10^3/UL	K/UL	THOU/mcL	X1000
10*3/uL	k/uL	THOUS/UL	X10X3
10?3/uL	KU/L	Thou/uL	X10^3/UL
10E3/uL	K/MM3	THOUSA	x10
10e3/uL	K/mm3	THOUSAND	X10?3/ul
10e9/L	LB	THOUSAND/UL	X10E3/UL
E9/L	PLATELET CO	U	X10E3
BIL/L	T/CMM	X 10-3/UL	K/A?L
bil/L	TH/MM3	X 10(3)/UL	K/B5L
CU MM	th/mm3	X10 3	1000/UL

Using the platelet test type as an example, **Figure 2** shows the kind of guidance provided to Data Partners.

Figure 2. Instructions to Data Partners for Harmonization of Platelet Test Result Data

1. Please remove any PLATELETS result values that cannot be converted to a numeric value.

2. For PLATELETS results only, please remove all records where the original result units are "%", "U/L", "U/ML", "IU/L", "IU/ML", or "MEQ/L".

3. Please set MS_Result_unit to "UNKNOWN" for all records where the original result unit is missing or blank, "NULL", "N/A" or "UNK". In addition, MS_Result_unit should be set to "UNKNOWN" for any records where Orig_Result_unit has values that are missing a numerator or a denominator.

4. Unless told otherwise by the Lead Team, please set MS_Result_unit to "K/UL", indicating "thousand per microliter", for all other PLATELETS records.

- To note, there are many ways of writing "thousand per microliter". Some helpful reminders are listed below:
  - "thousand" may be written as "K", "10*3", or many permutations of this.
  - One cubic millimeter of blood is equivalent to one microliter (UL). Cubic millimeter is often written as "MM*3" or "CU MM".
  - "billion per liter" is equivalent to K/UL.

5. For PLATELETS results only: for any records where Orig_Result_unit/Std_Result_unit can be directly converted to K/UL, please populate MS_Result_N using conversion factors listed in the below table.

6. Due to the many ways of expressing "K/UL", the Lead Team has provided an extended conversion table to help Data Partners populate Std_Result_unit and MS_Result_unit. For any records where Orig_Result_unit/Std_Result_unit can be directly converted to K/UL, please populate MS_Result_N using conversion factors listed in this attachment.

A summary is listed below, but please refer to the attachment for any result units not listed.

If Std_Result_unit =	***MS_Result_N =***	***MS_Result_unit =***
K/UL	*Orig_Result * 1*	*K/UL*
BIL/L	*Orig_Result *1*	*K/UL*
*X10*6/UL*	*Orig_Result *.001*	*K/UL*
K/ML	*Orig_Result *.001*	*K/UL*
K	*Orig_Result *1*	*UNKNOWN*
UL	*Orig_Result *1*	*UNKNOWN*
/UL	*Orig_Result *1*	*UNKNOWN*
10	*Orig_Result *1*	*UNKNOWN*

**If you find result units that look like they can be converted to K/UL and are not listed in the above table or the attachment, please contact the MSOC for guidance.*

For example, if you have a PLATELETS result of "200 X10^3/mm3 " in your source data, the variables for your MSCDM Laboratory Result Table should be as follows:

Orig_Result = 200
Orig_Result_unit = X10^3/mm3
Std_Result_unit = K/UL
MS_Result_N = 200
MS_Result_unit = K/UL

A similar approach to characterization and harmonization was used for each of the seven test types extensively evaluated in Year Five. Participating Data Partners completed data characterization and the workgroup provided guidance for harmonizing all seven of these laboratory results test types. Completion of data harmonization was confirmed in collaboration with the Data Model and Quality Assurance Group. Thus, 13 laboratory result test types are available, harmonized, and ready for use in Mini-Sentinel assessments.

Although the focus in Year Five was on laboratory data, data quality checks were also routinely conducted on both the structure and results of the vital signs table.

5. Querying Laboratory Result Data

In Year Four, investigators were able to query the laboratory result table to determine if a valid test result was present in the MSDD, but could not yet query a result based on its value.

In Year Five, the MSOC enhanced modular programs via the **Combo Tool** to allow querying of results values. This tool which increases the complexity possible for queries allows rapid querying of laboratory result values (e.g., define a cohort or outcome based on user-defined values of a laboratory test result). While enhancements allow investigators to query the results of any laboratory test type, the Clinical Data Elements Workgroup recommends that, when using MS programming tools, only laboratory test results that have been harmonized are queried.

6. Use of Standards and Controlled Terminologies

Data Partners use a mixture of LOINC and local battery and component codes to identify laboratory test result types. The LOINC and local codes are mapped to the MSDD laboratory result test type standards. There is substantial variability in the extent to which Data Partners use LOINC versus local codes. Some Data Partners have LOINC codes available to identify all results for a specific laboratory test and some have no LOINC codes at all. The Clinical Data Elements Workgroup lead team is continuing to work with FDA and the Data Partners to assess whether more robust application of LOINC (or potentially other standards) is possible.

Laboratory test results can be qualitative (e.g., some urine and blood pregnancy results) or quantitative (e.g., blood glucose) and require standardization. Quantitative result units are frequently reported in different units (e.g., Units, U, IU) and must be standardized. This was done for the seven test types harmonized in Year Five. This standardization work is resource intensive. It must be done on a case-by-case basis to capture all possible values for assessment and mapping, and must be routinely re-evaluated as new test type codes are introduced to the data. In Year Five, re-evaluation was done for the six test types originally harmonized in Year Four. Qualitative pregnancy result units found in the data were standardized to “positive,” “negative,” “borderline,” or “undetermined.”

7. Potential Next Steps for Clinical Additions

Potential areas of future expansion work clinical data elements include:

- Characterize and harmonize the remaining clinical laboratory test type results already in the MSCDM
- Expand the types of laboratory test result data available in the MSCDM (e.g., lipid, thyroid)
- Continue to explore existing data elements to enhance understanding of availability, patterns, frequency, usefulness, and logic inconsistencies
- Explore the feasibility of incorporating inpatient laboratory test type data from one additional Data Partner
- Further enhance and refine programming tools to enable more robust feasibility assessments of laboratory test results and vital signs
- Develop a user guide to the clinical laboratory data captured, providing information such as the data's strengths and weaknesses, content of the tables, description of harmonization done, and which laboratory test types each Data Partner contributes.
- Develop, test, and implement a body mass index macro tool for use with adult height and weight data in the vital signs table

B. EXPANSION INTO INPATIENT DATA STREAMS

During Year Four, a partnership began with the Hospital Corporation of America (HCA) in an effort to expand the MSCDM to include inpatient data streams. Encounter-level inpatient claims received by many of the current large Data Partners do not enable the identification of specific inpatient drug administrations. Moreover, inpatient data streams are available for a minority of Data Partners accounting for less than 10% of enrollees in the MSDD. HCA is a national health care services company that includes about 165 hospitals and 115 freestanding surgery centers in 20 states. Approximately 4-5% of all inpatient care delivered in the country today is provided by HCA facilities.

In collaboration with HCA, the Data Group has developed inpatient pharmacy and transfusion tables for inclusion in the MSCDM. The inpatient pharmacy table includes NDCs, administration date and time, route of administration, dose, and patient and encounter identifiers. The inpatient transfusion table includes a transfusion product name, International Society of Blood Transfusion (ISBT) product code, transfusion start and end dates and times, patient location, and patient and encounter identifiers. The unique encounter identifier links each table with existing tables in the MSCDM.

Within HCA, work is underway to transform HCA data into the MSCDM format including the newly specified inpatient pharmacy and transfusion tables. The assessment of feasibility will occur in the context of a specific evaluation: the association of red blood cell/plasma/platelet transfusion with transfusion related acute lung injury (TRALI).

C. OTHER REVISIONS TO THE MSCDM

1. New Variables

The MSCDM v4.0 introduced three variables: the Chart variable was added to the Enrollment table to indicate whether medical charts can be requested for use in validation studies; Zip and Zip_Date were added to the Demographic table to allow for assessment of socio-demographic confounders. The Zip

variable captures the most current 5-digit zip code for an enrollee whereas Zip_Date captures the earliest date at which that zip information is believed to be accurately captured. The Zip and Zip_Date variables are populated for approximately 84.5% of all PatIDs in the MSDD and chart variable is currently populated for approximately 96.3% of the MSDD. As of Year Five, the routine querying programs (i.e., modular programs and CIDA) do not have the capability to query Zip and Zip_Date; these functionalities will be added to querying capabilities in the near future. Modular programs can restrict cohorts based on the chart value variable.

2. Variable Value Revisions

The MSCDM v4.0 includes several text improvements to the laboratory results table. For influenza tests, several valid specimen source values were added to allow a more specific capture of the method used to collect nasopharyngeal specimen, and a generic nasopharyngeal specimen source code has been retired. Pregnancy tests now allow unknown specimen sources. Two LOINC codes were added for ANC, HgbA1C, and ALP tests, and one LOINC code was added for quantitative d-dimer tests and quantitative pregnancy tests.

3. New Tables

Prior to Year Five, the Mini-Sentinel PRISM Program developed a birth table and a fetal death table, designed to be populated with data sourced from governmental vital records agencies. PRISM also developed a Mother Baby Internal Linkage table, which identifies links between mothers' deliveries and infants in Data Partner data. Some Data Partners populated this table in Year Four and some in Year Five. During Year Five, work has continued with the four PRISM Data Partners (i.e., Aetna, HealthCore, Inc., Humana, and Optum) to obtain access to birth and fetal death data from additional jurisdictions and begin the transformation of that data into the birth and fetal death tables. The transformation is done using programming code written by the MSOC, explicitly for this purpose, which can be modified for reuse in similar data transformations.

D. LESSONS LEARNED

1. Clinical Data Elements

Incorporation of clinical data into the MSCDM and the subsequent use of those data for safety surveillance require careful attention to how the data are collected, captured, standardized, and stored as well as to the sub-populations that have clinical data available. To that end, MSOC will need to continue to develop user guides and provide education. The availability and missingness of laboratory test result data and recommendations for use of the data are areas for possible additional work. Exploring additional inpatient data is also a suggested future focus. Such information will enable other Mini-Sentinel teams and workgroups to make informed use of these clinical data tables.

2. Expansion into Inpatient Data Streams

As with Incorporation of any new data streams, adding inpatient data to the MSCDM requires identification of the data elements most relevant for the intended use and careful attention to how the data are captured in source systems and standardization across those systems. In addition, engaging local experts with deep knowledge of source systems is essential.

3. Other Revisions to MSCDM – New Tables

Obtaining vital records data is complicated and requires many staff resources and time for managing the application processes and communications. Each Data Partner-vital records agency pair is unique and close contact, between the two, is required.

IV. MINI-SENTINEL DISTRIBUTED DATABASE (MSDD)

A. DATA QUALITY ASSURANCE REVIEW AND CHARACTERIZATION

1. Overview

For each data refresh, the Data Partner updates its data in accordance with the MSCDM to produce a new ETL (Extract-Transform-Load). Most Data Partners refresh their data on a quarterly basis, while others do it either once or twice a year. The data refresh process is described in detail in Section III of our Year One Common Data Model Report. The Data Partner's new ETL is run through a series of quality assurance (QA) programs developed by the MSOC which return aggregated data that capture deviations from the MSCDM specifications, site-specific idiosyncracies, and changes in data over time. Before a new ETL is approved for use, it undergoes a QA review and characterization process resulting in communication between the MSOC and the Data Partner to resolve issues, as needed.

2. Roles and Responsibilities

The Data Model and Quality Assurance Group leads the data QA review process. After every data refresh, the Data Partner runs the data QA review and characterization programs against their new dataset. Analysts in the group review aggregated data output by the QA programs, draft a findings report, and communicate with the Data Partner to determine next steps. Depending on the findings, the MSOC may approve the refresh, approve with specifications for corrections to be made in the following refresh, or reject the refresh, requiring that the Data Partner apply corrections and regenerate the ETL. Regeneration of the ETL requires a complete new QA review. The specific steps included in the refresh process are described in Mini-Sentinel Standard Operating Procedure for Data Quality Review and Characterization. The following is a high-level summary of the data QA process:

- Data Partner performs a data refresh and produces a new ETL
- Data Partner executes QA programs against the new ETL
- Data Partner reviews outputs from the QA programs, revises ETL as necessary, and re-runs QA programs
- Data Model and Quality Assurance Group reviews QA program output, within and across ETLs for the Data Partner
- Data Model and Quality Assurance Group provides QA findings report to Data Partner
- Data Model and Quality Assurance Group and Data Partner review and discuss data QA findings report, agreeing to any necessary changes and their timeline
- Data Model and Quality Assurance Group approves the data QA

In Year Five, the MSOC implemented **Common Components** (CC) one aspect of which is ETL version control. Due to the interaction between the CC and QA programs, the updating and testing of CC was added as an additional step to the ETL approval process:

- Data Partner updates Common Components
- Data Model and Quality Assurance Group sends a Common Components test package to the Data Partner to validate Common Components update
- The Data Partner runs the Common Components test package and reviews results
- Data Model and Quality Assurance Group reviews and approves the Common Components test output and approves the ETL for Query Fulfillment use

After the ETL is approved, the Data Partner executes a SAS program that creates Summary Tables. The Data Partner sends the log from the SAS program to the MSOC. Once it is approved, the Data Partner links the Summary Tables to the Mini-Sentinel Query Tool and executes a "metadata refresh dates" query against its data. This query provides information to MSOC on the range of dates for which data is available for each query type for the Data Partner.

3. Data QA Review and Characterization Specifications

The Mini-Sentinel project relies on the comprehensiveness and quality of the data available in the MSDD. The Data Model and Quality Assurance Group 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 that MSDD data meet quality expectations, the Scientific Operations Center developed a series of measures to check data quality and characterize the breadth and depth of the data available for querying. These measures address missing data, deviations from the MSCDM (e.g., invalid values, invalid date ranges), and logical inconsistencies. The design and scope of the data QA review and characterization process balances adherence to the MSCDM with the expected variability across Data Partners, based on differences in manner of data capture and data source (e.g., administrative and claims data, electronic health record data). The data quality review done after each data refresh is organized into four levels of data characterization, based on the complexity of the checks performed. A description of the data characterization approach can be found in Data Quality Review and Characterization Programs.[ii]

a. Level 1 Data QA Review and Characterization

The Level 1 data checks review completeness and content of each variable in each table to ensure that formats and values of required variables conform MSCDM specifications. The data QA review program verifies 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. Its value must be in the range of January 1, 1885, through the date on which the demographic table was created. Categorical variables must include only the values specified in the MSCDM. **Table 2** illustrates several Level 1 data QA review and characterization items for the dispensing table.

[ii] Version 3.1.2 is the version referenced in this report. The link will bring the reader to the version current at the time of reading. Information about prior versions will be available at the link.

Table 2. Level 1 Data QA Review and Characterization: Example for the Dispensing Table

	Variable Name	Description of Error or Data Characteristic	Error Code
1	PatID	PatID variable is not character type	DIS1.1.1
	PatID	PatID variable has missing values	DIS1.1.2
	PatID	PatID variable has values that are not left-justified	DIS1.1.3
	PatID	PatID variable contains special characters	DIS1.1.4
2	RxDate	RxDate variable is not a SAS date value of numeric data type	DIS1.2.1
	RxDate	RxDate variable is not of length 4	DIS1.2.2
	RxDate	RxDate variable has missing values	DIS1.2.3
3	NDC	NDC variable is not character data type	DIS1.3.1
	NDC	NDC variable is not exactly 11 characters in length	DIS1.3.2
	NDC	NDC variable has missing values	DIS1.3.3
	NDC	NDC variable contains special characters or non-digits	DIS1.3.4
4	RxSup	RxSup variable is not numeric type	DIS1.4.1
	RxSup	RxSup variable is not of length 4	DIS1.4.2
	RxSup	RxSup variable has negative, missing, or zero values	DIS1.4.3
5	RxAmt	RxAmt variable is not numeric type	DIS1.5.1
	RxAmt	RxAmt variable is not of length 4	DIS1.5.2
	RxAmt	RxAmt variable has negative, missing or zero values	DIS1.5.3

b. Level 2 Data QA Review and Characterization

Level 2 data checks 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, PatID, 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 members for whom enrollment spans are created, corresponding demographic information exists. The converse PatID matching is also checked to determine how many PatIDs with demographic information do not have enrollment information. This represents a data characteristic as opposed to a data error because some Data Partners provide demographic information on un-enrolled members. **Table 3** illustrates several Level 2 data QA review characterization items for the enrollment table.

Table 3. Level 2 Data QA Review and Characterization: Example for the Enrollment Table

	Variable Name	Description of Error or Data Characteristic	Error Code
		Record(s) have duplicate key value combinations (with respect to table definition)	ENR2.0.0
1	PatID	At least one PatID in the DEM table is not in the ENR table	ENR_DEM2.1.1
	PatID	At least one PatID in the ENR table is not in the DEM table	ENR_DEM2.1.10
2	Enr_Start	Enr_Start is after Enr_End	ENR2.2.1
	Enr_Start	Enr_Start occurs more than once in the file in combination with PatID, MedCov, and DrugCov	ENR2.2.3
3	Enr_End	Enr_End occurs more than once in the file in combination with PatID, MedCov, and DrugCov	ENR2.3.4

The data QA review and characterization programs generate Level 1 and Level 2 data checking output, which is sent to MSOC for review. Anomalies are reported to the Data Partners to determine whether the issues need to be fixed or are part of the underlying data characteristics. 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 will not signal an alert in the next data checking process.

c. Level 3 Data QA Review and Characterization

In contrast to the Level 1 and Level 2 data checks, the Level 3 data checks "profile" the data, focusing on characterizations that do not require a specific outcome or True/False finding. Rather, these checks provide high-level qualitative and quantitative counts and proportions for analyzing patterns, trends and data characteristics over time and across Data Partners. For example, trends in the number of outpatient dispensings per person or the rate of hospitalizations should follow similar patterns across Data Partners, and any obvious divergence from the general trend requires investigation. This profiling characterizes specific data variables for each Data Partner and aggregates information for cross-institutional comparisons. Level 3 data characterizations also evaluate trends to help identify data gaps and unusual patterns both within an ETL and across a Data Partners' ETLs. Examples of trends within a single ETL include:

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

Examples of trends across ETLs include the number of members and number of records—both of which are expected to increase with each ETL. Other Level 3 data characterization topics include counts of procedures per encounter by encounter type and year; and counts of 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.

Examples of Level 3 data characterizations for the dispensing table are:

- Overall table statistics
  - Number of records in the table
  - Number of unique PatIDs
- Distribution of dispensing date (RxDate)
  - Dispensings overall, by month, and by year, within and across ETLs
- 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 MSOC 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 might indicate an issue with how the MSCDM was populated. Or, if the age proportions at one Data Partner are substantially different from the other Data Partners, it might reveal a difference in the underlying populations. Active participation from the Data Partners is essential to address unexplained variability. This level of data check is not intended to find 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.

d. Level 4 Data QA Review and Characterization

Level 4 data checks provide more targeted data analyses and profiling. Level 4 checks can be used to look for nonsense diagnoses in the data and variations in care practices across Data Partners. The checks inspect:

- Number of encounters with a hysterectomy procedure by sex
- Number of encounters with an ovarian cancer diagnosis by sex
- Number of encounters with a prostate cancer diagnosis by sex
- Number of encounters with a pregnancy diagnosis or procedure by sex
- Rates of emergency department encounters that become in-patient hospital encounters

4. Data QA Review and Characterization Revisions

The MSOC released four new versions of the data QA review and characterization programs during Year Five, to incorporate feedback from Data Partners and expand commonour knowledge of the MSDD. Versions 3.1.1 and 3.1.2 were released in September 2013 and added bug fixes and code enhancements. Version 3.1.3 was released in February 2014 to omit local procedure codes in the QA output. Version 3.2 was released in May 2014 to integrate Common Components as well as add minor enhancements. The

programs and release notes are available on the Mini-Sentinel website within the Distributed Database and Common Data Model Section.

5. Reporting

Results of data QA review and characterization activities are shared with the Data Partners. Two annual companion documents—the Mini-Sentinel Data Quality and Characterization Procedures and Findings Report and the MSDD Summary Report[iii]—provide details of the data QA review and characterization activities and results across all Data Partners.

6. Data Completeness and Availability

A data availability and completeness report is generated on a quarterly basis. The data availability graphs provide an MSDD table-centric overview of: 1) which Data Partners have data available for the five main MSDD tables (enrollment, dispensing, encounter, diagnosis, procedure); and 2) date ranges covered. The data completeness graphs provide a Data Partner-centric overview of the same data availability information, overlaid with vertical lines to indicate the first and last month of stable, complete data for each Data Partner. Updated reports are not posted on the public website, but are shared with the FDA.

7. Principal Diagnosis Flag (PDX) Variable Investigation

In Year Four, in response to several queries by the FDA and Mini-Sentinel workgroups, the Scientific Operations Center led a detailed investigation into how Data Partners populate the principal diagnosis flag (PDX) variable in the MSCDM Diagnosis table. A comprehensive survey and related distributed SAS program were developed and sent to Data Partners. The survey responses and data generated by the SAS program were reviewed and the findings were reported in Year Five. The investigation has helped guide use of the PDX variable and led to conversations with Data Partners that resulted in more specific guidance for populating the variable. SAS programs are in development in Year Five to verify compliance with this guidance and to evaluate the impact of guidance changes on the distribution of PDX values.

8. Data Partner ETL Survey

In Year Five, two questionnaires were developed and distributed to Data Partners to gather information about their data and better understand the methods and algorithms used to transform the data into the MSCDM. Information gathered from these questionnaires will be used to inform the use of data and enhance MSCDM guidance provided to the Data Partners.

The core questionnaire queries Data Partners as to data transformation practices and methods used to populate the enrollment, demographic, dispensing, encounter, diagnosis, and procedure tables and their variables. Examples of questions in the core questionnaire include the Data Partners' plan for transitioning to ICD-10 and its expected impact on the MSDD, and whether the Data Partner limits the

[iii] The annual Data Quality and Characterization Procedures and Findings document is no longer posted to the public website, but is shared with the FDA.

number of diagnoses per claim reported in the MSDD. The clinical data elements questionnaire, developed in consultation with the Clinical Data Elements Workgroup, focuses on the scope of the laboratory results data. This questionnaire gathers details about the sources of laboratory data, information used to extract relevant data, and any known limitations to which records or populations are included in the MSDD dataset.

The responses to the ETL questionnaire were summarized in a report that the Data Model and Quality Assurance group uses to improve their understanding of the variation within the MSDD.

B. INCORPORATION OF NATIONAL DATA STANDARDS AND CONTROLLED TERMINOLOGIES

MS Scientific Operations Center is committed to adoption and use of relevant national terminology standards related to electronic health care data. The primary activities under this task are incorporation of standards into the MSCDM, including plans for changing standards such as the approaching adoption of ICD-10 coding, and engagement with standards bodies, as directed by FDA.

1. 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 the year 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 transparent and easily-understood data model that all Data Partners can implement consistently within their existing electronic data capture systems. Currently, the Mini-Sentinel Data Partners use a limited yet comprehensive set of controlled terminologies to capture medical encounter, pharmacy dispensing, demographic, laboratory results, and health plan enrollment information. The information in MSDD represents the values found in the source files and does not include complex clinical mappings between coding standards or terminologies. Any necessary mappings can be done using the Mini-Sentinel analytic tools on a case-by-case basis. This approach minimizes the implementation and storage of unnecessary mappings, obviates the need to maintain multiple mappings that may or may not ever be used, and enables use of query-specific mappings based on the most recently available 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 and the Vaccine Safety Datalink. As described in the Mini-Sentinel Year One Common Data Model Report, 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. The current version of the MSCDM is available online and is updated as needed to improve clarity or add new data areas. The MSCDM was designed to accommodate other coding terminologies such as ICD-10 (see below for more information on 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, diagnoses are recorded on health insurance claims submitted for reimbursement and/or in electronic health record systems for Mini-Sentinel Partners that operate as integrated delivery systems. Each of our Data Partners use this standard terminology. The data model allows for inclusion of ICD-10 or any other diagnosis coding terminology. The data model was recently updated to specifically accommodate SNOMED CT codes.

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, drug infusions, and inpatient surgical procedures. Each of our Data Partners uses ICD-9 procedure and HCPCS codes. The data model allows for inclusion of ICD-10 or any other procedure coding terminology. Some data partners have non-standard local codes that can be included in the MSDD. The data model was recently updated to specifically accommodate Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT), LOINC, and National Drug Code (NDC) codes.

In addition, the Mini-Sentinel State Vaccine Table accommodates both CVX (Health Level 7 Table 0292, Vaccine Administered) and MVX (Health Level 7 Table 0227, Manufacturers of Vaccines) codes describing vaccine administration and manufacture. This table is created via linkage to selected state immunization registries to facilitate vaccine-specific activities. The CDC's National Center of Immunization and Respiratory Diseases maintains Health Level 7 standards for vaccine administration that are based CVX and MVX codes. CVX codes refer to the vaccine administered and MVX codes refer to the manufacturer.[vii]

Outpatient Pharmacy Dispensings: Pharmacy dispensings are identified using NDCs that are recorded by pharmacies at the point of dispensing to the patient. Each of our Data Partners uses this standard pharmacy dispensing terminology. Medications dispensed in the inpatient setting are not currently available from the Data Partners and are not included in the Dispensing Table.

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

[iv] CDC/National Center for Health Statistics. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Available at: http://www.cdc.gov/nchs/icd/icd9cm.htm

[v] Centers for Medicare & Medicaid Services. HCPCS - General Information. Available at: http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html

[vi] American Medical Association. About CPT®. Available at: http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.page?

[vii] Centers for Disease Control and Prevention. Immunization Information Systems. Available at: http://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx

[viii] CDC/National Center for Health Statistics. International Classification of Diseases, Tenth Revision (ICD-10). Available at: http://www.cdc.gov/nchs/icd/icd10.htm

Laboratory Results: Our Data Partners use a mixture of LOINC and local codes to identify laboratory test result types such as influenza A, influenza B, creatinine, and pregnancy. The local LOINC and local codes are mapped to the Mini-Sentinel laboratory result test type nomenclature. To the extent possible, LOINC codes are used to identify laboratory result types. Laboratory test result units also must be standardized to a set of uniform unit types. Laboratory test results can be numeric or text. For example, '+', '++', 'POS', and 'positive' are all potential pregnancy result units found in the source data. To enable distributed querying those results units must be standardized. In addition, numeric results could be measured in different units such as per liter or per microliter, and those units could be represented in a variety of ways (e.g., 'k', 'K', and '10e3' refer to thousands and 'uL', 'UL' U L' 'mcl', and 'cumm' are variations of a microliter). The MSCDM uses a standard abbreviation of 'UL' for microliter to enable distributed querying. Some data partners have non-standard local codes that can be included in the MSDD.

Mini-Sentinel investigators recently published a paper describing the laboratory working group activities to standardize laboratory result values.

Although that data model has been updated to accommodate a wide range of coding terminologies (eg, SNOMED-CT in Diagnosis and procedure tables, NDCs in the procedure table) that are expected to be increasingly adopted by electronic health record systems and some health plans, the Mini-Sentinel Data Partners do not uniformly capture information using all terminologies. MSOC will continue to work with FDA and the Data Partners to assess inclusion of these and other standards as possible.

2. Engagement with National Standards Organizations

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[ix] related to data capture and transmission promulgated by the Office of National Coordinator for Health Information Technology (ONC) have the potential to standardize 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 Data Partners and the Mini-Sentinel distributed querying approach. All uses of Mini-Sentinel are "secondary uses" of electronic health data and are therefore not directly related to approaches and standards targeting point-of-care transmission of health information. 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, use by the Mini-sentinel Data Partners, and the needs of the FDA within the system.

FDA has identified the ONC Standards & Interoperability (S&I) Framework[x] 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

[ix] Office of National Coordinator for Health Information Technology (ONC). Meaningful Use Regulations. Available at: http://www.healthit.gov/policy-researchers-implementers/meaningful-use

[x] ONC Standards & Interoperability (S&I) Framework. Available at: http://www.siframework.org/

Framework activities, especially the S&I Framework Query Health Technical and Clinical Workgroups, and will remain engaged with those activities. Prior activity in this area included a Query Health pilot project to investigate the potential for incorporating inpatient and ambulatory electronic health record data querying within the Mini-Sentinel framework. The pilot focused on a widely-used standardized clinical data model – Informatics for Integrating Biology and the Bedside (i2b2) - and a newly-developed clinical querying approach called the Health Quality Measure Format (HQMF). A video of the integration[xi] and a related poster presentation[xii] are available online. Finally, the Mini-Sentinel Distributed Query Tool (based on PopMedNet™) adheres to ONC Query Health distributed querying standards. Additional information on Mini-Sentinel activities related to national data standards is provided in Section VII, D that described engagement with other national distributed networking initiatives.

3. Impact of Transition to ICD-10-CM

Although due to the extension of ICD-10-CM implementation requirements, we do not expect to observe ICD-10-CM coding recorded in the MSDD until October 2015. As mentioned in **Section IV.B**, the existing MSCDM and the existing modular programs can accommodate ICD-10-CM without any changes to the data model or programs. The data model uses an indicator variable for both diagnosis and procedure codes that allow data partners to indicate the type of code being used for the specific observation. The combination of the indicator variable and the code are used together determine the type of code recorded. For example, the variables "DX" and "DX_CODETYPE" together are used to identify the exact nature of a code in the diagnosis table. The "DX_CODETYPE" variable is used to indicate whether the code recorded is an ICD-9-CM, ICD-10-CM or any other type of code.

So although the MSCDM can accommodate use of new code types, the widespread adoption of a new coding standard will have implications for Mini-Sentinel. For example, widespread adoption of ICD-10-CM will require work on developing new HOI algorithms or validating mappings between ICD-9-CM and ICD-10-CM based algorithms. Since Mini-Sentinel uses longitudinal data, another complication is the potential need to use two different algorithms for analyses that span coding terminologies. These issues are not unique to Mini-Sentinel, but will be issues for all users of electronic health data, especially longitudinal secondary users of these data. MSOC will remain engaged with other stakeholders (e.g., federal agencies) who also use these data to help identify options and solutions for the adoption of new coding standards. Moreover, the Data Model and Quality Assurance Group will keep monitoring the occurrence of ICD-10-CM codes within the MSDD, and will disseminate any relevant information to FDA, core leaders, and MS collaborators to ensure priorities are established.

C. LESSONS LEARNED

Through communication with the Data Partners, the MSOC continues to improve MSCDM guidance and quality of Mini-Sentinel data. Selected lessons from Year Five:

[xi] YouTube. PopMedNet - i2b2 Integration for ONC Query Health Pilot. Available at: http://www.youtube.com/watch?v=sqDAo6E-b1o

[xii] PopMedNet™. Distributed Research Network Technologies for Population Medicine. Available at: http://www.popmednet.org/?page_id=39

1. Data Changes and Quality Improvements

Given the number and variety of Data Partners that contribute data to the MSDD and the long-term nature of the project, changes in data source and data trends are expected over time. For example, a Data Partner's transition to a new electronic medical system may necessitate a more rigorous Quality Assurance review and discussions with the Data Partner to understand the implications for the data. Although every effort is made to update data as soon as possible, it is sometimes necessary to defer a data refresh and engage closely with the Data Partners in order to ensure completeness and maintain quality standards.

2. MSCDM Guideline Clarity

The MSCDM specifications and guidelines for variables are written for programmers versed in the SAS programming language. For Data Partners who transform data from a non-SAS source into Mini Sentinel's SAS datasets, misunderstanding of the MSCDM specifications may create unintended results. Improving clarity regarding variable length and format in the MSCDM specifications will better inform Data Partners how to accurately transform their source data into the MSDD.

V. MINI-SENTINEL ANALYTIC TOOLS

A. PROGRAMMING TOOLS

1. Overview

Mini-Sentinel programming tools are SAS macros that can be executed, alone or in combination with other tools, against MSCDM-compliant data. These tools allow for rapid querying of the MSDD and standardize routine programming procedures.

Mini-Sentinel programming tools can be categorized as toolkit macros or rapid response query tools. Toolkit macros are SAS programs written to standardize routine programming procedures (e.g., extract dispensings from the MSDD dispensing table, create continuous enrollment spans). These macros can be used in combination with each other and other programming code as building blocks en route to developing more complex and comprehensive programs. Each toolkit macro is a self-contained program that performs a discrete function. Rapid response query tools utilize multiple toolkit macros to answer a specific *question* of interest (e.g., rate of occurrence of an outcome during exposure to a drug).

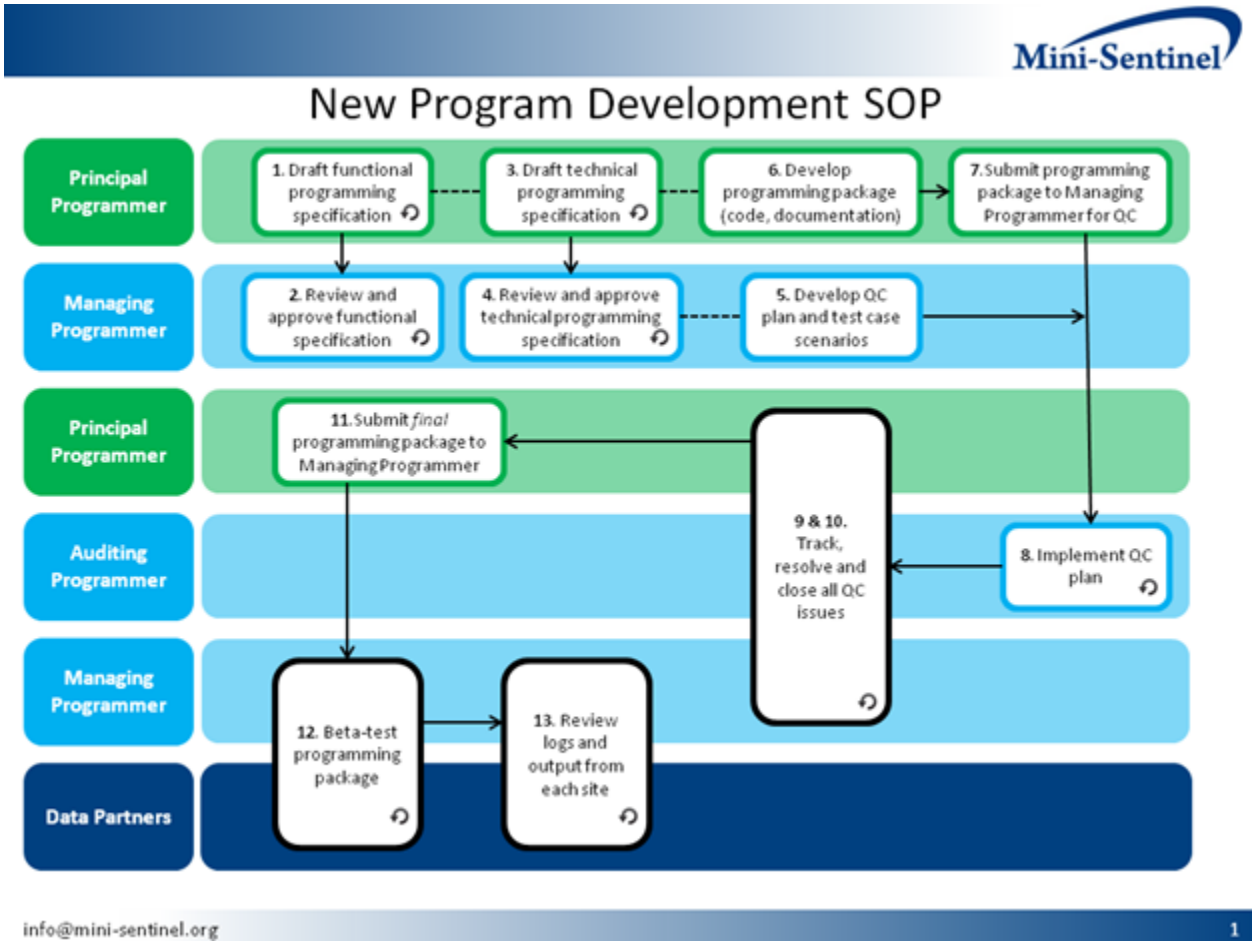
In Year Five, the MSOC began extensive integration and enhancement of the cohort identification, characterization, and descriptive analysis rapid query assessment tools. The goal is to replace the current suite of six SAS modular programs (MP) with one MP called the "**Cohort Identification and Descriptive Analysis**" (CIDA) tool. Maintaining a single MP will both speed up the programming development process and ensure consistency in cohort identification across all queries. In Year Five, the MSOC integrated MP3, MP6, and MP9 into the new MP, and plans, in future years, to integrate MP4, MP7, and MP8.

At the end of Year Five, the MSOC had a suite of four MPs responsible for cohort identification and descriptive analysis: the CIDA tool, MP4, MP7, and MP8. In Year Five, the MSOC also integrated an

Analytic Adjustment and an **Alerting and Sequential Analysis tool** into the suite of programs used for rapid response queries. MSOC also developed and released several new toolkit macros, many of which are also integrated with the CIDA tool.

All new tools and enhancements to existing tools underwent the Mini-Sentinel New Program Development process (see **Figure 3**).

Figure 3. SAS Program Development Process Flow



2. Roles and Responsibilities

The development and revision of Mini-Sentinel programming tools require careful planning for use of internal and external resources. The Infrastructure Group is responsible for the programming tool development process. During Year Five, both internal and external resources were used for SAS programming, testing, and Quality Compliance (QC). Data Partners supported the effort with the

validation of all new and enhanced tools. The roles and responsibilities of each group are described below.

Infrastructure Group:

- Prepare program development plan (what features to add and when)
- Identify new features of potential interest to FDA and workgroups
- Assess feasibility of new features, modules, and programs requested by FDA or workgroups
- Prepare specification and QC documents
- Ensure compliance with the Standard Operating Procedure for SAS Program Development
- Coordinate exchange of information between FDA, Mini-Sentinel Lead Team, Data Core, Data Partners, and External Programmers
- Update Data Core and Data Partners on status of programming tool development
- Hold training webinars for FDA, Data Core, and Data Partners
- Keep documentation and query request forms up-to-date; share with FDA as needed

External Programmers:

- Implement proposed new programming following specifications from Infrastructure Group
- Implement QC plans from Infrastructure Group
- Provide support to the MSOC, FDA, and Data Partners with interpretation and clarification of results

Data Partners:

- Test and validate new query tool releases
- Provide feedback on efficiency and functionality of tools

3. Rapid Response Query Tools

a. Cohort Identification, Characterization, and Descriptive Analysis Tools

Cohort Identification and Descriptive Analysis (CIDA) Tool: The CIDA tool identifies cohort(s), characterizes cohort(s)s using descriptive output tables and statistics, and performs minimally adjusted analyses (i.e., calculates incidence rate ratios comparing two identified cohorts, adjusted for age group, sex, year and/or data partner). The CIDA tool can be used to create simple cohorts to determine background rates of disease and prevalent and incident drug use, and also create more complex cohorts that identify an exposure, create treatment episodes of exposure based on dispensings days supplied (or create a user-defined exposure period), and look for the occurrence of a health outcome of interest (HOI) during exposed time. This functionality is a composite what was available in MP3, MP6, and MP9.

The CIDA tool includes several new features to enhance querying capabilities for both one-time assessments and prospective surveillance requests.

- **Exclude PatIDs:** The CIDA tool allows for the exclusion of members for specific types of data requests. The MSOC can exclude members from a request if they ever had a value of CHART=N in the MSDD enrollment table. This is useful for users who are interested in getting results/estimates for a subset of the population for which medical charts can be requested. Additionally, DPs have the ability to restrict certain members from consideration depending on

the type of output generated by a request. For example, a Data Partner can exclude administrative services only (ASO) members from requests that require the return of patient-level data to the MSOC.

- **Censoring:** Users can choose to censor treatment episodes or user-defined follow-up periods based on factors including: indication of death from the MSDD encounter or death table; hospitalization; or any NDC, procedure code, diagnosis code, or laboratory result value.
- **Laboratory Results:** Users can define exposures, outcomes, and inclusion/exclusion criteria using laboratory result values. Options include the ability to specify a single or a range of allowable values, and a user-specified algorithm for selecting which MSDD laboratory result table date value to use (e.g., lab date, result date, order date, in any user-defined hierarchical order).
- **Coverage Type and Enrollment Gap:** The type of coverage required and the allowable enrollment gap can now be specified by scenario, rather than by request. This means that these parameters can be adjusted in sensitivity analyses without requiring multiple executions of the program.
- **Combo Tool Integration:** The standalone programming tool "Combo Tool" was integrated with the CIDA tool, to allow for the specification of complex exposures, HOIs, and inclusion/exclusion criteria. Users can include validated algorithms and multiple criteria to specify an event of interest, including the ability to define the event at the start or end of a hospitalization of enrollment period.
- **Integration of the CIDA tool with Analytic and Alerting and Sequential Analysis tools:**
    - **Define Multiple Time Periods or "Looks"**: The CIDA tool can define multiple time periods by Data Partner to support prospective surveillance activities. One request package can handle requests for multiple time periods.
    - **Comorbidity Score:** This module was enhanced and integrated with the CIDA tool. It now uses the **combined Charlson/Elixhauser comorbidity score** rather than the Deyo Adaptation of the Charlson Comorbidity Index. The output was also enhanced to provide the information needed to include the comorbidity score in the propensity score matched model.
    - **Medical Utilization:** This module's output was enhanced to provide the information needed to include utilization metrics in the propensity score matched model.
    - **Covariate Specification:** There is now an input file available to specify pre-defined covariates for the propensity score matching tool. Users define the number of days prior to index date to observe covariates; covariate codes can be specified using wildcards, exact matching, and ranges of values.
    - **Analytic Datasets:** All analytic datasets required to perform propensity score matched analyses are created by the CIDA tool. Once created, the propensity score matching module can use the output to perform all subsequent analyses. Output is generated by time period, or "look," as needed.

Modular Program 4: In Year Four, MP4 was used to characterize concomitant use (secondary exposure following and overlapping a primary exposure) of outpatient pharmacy medication(s) and/or medical procedure(s), observed among members with or without a pre-existing condition, during a period defined by a start and end date. In Year Five, FDA requested an enhancement to MP4 to characterize the frequency of select event(s) during episodes of concomitant use. To achieve this, the event observation functionality of MP3 was added to MP4. In addition, the way that primary exposure, secondary

exposure, and concomitant exposure were defined was enhanced. The enhanced MP4 outputs metrics for 3 cohorts in each run: 1) a primary cohort, examining the risk of adverse events during primary exposure treatment episodes; 2) a secondary cohort, examining the risk of adverse events during secondary exposure treatment episodes; and 3) a concomitant cohort examining the risk of adverse events during concomitant exposure treatment episodes. In Year Six, several additional parameters will be added to allow increased flexibility to define concomitant exposure.

Modular Program 7: No enhancements to MP7 were made in Year Five. MP7 characterizes the "Top #" (user-defined) most frequently observed diagnosis, procedure, and drug codes during a user-defined period before and after an index date. Index event can be defined using any type of code, and results are provided for both prevalent and incident patients of the index event code(s). Standard output provides "Top #" rankings using both number of users and events, and rates for both prevalent and incident use of each most frequently used code are provided.

Modular Program 8: In Year Five, the MSOC published MP8, a program that characterizes the uptake, use, and persistence of new molecular entities (NMEs). New use of each NME can be defined by choosing options (e.g., length of pre-initiation enrollment, episode gap). Metrics reported include: monthly uptake rates, exposure to NMEs by number of treatment episodes, length of treatment episode (by first episode, second, etc.), gap (in days) between valid treatment episodes, and survival analysis.

b. Analytic Adjustment Tools

Propensity Score Matching Tool: Following the Active Surveillance Workgroup's[xiii] Year Four development of the propensity score matching module, in Year Five, the MSOC integrated the module with the CIDA tool for use in routine queries and prospective surveillance activities.

The propensity score matching tool uses the cohort(s) identified and output generated by the CIDA tool to perform propensity score matched analyses. Users can specify a model using pre-defined covariates and/or determine covariates based on a high-dimensional propensity score selection strategy. Users also determine the matching ratio (either 1:1 fixed or 1:100 variable matching) and the matching caliper (maximum allowed difference in propensity scores between treatment and control patients; options include 0.01, 0.025, and 0.05).

Additional options are available for the calculation of the high dimensional propensity score (hdPS). Users determine the number of covariates, by code type, to consider for inclusion in the hdPS model (e.g., drug, ICD9 diagnosis, ICD9 procedure, HCPCS, CPT). In addition, users can specify the number of covariates to keep in the final model.

The program automatically generates tables of patient characteristics for the unmatched cohort and for each matched cohort, stratified by exposure group and Data Partner. Tables include measures of covariate balance, including absolute and standardized differences, which indicate balance in specific variables, and the Mahalanobis distance, which provides a measure of balance across all variables while

[xiii] This workgroup was created under Mini-Sentinel's Year Three – Five Base Contract activity "4.10 Create Program for Routine Surveillance of Newly Approved Products".

accounting for their correlation. The tables also include the number of patients in each exposure group, the number matched from each group (where appropriate), the number that experienced outcomes, and the mean person-time of follow-up. The program also automatically generates figures depicting the propensity score distributions for each exposure group, separately for each Data Partner. Figures include c - statistics for each propensity score model. Using summarized data generated in the data extraction step, the program can estimate both hazard ratios (with 95% confidence intervals) and incidence rate differences (with 95% confidence intervals). Calculated confidence intervals do not account for repeated looks or correlation in the data across looks.

Incidence Risk Ratio (IRR): This program produces an automated comparison of two cohorts and their incidence rates. Cohorts are identified using the CIDA tool. The IRR tool allows for quick assessment of both crude and adjusted incidence rate ratios for the two groups by producing incidence rate ratio estimates and their corresponding 95% confidence intervals. It provides functionality to control for age, sex, year, and Data Partner within the adjusted rate ratio calculations. The tool utilizes a Poisson regression and a large sample approximation for calculation of the IRR.

c. Alerting and Sequential Analysis Tools

Binomial MaxSPRT: In Year Five, the MSOC began integration of the binomial maximized sequential probability ratio test (maxSPRT) program (Kulldorff et al. 2011) into the existing suite of tools.[11] The program can be used for sequential analysis in PROMPT requests that perform propensity score matched analyses. It can be used to alert investigators of potential excess risk of an outcome for an exposed cohort relative to a 1:1 matched comparator cohort. The program requires further integration with programming tools in Year Six to streamline execution and reduce manual data entry.

4. Toolbox Macros

Combined Elixhauser-Romano (CCI) tool: This program accounts for comorbid conditions by calculating a combined comorbidity score using an algorithm that applies a weight to ICD-9 diagnosis codes appearing in a patient's health history. The algorithm and program are based on the Gagne (2011) combined comorbidity score.[9]

Medical Utilization: This program uses encounter data from the MSDD to count the number of distinct visits from a reference date and within a look back period. Any encounter type(s) can be specified. The program provides the option of reporting total number of visits or visits by encounter type.

Continuous Enroll: The main purpose of this program is to determine whether individuals have health insurance plan enrollment before and/or after an index date. The macro allows the user to supply the pre- and post-index date period as any number of days, months or years on either end of the index date. Another feature of this macro reconciles overlaps and gaps in plan enrollment prior to joining the enrollment file with the index date file. This feature allows the user to specify the maximum number of days allowable as a gap in enrollment.

Combo Tool: Defining health outcomes of interest and medical product exposures sometimes requires complex algorithms. For example, a health outcome of interest could be defined as two ICD-9-CM codes occurring during the same medical encounter, with a specific lab test result occurring within seven days

of that encounter. The SAS "Combo Tool" allows for the identification of such complex combinations of events within the MSDD.

Process Wildcards: This program expands ICD-9 diagnosis codes supplied by a user as ranges (*e.g.*, 310-312) and/or wildcards (*e.g.*, 250**) to all possible values within those ranges or wildcards.

Extract Meds: This program extracts diagnosis and/or procedure claims from the MSCDM diagnosis and/or procedure tables. The extraction matches on code and code type supplied in a user-specified file.

Extract Drugs: This program extracts dispensing claims from the MSCDM dispensing table based on 9 and/or 11 digit NDC codes supplied in a user-specified file.

Extract Deaths: This program extracts deaths from the MSCDM death table and death-related encounters from the MSCDM encounter table.

B. SUMMARY TABLES

1. Overview

Another analytic tool used by MSOC is the Mini-Sentinel Distributed Query Tool, described in greater detail in the next section. This software application allows MSOC to quickly create and securely distribute simple queries on counts, prevalence, and incidence of drug products, diagnosis codes, and procedure codes to network Data Partners. Data Partners are then able to quickly review, execute, and securely return results of those queries to the requestor within two business days via a web-based Portal. Queries are run against each Data Partner's "Summary Tables" rather than against the entire MSDD.

All Data Partners create a set of 12 summary tables using a distributed program that runs against their MSDD. Summary tables are refreshed with each Data Partner data refresh. Summary tables include prevalence and incidence counts of dispensings, procedures, diagnoses, and enrollment stratified by year, sex, age group, and where applicable, care setting. Specifically, the eight prevalence summary tables represent prevalence counts of diagnoses (3-, 4-, and 5-digit ICD-9-CM), procedures (3- and 4-digit ICD-9-CM and HCPCS), and drug exposures (ingredient name and drug category). The three incidence summary tables represent incidence counts of diagnoses (3-digit ICD-9-CM) and drug exposures (ingredient name and drug category). An enrollment summary table provides enrollment information by coverage type. The code set used for the specifications for HCPCS, ICD-9-CM Diagnosis (3-, 4-, and 5-digit) and ICD-9-CM Procedure (3- and 4-digit) query types are provided by Optum Insight, Inc. Summary tables and the Query Tool are not currently set up for ICD-10-CM diagnoses and procedures.

Summary tables are stored locally by each Data Partner. Summary table queries (specified as SQL queries) are distributed using the secure Mini-Sentinel Query Tool, executed locally, and returned using the Query Tool software. A description of each summary table is provided here:

Enrollment Summary Table: Provides a count of unique members and days covered stratified by age group, sex, year, drug coverage status and medical coverage status. The count of unique members or days covered can be used as denominators to calculate crude prevalence rates.

Prevalent Summary Tables:

Prevalent ICD-9-CM Diagnosis Summary Table (3-Digit): Provides a count of unique members with a specific 3-digit diagnosis observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and 3-digit ICD-9-CM code.

Prevalent ICD-9-CM Diagnosis Summary Table (4-Digit): Provides a count of unique members with a specific 4-digit diagnosis observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and 4-digit ICD-9-CM code.

Prevalent ICD-9-CM Diagnosis Summary Table (5-Digit): Provides a count of unique members with a specific 5-digit diagnosis observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and 5-digit ICD-9-CM code.

Prevalent ICD-9-CM Procedure Summary Table (3-Digit): Provides a count of unique members with a specific 3-digit procedure observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and 3-digit ICD-9-CM code.

Prevalent ICD-9-CM Procedure Summary Table (4-Digit): Provides a count of unique members with a specific 4-digit procedure observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and 4-digit ICD-9-CM code.

Prevalent HCPCS Summary Table: Provides a count of unique members with a specific HCPCS code observed during the period and a count of events experienced within each stratum. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year, and HCPCS code.

Prevalent Generic Name Summary Table: Provides a count of unique members who had a drug dispensing during the period, a count of dispensing received by all of these members, and total days supplied by strata. Counts are stratified by generic drug name, age group, sex, quarter-year, and year.

Prevalent Drug Category Summary Table: Provides a count of unique members who had a drug dispensing during the period, a count of dispensing received by all of these members, and total days supplied by strata. Counts are stratified by drug category, age group, sex, quarter-year, and year.

Incident Summary Tables:

Incident ICD-9-CM Diagnosis Summary Table (3-Digit): Provides a count of unique members with a new specific 3-digit diagnosis observed during the period and a count of events experienced within each stratum. A new diagnosis was defined in three different ways: 1) the member has not had the diagnosis code in the prior 90 days, 2) the member has not had the diagnosis code in the prior 180 days, and 3)

the member has not had the diagnosis code in the prior 270 days. The counts are stratified by setting of visit (inpatient, outpatient, emergency department, any), age group, sex, year.

Incident Generic Name Summary Table: Provides a count of unique members who had a new drug dispensing during the period, a count of dispensing received by all of these members, and total days supplied by strata. New use was defined in three different ways: 1) the user does not have a dispensing of that particular drug in the prior 90 days, 2) the user does not have a dispensing of that particular drug in the prior 180 days, and 3) the user does not have a dispensing of that particular drug in the prior 270 days. Counts are stratified by generic drug name, age group, sex, quarter-year, and year.

Incident Drug Category Summary Table: Provides a count of unique members who had a new drug dispensing during the period, a count of dispensing received by all of these members, and total days supplied by strata. New use was defined in three different ways: 1) the user does not have a dispensing of that particular drug category in the prior 90 days, 2) the user does not have a dispensing of that particular drug category in the prior 180 days, and 3) the user does not have a dispensing of that particular drug category in the prior 270 days. Counts are stratified by drug category, age group, sex, quarter-year, and year.

2. Roles and Responsibilities

The Infrastructure Group is responsible for developing and maintaining the SAS programs used by Data Partners to create summary tables. Any time a revision is made to this program, usually as a result of an FDA requested enhancement or Data Partner suggestion for improvement, it is reviewed and tested in accordance with the Mini-Sentinel SAS Program Development SOP. This phase involves internal testing, beta-testing by several Data Partners, and iteration until the program is accepted as final.

MSOC staff send a summary table generation package to each Data Partner each time a data refresh is approved. The package includes SAS programs and lookup tables. Data Partners run the package and return their SAS logs to MSOC for review. Once the logs are reviewed and approved, MSOC staff send the Data Partner a standard set of 16 test queries. These test queries touch all 12 summary tables. Data Partners run the test queries, review the output, and upload results. Finally, MSOC staff examine test query results, follow-up with the Data Partner about any unexpected results, and approve when appropriate. Each Data Partner always has a set of summary tables ready and available for querying when query requests are made by members of the FDA.

The lookup tables included in the summary table package are kept up to date by MSOC programmers. They are lists of all NDCs, diagnosis codes, procedure codes, and HCPCS (provided by Ingenix, Inc.) and include a text description of each code. The most recent lookup tables are sent to Data Partners with the summary table package. Lookup tables provide a crosswalk between the code, which appears in each Data Partner's MSDD, and the description so that descriptions appear in the Query Tool.

FDA regularly submits summary table requests. The Query Fulfillment Group manager logs the request in the request tracker and assigns it an identification number and an analyst. This analyst works with the requester as needed to address any potential issues and finalize specifications for the request. Queries are sent to Data Partners and results are returned within two business days. The analyst then aggregates data from all Data Partners and drafts a summary report. This report is reviewed by the Query

Fulfillment Group manager and an epidemiologist, before being sent to the requester. MSOC staff are available to answer any questions about the report.

3. Summary Table Revisions

Aside from technical enhancements to the Query Tool described below, there were no revisions made to the summary table creation and request processes during Year Five as the MSOC concentrated on building and enhancing more complex querying tools such as Modular Programs and the PROMPT modules.

C. LESSONS LEARNED

1. Programming Tools

The focus of Year Five was major enhancement of existing programming tools, supporting documentation, and request forms.

During the first four years of the Mini-Sentinel, cohort identification and descriptive analysis programs (i.e., modular programs) were able to define events of interest (e.g., exposures, outcomes, inclusion/exclusion criteria) using a list of NDCs, diagnosis, or procedure codes. Feedback from workgroups and FDA indicated that the ability to specify complex algorithms was needed to properly define many HOIs. The challenge for the MSOC was to develop a reusable tool that could allow for specification of an unlimited number of complex algorithms without requiring de novo programming to code each specific HOI. Year Five marked the introduction of the Combo Tool, a programming tool that can define an event using any combination of NDCs, diagnosis and procedure codes, laboratory result values, enrollment periods, and encounter start and end dates and use them in combination and in relation to each other to define an event. This is an extraordinary enhancement to our suite of programming tools, essentially allowing for the specification of an unlimited number of complex HOI algorithms without ever requiring the time and resources to code each individual algorithm.

Year Five also marked the integration of several Year Four PROMPT-related analytic programming tools. The integration of these tools enhanced our routine queries by allowing for more complex analytic adjustment for confounding and the ability to perform sequential analysis. Supporting these new tools in distributed programs across our network of 18 Data Partners, in a production-like fashion, included supporting the programming language, Java, which we had only previously supported in small pilot projects. Providing support for Java, across multiple operating systems and Java versions using query request scenarios, was challenging. Although we were ultimately successful in this endeavor, we were reminded that software dependencies have the potential to create additional, sometimes unnecessary, challenges, which can result in analytic and reporting delays. The MSOC will make it a priority to create as few new software dependencies as possible, while still being able to maintain and enhance a high quality, nimble analytic environment.

As our programming tools increased in complexity in Year Five, it became evident that our supporting training materials and requester query request forms needed to be enhanced to support understanding of available tools and ease of use. Year Five saw the introduction of the MS Query Request Form, a single form used to specify needs for most routine query requests. The MS Query Request Form replaced MP 3, 6, and 9 query request forms, and allows requesters to specify needs for propensity

score matched analytic requests. The format and terminology of the form was revised to mimic a study protocol and to be clearer to investigators. The MSOC also created several PowerPoint slide presentations describing an overview of available analytic tools and new features. Presentations were given on Data Core and Data Partner calls, and posted on the FDA intranet site for reference.

Lastly, in Year Five, the MSOC began a long-term consolidation and integration of programming tools, combining modular programs 3, 6, and 9 into the CIDA tool and integrating CIDA with the propensity score matching tool. The purpose of consolidation is to decrease the time needed for implementation of new features, by reducing the number of programs that must be modified, go through the QC process, and be beta tested.

2. Summary Tables and Distributed Query Tool Software

The Mini-Sentinel Distributed Query Tool has proven very useful in quickly generating high-level information regarding exposures, diagnoses, procedures, and enrollment. Many requesters find Summary Table requests helpful in examining whether there are enough patients with a specific exposure to warrant a more in-depth analysis, such as a modular program request, a PROMPT request, or a workgroup dedicated to the examining the exposure. To date, the Query Tool has been used to issue over 300 summary table queries that generated information on over 1,200 drug exposures, diagnoses, and procedures.

In Year Five, there was a focused effort to expand Query Tool functionality beyond Summary Table queries to serve a broader spectrum of Mini-Sentinel's complex query fulfillment needs. The PopMedNet team worked closely with the MSOC to better understand the query fulfillment processes, workflows, and metrics. This analysis of Mini-Sentinel processes and Query Tool expansion will continue into Year Six.

VI. MINI-SENTINEL INFRASTRUCTURE

A. NAMING CONVENTION

The large volume of requests that the MSOC handles necessitated fine-tuning of request file naming in Year Five. A standardized approach was designed and implemented to assigning each data request a unique ID that allows for easy tracking and reporting for FDA, MSOC, and Data Partner staff. This is the backbone of a system that will be rolled out to simplify related Mini-Sentinel organizational systems such as hard drive folder structure, invoicing, and health outcomes algorithm library. A unique request ID is now made up of various tokens such as (1) project ID (made up of short names for the FDA/Mini-Sentinel task order and activities), (2) type of request or tool used (e.g., routine tools such as modular programs or ad hoc programs) , and (3) sequential run ID. For example, the Methods Development in Laboratory Data Workgroup part of Task Order 99 (Project ID: "**to99_meth_lab**") distributed a second data request (Sequential Run ID: "**wp002**") involving ad hoc programming (Type of Request: Ad Hoc -> "**ah**"), so a unique ID of "**to99_meth_lab_ahr_wp002**" was used to identify and label all materials related to this data request. Zip files containing data request packages and returned data are named with the request ID. The uniformity of the construction of request IDs allows programming code to easily parse out and use any of the pieces of information they contain, as needed.

B. COMMON COMPONENTS

During Year Five, the MSOC, in collaboration with the Data Partners, developed and beta tested Common Components, a SAS program which makes Data Partner site and ETL metadata available to Mini-Sentinel distributed SAS programs. The impetus for this initiative was a Year Four survey of Data Partner and MSOC staff soliciting ideas for improvements to the data request process. Common Components streamlines the work required by Data Partner's when executing distributed programs and provides assurance that programs are being run against the correct ETL version.

Rollout of Common Components began in May, 2014 and was integrated with the rollout of release of QA package version 3.2. In Year Six, all Mini-Sentinel requests will make use of Common Components. Once all Data Partners have successfully installed Common Components and run QA package v3.2 or higher, Modular Programs and Summary Table update programs will be integrated with Common Components. New workgroups will be trained to write their SAS programs to make use of Common Components, as well.

C. MINI-SENTINEL SECURE PORTAL

1. Function

To allow for secure electronic transmission of data and information between MSOC, FDA, workgroup/evaluation projects, Data Partners, and other Mini-Sentinel collaborators, MSOC implemented a secure portal accessible via secure web-based interface (i.e., using a web browser) or secure file transfer protocol (sFTP) software using usernames and strong passwords as well as a sophisticated system of group permissions and folder access. Any approved members of the Mini-Sentinel community can securely transfer documents in a section specifically assigned to them (or their group/organization).

During Year Five, MSOC implemented one major upgrade to the secure portal system to comply with the more stringent security requirements of the Transport Layer Security (TLS) 1.2 protocol. With this upgrade additional maintenance and administration tasks otherwise performed by external vendors have now been streamlined and are handled by MSOC staff (e.g., addition/deletion of users or groups, changing user/group permissions, creation of folders, creation of frequent reports with list of users for certain organization or groups).

2. Future Work

- Explore how the current secure portal application can be integrated into the existing Single-Sign-on and PopMedNet architecture to offer Mini-Sentinel users a one-stop-shop for secure communications and reduce maintenance and administrative burden (e.g., user and group authorization, audit).
- Additional security upgrades (as needed)

D. TESTING ENVIRONMENT AND SYNTHETIC DATA

1. Function

The Mini-Sentinel testing environment is a set of high-performance workstations hosted within the SOC with access to programming and editing applications (e.g., SAS, program editor, visual analytics, data processing and formatting) as well as a synthetic version of the MSDD with data for five million fictitious members spanning six calendar years. The workstations are used by SOC staff to develop, test, check for quality compliance, and validate SAS programs for infrastructure projects (e.g., modular programs, PROMPT, summary tables) or for workgroups.

During Year Five, the SOC Infrastructure and Programming Groups worked under the supervision of FDA with the Department of Health and Human Services staff to acquire license for additional SAS modules. These additional modules will enhance analytic capabilities (e.g., by using more efficient SAS procedures to implement rapid querying programs).

2. Future Work

- Enhance the pool of synthetic data
- Explore cloud environments so that more Mini-Sentinel developers have access to the testing environment.

E. MINI-SENTINEL DATA CATALOG V2: THE TASK ORDER MATRIX

The Mini-Sentinel Data Catalog (MSDC) v1 continued to be used during Year Five to track data requests and produce metrics about them. Much of the functionality of the MSDC is being transferred to the Query Tool, so that all Mini-Sentinel requests and related metadata can be centralized in that one system.

MSDC v2, called the Task Order Matrix was developed to support enforcement of the naming conventions discussed in **Section VI. A.** above. The Matrix allows the assignment of the short names used to represent FDA/Mini-Sentinel task orders and activities and uses them to build project IDs according to the convention. It also keeps track of task order start and end dates, and contract numbers. The current version of the Matrix also provides functionality for keeping track of workgroup participants, roles, and contact info.

F. MINI-SENTINEL DISTRIBUTED QUERY TOOL

1. Overview of Query Tool

Mini-Sentinel is one of the distributed networks that use a version of PopMedNet. The Mini-Sentinel tool, called the Mini-Sentinel Distributed Query Tool (Query Tool), is used to create and securely distribute data queries to Data Partners and enable Data Partners to review, execute, and securely return the results of those queries.

The Mini-Sentinel distributed network is hosted in a private cloud environment in a Federal Information Security Management Act of 2002 (FISMA)[xiv] compliant TIER III data center. The Query Tool is based on the PopMedNet™ software platform. The implementation design and architecture are detailed in the PopMedNet User's Guide and technical and security specifications can be found in PopMedNet Security Specifications Overview.

Query Tool architecture is consistent with the standards promulgated by the Standards and Interoperability (S&I) Framework supported by ONC. Mini-Sentinel staff works actively with the S&I Framework Query Health team and participated in the ONC Query Health Initiative as a pilot program. The pilot investigated the potential for including additional data sources on the Query Tool system[xv]. The selected data source was i2b2 (Informatics for Integrating Biology and the Bedside) – a widely used data repository and analysis platform. The PopMedNet team worked with Beth Israel Deaconess Medical Center, a clinical data partner with an existing i2b2 installation, to pilot end-to-end querying using the PopMedNet-i2b2 adapter. A demonstration video was created to show a successful query.

The Query Tool is designed to be flexible and configurable. For example, Data Partners can take advantage of multiple automation settings (e.g., set summary tables to run automatically when request is received), which can help to make the query response process more efficient. The Query tool also provides workflow decision points which allow users to review or reject requests and results and add comments. Data Partners review the query before it is run against their local data, review the results once the query is run, and have the option to send comments with the results they return. There are also configuration options (e.g., minimum cell count released, enable automatic notification of request received or results uploaded).

Multiple types of queries are available. Prior to Year Five, menu-driven queries could be run against the **Summary Table** data. In Year Five, additional menu-driven **query types** were added, including Request Metadata and Data Checking queries. Additionally, more complex queries, like **Modular Program** queries, that are built elsewhere are now distributed using the Query Tool.

The PopMedNet team continues to improve the processes involved in operation of the Query Tool. During Year Five, we implemented additional processes for managing the software development, testing, and software releases. New, more formal, processes for documenting system requirements will continue to be enhanced and standardized in Year Six. The software release process now includes more robust testing, including utilizing standard scripts for user acceptance testing.

The PopMedNet team is collaborating more closely this year with the MSOC to design and develop new features and functionality of the Query Tool. Additionally, as we make improvements to the PopMedNet software, we actively ensure that the system changes can be utilized across other networks using the same platform as appropriate. This allows for interoperability across PopMedNet networks, scalability, and efficient use of resources. For example, new reporting features developed for the new PCORnet distributed research network have been implemented and utilized by the Query Tool.

[xiv] National Institute of Standards and Technology, Computer Security Resource Center (CSRC). Federal Information Security Management Act (FISMA) Implementation Project. Available at: http://csrc.nist.gov/groups/SMA/fisma/index.html

[xv] Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, Murphy SN.
Query health: Standards-based, cross-platform population health surveillance. *Journal of the American Medical Informatics Association.* 2014; 21: 4, 650-656. Available at: http://jamia.bmj.com/content/early/2014/05/05/amiajnl-2014-002707

2. Network Implementation

The distributed querying network was established in partnership between the MSOC, Mini-Sentinel information technology vendor, and the Data Partners. The implementation process involves establishment of multiple "staging" networks that allowed for thorough testing of governance, security, performance, and querying capabilities of the software platform prior to deploying new releases to the Mini-Sentinel production environment. Standard systems development life cycle processes and procedures have been implemented and are continuously improved. Agile software development is used, allowing for frequent software releases.

The Query Tool software contains two separate but integrated components, a central web-based portal, and DataMart clients installed locally at each Data Partner. Having separate components means, new features, enhancements, and fixes to the portal, can be released without effecting Data Partners' local environments. We have limited releases requiring Data Partner installation of DataMart upgrades to twice a year, to address the barriers some Data Partners face when installing external software. A new web-based DataMart, that does not require local installation, has been developed and is undergoing testing and review of implementation options. We expect to start deploying the web-based DataMart in Year Six.

In Year Five, we continued to enhance the software documentation, including creation of an online wiki, which is continuously updated. Procedural changes have substantially improved the Query Tool development process including:

- Implementation of protocols for documenting system requirements
- Documentation of Mini-Sentinel use cases for the Query Tool
- Creation of user acceptance testing scripts
- Improvements to the software release process

3. Enhancements for Mini-Sentinel Query Tool Version 3.2 – 5

The Query Tool software platform undergoes ongoing improvements to better conform to software development standards, enable modularization of enhancements, improve scalability and extensibility, make the system easier to maintain, and simplify system modifications. Enhancements were also made, in Year Five, to better align our infrastructure with national querying standards described by ONC S&I Framework Query Health.

In Year Five, there were three major software releases (versions 3.2, 3.3, and 4.0). One release required Data Partners to install an updated version of the DataMart Client software. All three releases included enhancements to the portal, which provided additional features and functionality and bug fixes. Release notes can be found in PopMedNet Software Release Notes.

Each release allows for a broader and more efficient use of the Query Tool software by improving the system extensibility and technical architecture:

Maintainability: Ongoing platform upgrades are necessary to maintain the Query Tool and improve its efficiency and sustainability as system activities and requirements grow and use increases.

Workflow: New functionality has been designed and is under development to improve workflows. Several processes have been identified and developed for the Query Tool to allow for additional automation, auditing and tracking, and process improvements for developing new requests. For example, new workflow components have been developed that allow for additional user groups (e.g. MSOC project managers, FDA users, etc) to use the Query Tool. We expect to have this new functionality released and these new user groups trained and using the Query Tool for multiple uses in Year Six.

Scalability: The Query Tool can cultivate and support new networks, projects and users. With the architecture improvements and workflow engine functionality, the Query Tool can now support additional customized workflows for different user groups and request types (e.g., develop a customized user interface for specific QA activities). Additionally, the Query Tool infrastructure was improved to support increases in both volume and complexity of new users, Data Partners, data sources, query activity, and metadata capture without performance concerns or database constraints.

Extensibility: The plug-in design allows for development of new features that can be added without impacting other parts of the system.

Major features, fixes, and enhancements designed and developed in Year Five are outlined below:

General Infrastructure:

- Improved usability of secure portal global access controls and related security
- Improved secure portal security architecture to enable scalability
- Improved the email notification functionality (e.g., improve reliability/monitoring of the email service, create the ability to send email to specific users when requests are assigned or needing review,)

Web-based Portal:

- Performance and Stability improvements
  - Infrastructure improvements increased the usability of access controls, which increases system performance.
  - The existing technical architecture for menu-based queries has been significantly improved, with a new Question Engine framework that allows for efficient scalability and extension of the menu-based Query Tool interface to new query types and data sources.
- Network administration:
  - Project-based request processing and administration: Groups, organizations, DataMarts, request types, and users can be organized into a project. Access controls can be enforced by project.
  - Strengthened password selection criteria at registration
  - Email notification functionality added: Enables Network Administrators to send custom messages to the entire network. Users receive an email message, a copy of which can be viewed in the new "Messages" panel on the Portal home page.
  - Organizations, DataMarts, and Projects and their associated access controls can be copied to create a new entity.
  - Network Administrators can revoke user access to the Portal and DataMart Client while retaining revoked user profiles and history.
- Documentation:

  - Enhancements to the Log-in and Resources pages with additional content, including links to online resources specific to the Mini-Sentinel Query Tool.
  - Tip text added throughout the Portal with links to relevant wiki pages.
- Creating requests:
  - Investigators can define the task order, activity, and project associated with their requests
  - Bugs fixed related to viewing of draft requests
- Reviewing results:
  - Comments entered by Data Partner are available to the investigator in the web-based portal
  - Fixed a bug which truncated the decimals places displayed for summary table results on the Portal and in exported files.
- Organization and DataMart metadata capture: Enables organizations to describe their organization and DataMarts with a finer level of detail.
- Searching and Reporting:
  - Request Search: Allows users to query previously entered request metadata (e.g., task order, activity, activity project, requestor, date ranges, status of request). Detailed and summary views are available for viewing on the portal and for export.
  - System Search: Enables users to search system metadata (e.g., for organizations, DataMarts, and registries). Detailed and summary views are available for viewing on the portal and for exporting.
  - Network Activity Report: Generates a summary of the number and types of requests submitted within a network. Users may select specific projects or time periods for reporting.
- Single Sign On: A single sign on page was created for Mini-Sentinel web-based applications, which allows a user to sign in to all available applications through a secure landing page and gain access. MSOC is piloting this feature with single sign on for the Query Tool and new Survey Tool, with an additional sign on required to access the Mini-Sentinel Secure Portal.

DataMart Client:

- Performance and Stability improvements : The new Question Engine framework described above involves changes to the DataMart Client infrastructure which allow for improved query processing. The DataMart Client was also modified and will be released in Year Six to ensure greater compatibility with multiple versions of the software. This is important, as not all Data Partners upgrade the DataMart software at the same time. User interface: improved usability, including page scrolling options
- DataMart Client proxy server compatibility added: This fixes a bug preventing the DataMart from connecting to a data source that uses an HTTPS connection to a proxy server.
- MS SQL Server compatibility: Summary table changes were made to allow the DataMart Client to connect to data stored in Microsoft SQL Server, in addition to Microsoft Access, which was previously the only option. Improvements to DataMart Client exception handling, including log reports
- Delete files from DataMart Client response: Files may now be deleted from the DataMart Client after they have been added in response to request, but before the response is uploaded to the Portal.
- New Web-based DataMart Client developed: The web-based DataMart Client (WDMC) provides a web-based platform for administering DataMarts. The features and functions of the existing desktop DataMart Client are provided while simplifying support, maintenance, and deployment. The WDMC includes a user interface, application services, DataMart Client web service client,

request service web API, request model adapter service, network connection manager, and data service. These features can be configured by Data Partners to suit their needs and security requirements.

Query Requests:

- Mini-Sentinel Query Interface (Question Engine Enhancements)
  - Designed specifications to adapt the existing Question Engine (Query Composer) to create inputs for SAS programs.
  - Designed new functionality that provides the Query Tool with information about the MSCDM and allows for reuse of existing queries
  - Made additional enhancements to the network activity and audit reports
- SQL Distribution: Allows users to submit raw SQL code to data partners.
- Menu-based queries:
  - Data Checking Request: A new data characterization query, allows the MSOC to quickly assess data availability and characteristics across Data Partners of race, ethnicity, diagnosis codes, and procedure codes and availability of NDC codes. New functionality has been designed to further enhance the data characterization queries, including adding additional access control permissions, which will allow for broader use of this analytic tool.
  - Searching and exporting of queries distributed during user-specified time periods: This query allows for easier tracking of query requests (counts and details) by type of query, status, workplan, task order and activity, and requester.
- Modular Program Distribution: Enables users to distribute modular program packages and define the metadata about the request, including the task order, project, and description of the request. Files can be added to the request from the user's local computer or from an external sFTP site. Request metadata from signature files is also now imported and populated into the Query Tool Portal and can be used for reporting purposes.
- Data Checking Request: Enables users to issue requests against the tables created by the MSOC's data QA process. The output includes a variety of charts, graphs, and reports to illustrate the data characterization. In Year Five, new functionality has been identified and designed to allow for additional access control levels, allowing additional users, including FDA and workgroups. These enhancements are expected to be developed in Year Six.
- Query request workflow
  - Developed the specifications for an end-to-end web-based query request workflow that allows input and monitoring of all query request steps, from query initiation to query package approval to query response.
  - Provided additional Query Tool access to the groups that already had access (i.e., the MSOC) and prepared for adding the groups that previously had no access (i.e., FDA and workgroups). New user groups will have the ability to use the Query Tool in Year Six.
  - Documented the requirements for creating additional efficiencies to the request development process, such as the ability to easily triage and track tasks and requests assigned to users.

4. Deploying Platform Enhancements to the Data Partners

Query Tool software releases involve several steps starting with requirement gathering sessions and resulting in functional and technical documentation. Design, development, and testing phases are a

collaboration between MSOC and sub-contractor software development team. The two teams hold their own regular weekly meetings as well as participating on joint meetings at least twice a month to review project status, conduct demonstrations of features under development, and plan for upcoming work and release schedules. This level of collaboration allows for standardized software development processes which ensure successful software releases of the Query Tool with minimal impact to users.

The MSOC and the software development team provide ongoing support as new users are added, questions arise, and enhancements are requested and developed. All software upgrades and revisions are accompanied by Release Notes to inform Data Partners of the changes. All 18 Mini-Sentinel Data Partners currently use the Mini-Sentinel Distributed Query Tool.

G. CODE LOOKUP TOOL

All Mini-Sentinel data requests involve defining medical events (e.g., exposures, outcomes, covariates, preexisting conditions) using lists of medical codes of various types (e.g., NDC drugs, ICD-9-CM diagnoses and procedures, HCPCS procedures). These codes are pulled from resource tables using SAS programs. Code lists are created frequently and it has been a time-consuming, mostly manual process. By the end of Year Five, specifications were written, a web-based Code Lookup Tool (CLT) built, and beta testing begun. The CLT is a web-based database application designed to allow users to search any code lookup resource of any type and easily build comprehensive code lists needed for data requests. The first version of the tool allows for the building of lists of drug codes only based on one resource table. Future versions will allow for the building of other types of code lists (e.g., procedures, diagnoses) from multiple sources.

H. ALGORITHM LOOKUP TOOL

Mini-Sentinel activities regularly involve developing, reviewing, and/or utilizing algorithms for identification of specific cohorts, health outcomes of interest (HOIs), and confounders. These algorithms need to be catalogued in a way that allows users easy access to information about where and how they have been used in past activities and how to use them in subsequent activities. In Year Five, specifications were written and a prototype developed for the Algorithm Lookup Tool (ALT). The ALT was designed as a flexible, expandable web-based database with user-friendly data storage, search, and reporting functionality.

I. DATA REVIEW TOOL

The Data Review Tool (DaRT) is an Excel workbook comprised of VBA macros that was developed as an internal tool for doing QA. The MSOC Data QA is a complex process done for each Data Partner's refresh. It involves vetting over 200 SAS datasets generated by the QA program package against a list of error codes and producing a report for the Data Partner analyst. Prior to the introduction of DaRT, this was largely a manual process. The tool consolidates resources, datasets, and a "master error checklist" into one toolkit to allow the MSOC analysts easy access to all components needed to perform a systematic QA review.

J. MSDD HUB

The MSDD Hub is an Excel workbook that organizes and consolidates metadata about the MSDD (e.g., ETL number, QA program version used, and status of refresh process for each refresh; number of PatIDs and date range of data available by table for each Data Partner) into a single tool. An interactive table of contents enables navigation of topics. Easier access to information about the MSDD facilitates communication about MSDD related issues.

K. AUTOMATED REPORTING TOOL

The automated reporting tool continued to be used heavily in Year Five. It was used to produce every report provided to the FDA for an MP3 request. Compared to the previous manual method it requires no programmer time and substantially less analyst time to produce a report. Modifications were made to the tool this year to keep it in sync with changes to MP3.

L. SIGNATURE FILE

Each run of a modular program since Year Two and each run of a quality assurance program since Year Four has output a "signature" file containing metadata needed to track requests and understand the environment in which the request was executed. MPs output metadata about the request (e.g., names of input files, date range of data used, age groups queried, execution time). QA programs output metadata about the environment the program was run in (e.g., SAS version, operating system).

In Year Five, the Infrastructure Group added the IDs required by the new **naming convention** to MP and QA signature files. We also added metrics (e.g. counts of scenarios by MP) and environmental metadata (e.g. SAS version, operating system, SAS products licensed) to the MP signature file. The critical variable added to the QA signature file was the QA program version.

In addition to producing a run-level signature file for each separate macro call during MP execution, a request-level signature file is now output that sums calculations from the individual runs. The QA program outputs a request-level signature file that sums calculations from the program runs against individual MSDD tables.

M. LOG CHECKER

This SAS tool scans SAS logs for errors, warnings and notes of interest, and produces summary and detailed reports of what is found. It exists as a stand-alone tool and can also be integrated into any distributed Mini-Sentinel production SAS program. Specifications for the tool were written in Year Five.

N. PERSONAL HEALTH INFORMATION CHECKER

This add-on module automates the checking of SAS logs and output SAS datasets for indications of the presence of personal health information (PHI). Examples of PHI for which the tool checks include: names of variables that normally contain PHI (e.g., PatID, EncounterID, ProviderID, date, dt, DOB, variables specified by the user) and particular variable formats (e.g., date format). This tool is based on the one developed by the Virtual Data Warehouse (VDW) distributed data network (Bredfeldt 2013).[8] Specifications for the tool were written in Year Five.

O. REPLACE INDIVIDUAL IDENTIFIERS WITH RANDOM IDENTIFIERS

Some studies, whether for chart review preparation or statistical methods development, have need for individual-level, de-identified datasets to be returned. This tool facilitates the replacement of individual-level identifiers with random identifiers and generates a file, which remains with the data source, that contains a mapping of original identifiers to the random identifiers. Specifications for the tool were written in Year Five and programming was completed. The tool will be released in Year Six.

P. ZIP RESULT UTILITY

This tool produces a zip archive of distributed program files sent by Data Partners that ensures a consistent structure and naming of files returned to MSOC. Specifications for the tool are under development.

Q. CIDA RESULTS INTEGRITY CHECKER

This tool provides integrity checks across Cohort Identification and Descriptive Analytic (CIDA) output datasets within a request, in order to ensure that output generated is consistent across tables. Specifications for the tool are under deveoplment.

R. LESSONS LEARNED

In Year Five, there were multiple successful Query Tool software releases, however, there continue to be limitations on what and when we can implement due to the reliance on locally-installed DataMart Clients. Moving to a web-based DataMart, in Year Six, will alleviate the difficulties that some Data Partners experience downloading the current DataMart Client. This will make it possible to make more frequent system enhancements.

In Year Five, there was a focused effort to expand Query Tool functionality to serve a broader spectrum of Mini-Sentinel's complex query fulfillment needs. The PopMedNet team worked closely with MSOC staff to better understand the query fulfillment processes, workflows, and metrics. This collaborative analysis of Mini-Sentinel processes and Query Tool expansion will continue into Year Six.

VII. OTHER DATA CORE ACTIVITIES

A. COMMUNICATIONS

The MS Scientific Operation Center holds a regular meeting with Data Partners to maintain contact with them and to facilitate communication among organizations. The Year Four format of a monthly web conference continued throughout Year Five. Examples of the Year Five presentation topics included Signature File Enhancements, ETL Versioning, Mini-Sentinel on the Web and in the News, PopMedNet Tutorial, QA Process Changes and Enhancements, National Death Index (NDI) Linkage Project. In addition to these regularly scheduled meetings, the MSOC regularly communicates with Data Partners by email, phone, and teleconference to address questions as they arise.

B. SUPPORT TO WORKGROUPS

The MSOC continued to expand its work with workgroups. The MSOC helps ensure that workgroups utilize the MSDD effectively, efficiently, and properly. MS Scientific Operations Center members actively participate during workgroup meetings and are available by email and phone, if needed. In Year Five, the MSOC continued to advise workgroups on the use of Modular Program and Summary Table queries in feasibility studies. In Year Five, the MSOC completed 6 Modular Program requests and 1 Summary Table Request in support of workgroup activities. New in Year Five, MSOC members provided support to workgroups using the new prospective surveillance tools (PROMPT). MSOC members helped workgroups specify, package, test, distribute and compile results for two workgroups.

The MSOC reviews all workgroup plans to ensure that sensitive information is appropriately protected. The MSOC also maintains a secure system used to communicate sensitive information with Mini-Sentinel Collaborators. This system has been designed to be compatible with Mini-Sentinel Collaborators to continually facilitate data exchange.

C. DISSEMINATION ACTIVITIES

The success of Mini-Sentinel has led to many requests for information, requests for presentations, and other inquiries to describe how Mini-Sentinel works. Many of the questions about Mini-Sentinel are addressed on the Mini-Sentinel website and information seekers are directed to the appropriate webpage. Requests for Mini-Sentinel staff to present at professional meetings or other public venues are typically handled by the Data Core co-leads, the Director of Scientific Operations, and the Mini-Sentinel Principal Investigator.

1. Manuscripts

A complete list of manuscript and presentations is available in the Publications and Presentations section of the Mini-Sentinel website.

2. Meeting Presentations

Table 4 includes a list of key presentations related to the Mini-Sentinel Scientific Operations Center during Year Five.

Table 4. Meetings and Presentations

Date	Venue	Presentation Title	Presenter(s)
09/09/2013	2013 Northeast SAS Users Group (NESUG) Annual Conference	Continuous Enrollment Requirements in Epidemiologic Studies	Jennifer Popovic
10/17/2013	International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Annual International Meeting, New Orleans, LA	Distributed Research Networks and Applications in Safety and Outcomes Research	Kevin Haynes

Date	Venue	Presentation Title	Presenter(s)
11/23/2013	Monte Jade Science and Technology Conference, Boston, MA	Big Data For Medical Product Safety Surveillance	Darren Toh
12/10/2013	Mini-Sentinel Data and Programming Summit Boston, MA	Mini-Sentinel Programmers Annual Meeting Introduction	Richard Platt
		FDA's Mini-Sentinel Program to Evaluate the Safety of Marketed Medical Products	Jeff Brown
		Mini-Sentinel Data Group: Overview of Year Five Planned Activities	Nicolas Beaulieu
		Mini-Sentinel Modular Programs: Enhancements to Query Laboratory Result Values	April Duddy
		MSCDM Expansion Update	Lesley Curtis
		SAS Programming Standard Operating Procedures	Jen Popovic
		Mini-Sentinel SAS Programming Tools	Jen Popovic
		Common Components/Program Header/ Meta Data	Malcom Rucker
1/14/2014	Brookings Sentinel Initiative Public Workshop, Washington, DC	Prospective Routine Observational Monitoring Program Tools (PROMPT): Current Status of Development	Azadeh Shoaibi
		Prospective Surveillance of Anti-Diabetes Drugs and Acute Myocardial Infarction	Bruce Fireman
		Developing a PROMPT Surveillance Plan	Elizabeth Chrischilles
		Sentinel Prototype	Janet Woodcock
		Sentinel Initiative Public Workshop Panel: Overview of Enhancements Underway to Mini-Sentinel	Jeff Brown, Lesley Curtis
		Risk of Intussusception after Rotavirus Vaccination: Results of the Mini-Sentinel/PRISM Study	W. Katherine Yih
		CDER Use of Mini-Sentinel (MS) Tools, Approaches, and Resources in Analyses of Post-Market Drug Safety	Marsha Reichman
		State of CBER's Mini-Sentinel Activities	Michael Nguyen

Date	Venue	Presentation Title	Presenter(s)
		FDA's Mini-Sentinel Program to Evaluate the Safety of Marketed Medical Products: A Look Back, a Look Ahead	Richard Platt
1/28/2014	IVIG and Hemolysis Public Workshop	Evaluation of Immune Globulin and Hemolysis at FDA	Scott K. Winiecki
2/28/2014	Mini-Sentinel Planning Board Meeting	PCORnet: The National Patient-Centered Clinical Research Network	Rich Platt
3/03/2014	Data Partner Meeting	Mini-Sentinel on the Web and in the News	Tiffany Woodworth
3/20/2014	FDA Webinar	Adjusting for Time-Varying Confounding and Selection Bias in Longitudinal Studies	Darren Toh
3/25/2014	SAS Global Forum (SGF) San Francisco, CA	Programming in a Distributed Data Network Environment: A Perspective from the Mini-Sentinel Pilot Project	Jen Popovic
3/28/2014	Mini- Sentinel Workgroup Presentation	Metabolic Effects of Second Generation Antipsychotics in Youth (APY)	Tobias Gerhard, Marsha Raebel
4/03/2014	HMO Research Network Conference	Mini-Sentinel Data Quality Review and Characterization: Processes, Tools, and Future Directions	April Duddy
4/11/2014	Bordeaux Pharmacoepi Festival, Bordeaux, France	Prospective Surveillance of Newly Approved Medical Products: The U.S. Mini-Sentinel Experience	Darren Toh
4/17/2014	FDA Webinar	New Sequential Methods in a Distributed Data Setting	Andrea Cook
5/15/2014	FDA Webinar	Supplemental Information to Improve Confounder Adjustment	Sascha Dublin
6/02/2014	Data Partner Meeting	Mini-Sentinel Programming Tools: Combo Tool Functionality	April Duddy
6/19/2014	FDA Webinar	Medical Countermeasures	Matthew Daley, Gretchen Weiss, Arthur Davidson, Melissa McClung
7/17/2014	FDA Webinar	16 Health Outcomes of Interest for Surveillance Preparedness	Sean Hennessy, Charlie Leonard, Cristin Feeman
7/25/2014	Mini-Sentinel Planning Board Meeting	Assessment of febrile seizures after trivalent inactivated influenza vaccines during the 2010-2011 influenza season in PRISM Modular Program Level 1 Requests: Using the Cohort Identification and Descriptive Analysis (CIDA) Tool	Alison Tse Kawai April Duddy
7/29/2014	The Brookings Institution Webinar	Linking Data from Public Health Medical Countermeasure Campaigns with Electronic Health Records: The Mini-Sentinel Medical Countermeasure Post-marketing	Arthur J. Davidson, Matthew F. Daley

Date	Venue	Presentation Title	Presenter(s)
		Surveillance Project	
9/18/2014	FDA Webinar	Methods for Improving Adjustment for Confounding	Stan Xu, Susan Shetterly

D. ENGAGING WITH OTHER NATIONAL INITIATIVES

The MSOC is actively engaged in a number of national and international initiatives related to distributed networks using observational data to generate health care evidence and the standards needed to create such networks. The NIH Health Care Systems Collaboratory Distributed Research Network and PCORnet are two emerging networks designed to facilitate distributed querying of health care information to generate evidence. MSOC staff and leadership are actively involved in both networks, and are working to share information across networks. For instance, the NIH Health Care Systems Collaboratory Distributed Research Network includes several Mini-Sentinel Data Partners and the network has used publically available Modular Programs for distributed querying. The PCORI-funded PCORnet network common data model is based on the MSCDM and plans to leverage Mini-Sentinel tools to facilitate distributed querying. In addition, some lessons learned from development of the PCORnet common data model will be used to improve the MSCDM. For example, the PCORnet common data model extended the MSCDM by adding several data elements (e.g., raw values for several variables) and expanding data element value sets (e.g., additional options for missing values based on HL7 standards); these changes will be adopted by Mini-Sentinel during Year Six to improve the MSCDM and maintain consistency between the two models.

In addition, other initiatives that MSOC investigators and staff are collaborating on include: 1) a PCORI grant to develop a standards-based approach for data quality assessment; 2) the Agency for Healthcare Research and Quality-funded Electronic Data Methods Forum on issues related to data quality, governance and metadata standards for distributed querying; 3) the Learning Health System Initiatives at the University of Michigan to develop standards for a learning health system; 4) the M.I.T. New Drug Development Paradigms initiative to help stakeholder understand the potential uses of observational data; 5) the ONC Structure Data Capture initiative; and 6) the Reagan-Udall Innovation in Medical Evidence Development and Surveillance (IMEDS) program.

MSOC leadership are routinely invited to lecture at national and international meetings on all aspects of the Mini-Sentinel project.

VIII. MSDD QUERY REQUEST SUMMARY

A. MODULAR PROGRAMS

A total of 69 Modular Program requests were initiated in Year Five. Of these, 51 were completed as of September 22, 2014. Of these 51 completed requests: CDER initiated 33 requests; CBER 11 requests; CDRH 1 request; and workgroups 6 requests (**Table 5**). MP3 was used in 30 requests, MP4 in 2 requests, MP6 in 5 requests, MP7 in 2 requests, MP8 in 2 requests, and MP9 in 7 requests (**Table 6**). At the end of

Year Five, the newly available CIDA/propensity-score matching tool was used in 3 requests. The 51 completed requests involved between 3 and 108 modular program scenarios each for a total of 1,496. A scenario is defined as a unique set of query parameters. Modular programs allow for multiple scenarios to be run by Data Partners within a single request. Though it is possible to run any number of scenarios with one execution of a modular program, effectively communicating the large amount of data returned for numerous scenarios may require more than one report. The 51 completed requests generated 63 reports.

The complexity of the requests varied from a straightforward MP9 request using one run to analyze prevalent and incident drug use, to a complex request with pre-existing conditions and several outcome events using MP3 and the combo tool. For example, one request consisted of 24 scenarios to assess several events among drug users with a pre-existing condition and prior use of several drug products. The combo tool was used with a pre-existing conditions file to assemble a cohort of members with: 1) a pre-existing diagnosis; and 2) prior use of Drug A; and 3) prior use of Drug B; and 4) no evidence of prior use of Drug C. Prior to development of the combo tool, this analysis would have required de novo programming.

Table 5. Number of Modular Program Requests, Scenarios, and Reports by Requester in Year Five (September 23, 2013 to September 22, 2014)

Center/ Requester	Number of Requests Initiated	Number of Requests Completed	Number of Scenarios Completed	Number of Reports Completed
CDER	44	33	974	45
CBER	12	11	255	11
CDRH	1	1	52	1
Workgroups	12	6	215	6
Total	69	51	1,496	63

Table 6. Number of Completed Modular Program Requests and Scenarios by Modular Program in Year Five (September 23, 2013 to September 22, 2014)

Modular Program (MP)	Number of Requests	Number of Scenarios
MP3	30	1,045
MP4	2	28
MP6	5	126
MP7	2	17
MP8	2	104
MP9	7	95
Other*	3	81
TOTAL	51	1,496

*These include requests that used newly-developed CIDA/propensity score matching code, andwere distributed and executed as part of the Modular Program set of activities.

Data Partners typically have five business days to complete requests. However, MSOC occasionally distributed multiple requests concurrently but staggered the due dates to keep consistent with Data Partners' workload expectations. Of the 51 requests, 14 were completed on time by all Data Partners. Of the 37 remaining requests, the average number of days to completion past the due date was 7 and the median was 5 (including weekends and holidays). Overall, response time by Data Partners was within expectations.

All reports were created in Microsoft Excel® and typically included tables and figures of counts and rates both aggregated and stratified by sex, age, and year. The reports also included an overview describing their contents, a glossary, and specifications. Depending on the MP, parameters, and codes used, a report may have contained incident and prevalent data on drug use, diagnoses, and procedure use (e.g., number of users/patients, dispensings, diagnoses, procedures, total days supplied, eligible members (denominator), member days, users per eligible members, dispensings per user, days supplied per user, and days supplied per dispensing as well as events, days at risk, and events per days at risk (for MP3)). Additionally, reports presented the percent contribution of each Data Partner to the total as well as the percent within each Data Partner the number of users, dispensings, days supplied, eligible members,

member days as well as events and days at risk (for certain reports). Code lists and other content were included when appropriate.

The average time from receipt of all data to report submission was 8 days and the median time was 6 days (including weekends and holidays). The increasing use of modular programs has given requesters more experience with the capabilities of the programs, and in turn generated more complex requests. Complex requests usually require additional consultation with FDA regarding specifications, more "scenarios" and more data received from the Partners, and more complicated and/or number of reports. Additionally, some requests required investigation and revision of errors or unexpected data in the output at one or more of the 18 Data Partners, and prioritization of other requests and activities.

B. SUMMARY TABLES AND QUERY TOOL

A total of 47 summary table queries were performed to respond to 17 requests during Year Five (Table 7). Multiple queries are sent per request for a number of reasons. First, it can happen when the request examines codes that fall into more than one query type and/or care setting. For example, a single request could examine metformin HCL use along with diabetes diagnoses (ICD-9-CM diagnosis code 250). One query would be sent on metformin HCL while a second query would be sent on diabetes. Second, if the requester would like to examine diabetes diagnoses in more than one care setting (for example, outpatient, and inpatient), then a separate query would have to be sent for each care setting. Finally, if the requester wishes to examine both prevalence and incidence of a generic name, drug class, or three-digit diagnosis code, two different queries would have to be sent out—one for prevalence and one for incidence.

The 47 queries performed included 221 scenarios (code-care setting combinations, code-incidence/prevalence combinations, or drug product/class combinations), each stratified by age group, sex, and year. CDER initiated 14 requests, while CBER, the Intravenous Iron Workgroup, and PRISM submitted one request each.

Data Partners were typically given two business days to complete each query, and responded within the allotted time for the majority of queries. During Year Five, a Query Tool upgrade resulted in minor technical difficulties and a slight delay in responding to three of Year Five's requests at a few of the smaller Data Partner sites. Two rounds of reports for these queries were submitted to FDA—one before those sites were able to respond, and one which included results from those sites. In addition, MSOC occasionally waited for a Data Partner to update its data before distributing a particular request, especially if the request was for more recent data.

Sixteen out of the 17 total requests generated summary table reports during Year Five, for a total of 36 reports (**Table 7**). Most requests involved more than one 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. If a request involved both prevalence and incidence queries, a separate report was generated for each. For generic name queries and drug class queries, reports displayed counts of users, prevalence or incidence rates (users per 1,000 enrollees), days supplied per user, dispensings per user, and days supplied per dispensing. For diagnosis and procedure queries, reports displayed counts of patients, prevalence or incidence rates (patients per 1,000 enrollees), and the number of events per patient. All reports were

created in Microsoft Excel and included both pivot tables and figures along with an overview describing the tables and figures presented.

Table 7. Number of Summary Table Query Requests in Year Five (September 23, 2013, to September 22, 2014), by Requester

Center/ Requester	Number of Requests Initiated (Broad Categories)	Number of Requests Completed (Broad Categories)	Number of Queries Completed	Number of Code-Setting Combinations, or Number of Drug Combinations Completed*	Number of Completed Requests Involving Reports	Number of Reports Completed
CDER	14	14	33	109	13	30
CBER	1	1	8	40	1	2
IV Iron WG	1	1	3	22	1	3
PRISM	1	3	3	50	1	1
TOTAL	**17**	**19**	**47**	**221**	**16**	**36**

*For generic names, drug classes, and three-digit diagnosis codes, prevalent and incident counts are queried separately. If prevalent and incident use of Drug X were both queried, for example, that would add two to the count in this column.

Table 8 displays the number of queries completed during Year Five stratified by requester and query type. Diagnosis Code queries (86 in total), Generic name queries (88), and HCPCS queries (47) accounted for the bulk of activity.

Table 8. Number of Summary Table Queries Completed in Year Five (September 23, 2013, to September 22, 2014), 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
CDER	---	80	---	2	14	---	---	---	13	109
CBER	---	---	---	---	---	20	---	---	20	40
IV Iron WG	---	8	---	---	---	---	---	---	14	22
PRISM	---	---	---	2	16	32	---	---	---	50
TOTAL	**0**	**88**	**0**	**4**	**30**	**52**	**0**	**0**	**47**	**221**

C. AD HOC REQUESTS

Ad hoc requests are requests that cannot be addressed using existing tools. Additional work in the form of de novo programming is needed to fulfill the requirements of such requests. De novo programming must adhere to the SAS Program Development SOP that requires: 1) a formal specification of the program requirements; 2) MSOC development and testing; 3) quality compliance checks by independent, third party programmers; and 4) beta-testing by at least two Data Partners. Once this process is complete, the program is released by the MSOC for use.

The one de novo programming activity, in Year Five, was to complete a CBER request to determine the concordance between dates recorded in the MSDD and patient charts from a prior MP request and chart abstraction. The programming code pulled exposure and outcome event dates in the patient level files kept behind the Data Partners' firewalls. The dates were then compared to the dates recorded during the chart abstraction.

D. LESSONS LEARNED

1. Modular Programs

With the addition of several new modules and options to increase the flexibility of Modular Programs comes increased complexity and potential for introducing unintended effects on output and errors. The MSOC expects request development time to increase for more complicated requests to allow for additional review, development, and testing. This may include additional conference calls with requestors and Modular Program programmers to ensure request packages will answer the requestors' questions. To better handle this increased complexity and demand, the MSOC is continuing to expand internal capacity by training more personnel to develop requests. The MSOC is currently developing a more structured training program to facilitate onboarding of new personnel.

The MSOC has continued to review reference code databases and processes for pulling target codes throughout the Mini-Sentinel Project. The MSOC plans to incorporate additional reference code databases to ensure complete capture of drug, procedure, and diagnosis codes. The distinct code references will be cross-checked and consulted when developing code lists for modular programs and other requests. In addition, the MSOC will continue to use 9 digit NDCs as this allows for more sensitive identification of drug exposures. All codes are checked for duplication and MSOC will revert to 11 digit codes when discordant drug name and NDC duplicates are discovered. MSOC is also developing a library of code sets used to define exposures and outcomes for reference in developing future requests. This is particularly useful when developing covariate files for the propensity score matching module.

2. Summary Tables

In Year Five, the MSOC continued to improve the reports summarizing results both in terms of the information contained in the tables and figures and formatting for posting to the Mini-Sentinel website. The reports were significantly simplified to only display relevant information needed by requesters. Report templates have been created with these improvements and formatting changes so that report development is more efficient for new requests.

IX. POSTINGS TO MINI-SENTINEL WEBSITE

During Year Five, the MSOC continued posting reports generated from summary table and modular program requests to the Mini-Sentinel website. These reports, completed during Years Two, Three, Four, and Five, were approved for posting by FDA. No Data Partner-specific results are included in posted reports. In Year Five, a total of 43 reports were posted. All 43 reports appear in the "Assessments" tab on the website: 17 under the sub-tab "Exposures to Medical Products"; 19 under the sub-tab "Diagnoses and Medical Procedures"; and 7 under the sub-tab "Health Outcomes Among Individuals Exposed to Medical Products". The titles of the reports are shown below.

A. REPORTS

1. Summary Table Reports Under "Assessments: Exposures to Medical Products"

- Enoxaparin sodium use
- Injection enoxaparin sodium use

- Anti-seizure medication use
- Occurrence of selected generic drugs 11
- Multiple sclerosis medication use
- Multiple sclerosis medication injections
- Acetazolamide use
- Occurrence of selected biological generic drug products
- Propylthiouracil and methimazole use 2
- Occurrence of selected generic drugs 9
- Injection gammagard use
- Immune globulin use
- Analgesic use 3

2. Modular Program Reports Under "Assessments: Exposures to Medical Products"

- Quinine sulfate use
- Oral antifungal use
- Occurrence of selected generic drugs 10
- Antiepileptic drug use

3. Summary Table Reports Under "Assessments: Diagnoses and Medical Procedures"

- Acute kidney failure diagnoses
- Hemodialysis procedures
- Occurrence of selected dental HCPCS codes
- Cataract procedures
- Cholecystectomy procedures
- Herniorrhaphy procedures
- Myringotomy procedures
- Tonsil/adenoid removal procedures
- Nursing home facility HCPCS codes
- Occurrence of selected biological product HCPCS codes
- Unspecified anterior pituitary hyperfunction and hypertrophy of the breast diagnoses
- HIV and HBV infection diagnoses
- Bupropion HCL sustained release HCPCS codes
- Hemorrhagic disorder diagnoses

4. Modular Program Reports Under "Assessments: Diagnoses and Medical Procedures"

- Occurrence of kidney stones
- Chronic Kidney Disease diagnoses
- Nursing home HCPCS codes
- Intravenous immunoglobulin (IVIG) procedures
- Diabetes diagnoses and occurrence of laboratory results

5. Modular Program Reports Under "Assessments: Health Outcomes among Individuals Exposed to Medical Products"

- Pneumococcal conjugate vaccines & Kawasaki's disease
- Quinine sulfate, diltiazem & selected thrombotic events
- Antiepileptic drugs (AEDs) & kidney stones
- Immunoglobulin & hemolysis 1
- Immunoglobulin & hemolysis 2
- Immunoglobulin & hemolysis 3
- Anti-D, dexamethasone, prednisone, romiplostim & hemolysis

The MSOC is working with FDA to post the remainder of the reports that have been created for summary table and modular program requests following approval for posting by FDA, and will continue to work with FDA to post reports as new requests are completed and new reports are created.

B. OTHER POSTINGS

1. Mini-Sentinel Data Core Modular Programs

During Year Five, the MSOC posted to the Mini-Sentinel website revised versions of documentation and code for Modular Programs 3, 4, 6, 7, and 9, as well as documentation and code for the new Modular Program 8 (**Table 9**). Modular programs 1, 2, and 5 were combined to create Modular Program 9 and have been retired. The MSOC will continue to improve the Modular Programs to increase functionality and will post revised documentation and code as they are developed.

Table 9. Modular Program Documentation Posted in Year Five

Date	Document Title
2/3/14	Modular Program 3: Frequency of Select Events During Exposure to a Drug/Procedure Group of Interest (version 7.1)
7/14/14	Modular Program 3: SAS Code
3/13/14	Modular Program 4: Frequency of Select Events During Concomitant Exposure to a Drug/Procedure Groups of Interest (version 5.0)
6/11/14	Modular Program 4: SAS Code
2/3/14	Modular Program 6: Frequency and Duration of Treatment Following an Event of Interest (version 7.0)
6/11/14	Modular Program 6: SAS Code
3/13/14	Modular Program 7: Drug Use, Medical Diagnoses, and Medical Procedures Before and After an Exposure or Event of Interest (version 5.0)
6/11/14	Modular Program 7: SAS Code
6/27/14	Modular Program 8: Uptake, Use, and Persistence of New Molecular Entities (NMEs)
6/27/14	Modular Program 8: SAS Code
2/3/14	Mini-Sentinel Modular Program 9: Background Rates and Characterization of Health Outcomes of Interest among Individuals with or without Condition(s) of Interest (version 3.0)

Date	Document Title
6/11/14	Modular Program 9: SAS Code

2. Mini-Sentinel Toolkit Library

During Year Three and Year Four, the MSOC posted to the Mini-Sentinel website a library of standalone programming tools written to standardize routine programming procedures, such as selecting a cohort of members exposed to specific medical products, creating continuous treatment episodes, or identifying continuous enrollment periods. Each tool is a self-contained SAS® macro. These tools are used in combination to facilitate development of the Mini-Sentinel Modular Programs. The MSOC will continue to post new and revised programming tools as they are developed. No new or revised tools were posted to the website during Year Five.

All SAS programs posted to the Mini-Sentinel library include a user guide and documentation. Each standalone macro comes with examples and test datasets to be used as test scenarios to speed development work.

3. SOPs

During Year Five, MSOC posted one revised Standard Operating Procedure (SOP) to the Mini-Sentinel website, Data Quality Review and Characterization. Mini-Sentinel SOPs provide internal guidance for the operations and procedures of various Mini-Sentinel activities.

X. CONCLUSION

This report described the activities of the Mini-Sentinel Coordinating Center's Data Group during Year Five of the Mini-Sentinel project. As described throughout this report, significant progress was made in areas that are key to the sustainabilty and efficiency of the operations and infrastructure development to serve the various needs of the project, namely the data model expansion and improved quality assurance, multiple enhancements to web-based and core analytic infrastructures, and increased capacity to request fulfillment and workgroup support. For Year Six we look forward to continuing the progress of this unprecedented and significant public health initiative.

XI. 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.

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.

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.

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.

8. Bredfeldt CE, Butani A, Padmanabhan S, Hitz P, Pardee R. Managing protected health information in distributed research network environments: automated review to facilitate collaboration. *BMC Med Inform Decis Mak.* 2013; 13:39.

9. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of Clinical Epidemiology.* 2011; 64: 7, 749-759.

10. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, Wagner EH, Pardee R, Schmidt MM, Geiger A, Butani AL, Field T, Fouayzi H, Miroshnik I, Liu L, Diseker R, Wells K, Krajenta R, Lamerato L, Neslund Dudas C. Building a virtual cancer research organization. *Journal of the Natational Cancer Institute, Monographs.* 2005; 35:12-25.

11. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for Drug and Vaccine Safety Surveillance. *Sequential Analysis.* 2011; 30: 58-78.

12. Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, Curtis LH, Weiner MG, Archdeacon P, Brown JS. Electronic clinical laboratory test results data tables: lessons from mini-sentinel. *Pharmacoepidemiology and Drug Safety.* 2014; 23:6, 609-618.