

MINI-SENTINEL COORDINATING CENTER DATA GROUP YEAR FOUR REPORT OF ACTIVITIES

September 2012 to September 2013

Prepared by:

Mini-Sentinel Coordinating Center Data Group and Collaborating Institutions

February 2014

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

**Mini-Sentinel Coordinating Center Data Group
Year Four Report of Activities
September 2012 to September 2013**

TABLE OF CONTENTS

I. INTRODUCTION	1
A. OVERVIEW OF THE MINI-SENTINEL PROJECT	1
B. MINI-SENTINEL SCIENTIFIC OPERATIONS CENTER.....	1
1. <i>Responsibilities of the Data Infrastructure Group</i>	2
2. <i>Responsibilities of the Production Group</i>	2
3. <i>Responsibilities of the MSDD and Communications Group</i>	2
C. MINI-SENTINEL DATA CORE	4
1. <i>Overview</i>	4
2. <i>Members of the Data Core</i>	4
3. <i>Members' Terms and Selection</i>	4
4. <i>Data Partners</i>	4
D. DISTRIBUTED DATA APPROACH	4
II. OVERVIEW OF COMMON DATA MODEL.....	5
III. EXPANSION OF THE MINI-SENTINEL COMMON DATA MODEL.....	9
A. CLINICAL DATA ELEMENTS.....	9
1. <i>Overview</i>	9
2. <i>Roles and Responsibilities</i>	9
3. <i>Selection of Data Elements</i>	10
4. <i>Revisions and Implementation of the Data Model for Clinical Data</i>	11
5. <i>Querying Laboratory Result Data</i>	14
6. <i>Use of Standards and Controlled Terminologies</i>	14
7. <i>Potential Next Steps for Clinical Additions</i>	14
B. OTHER REVISIONS TO THE MSCDM.....	15
1. <i>MSCDM Tables: Text Revisions</i>	15
2. <i>MSCDM Variables: Expansion of Code Lengths</i>	15
3. <i>MSCDM Tables: Addition of Laboratory Result Guidelines</i>	16
4. <i>MSCDM Tables: Addition of Incident Summary Tables</i>	16
C. EXPANSION REPORT	16
1. <i>Priority Exposures and Health Outcomes of Interest</i>	16
2. <i>Capture of Priority Exposures and Health Outcomes of Interest in the Current MSCDM</i>	17
D. LESSONS LEARNED AND SUGGESTIONS FOR FUTURE WORK.....	18
1. <i>Clinical Data Elements</i>	18
2. <i>Expansion of the MSCDM</i>	18
IV. MINI-SENTINEL DISTRIBUTED DATABASE	18
A. DATA QUALITY ASSURANCE REVIEW AND CHARACTERIZATION	18
1. <i>Overview</i>	18
2. <i>Roles and Responsibilities</i>	19
3. <i>Data QA Review and Characterization Specifications</i>	19
4. <i>Data QA Review and Characterization Revisions</i>	23

5.	<i>Reporting</i>	23
6.	<i>Data Completeness and Availability</i>	23
7.	<i>Principal Diagnosis Flag (PDX) Variable Investigation</i>	24
B.	INCORPORATION OF NATIONAL DATA STANDARDS AND CONTROLLED TERMINOLOGIES	24
1.	<i>Incorporation of Standards into the MSCDM</i>	24
2.	<i>Engagement with National Standards Organizations</i>	26
3.	<i>Impact of Transition to ICD-10</i>	27
C.	LESSONS LEARNED	28
V.	MINI-SENTINEL ANALYTIC TOOLS	28
A.	MODULAR PROGRAMS	28
1.	<i>Overview</i>	28
2.	<i>Roles and Responsibilities</i>	29
3.	<i>Modular Program Revisions</i>	30
4.	<i>Other Modules</i>	33
5.	<i>Beta-testing</i>	34
B.	SUMMARY TABLES	34
1.	<i>Overview</i>	34
2.	<i>Roles and Responsibilities</i>	36
3.	<i>Summary Table Revisions</i>	37
C.	MINI-SENTINEL DISTRIBUTED QUERY TOOL	37
1.	<i>Overview of Query Tool</i>	37
2.	<i>Network Implementation</i>	39
3.	<i>Platform Enhancements for Mini-Sentinel Query Tool Version 3.2</i>	39
4.	<i>Deploying Platform Enhancements to the Data Partners</i>	42
5.	<i>Portal Enhancements</i>	42
D.	WEB-BASED LIBRARY AND TOOLKIT.....	43
1.	<i>Overview of Web-Based Library</i>	43
2.	<i>Description of Currently Available Tools</i>	43
E.	ELECTRONIC SUPPORT FOR PUBLIC HEALTH (ESP).....	44
1.	<i>Enhancing ESP's Driver to Create Fake Clinical Data</i>	44
2.	<i>Installation and Documentation</i>	45
F.	LESSONS LEARNED	45
1.	<i>Modular Programs</i>	45
2.	<i>Summary Tables and Distributed Query Tool Software</i>	45
VI.	MINI-SENTINEL INFRASTRUCTURE.....	46
A.	MINI-SENTINEL DATA CATALOG	46
1.	<i>Function of the Mini-Sentinel Data Catalog</i>	46
2.	<i>Expansion of the Mini-Sentinel Data Catalog during Year Four</i>	47
3.	<i>Future Work</i>	47
B.	IMPLEMENTATION OF MANTIS ISSUE TRACKING SYSTEM.....	47
C.	AUTOMATED REPORTING TOOL	48
1.	<i>Function of the Automated Reporting Tool</i>	48
2.	<i>Future Work</i>	48
D.	MINI-SENTINEL SECURE PORTAL.....	48
E.	TESTING ENVIRONMENT AND SYNTHETIC DATA	48
F.	LESSONS LEARNED	49
VII.	OTHER DATA CORE ACTIVITIES.....	50
A.	COMMUNICATIONS.....	50

B. SUPPORT TO WORKGROUPS	50
C. DISSEMINATION ACTIVITIES	51
1. <i>Manuscripts</i>	51
2. <i>Meeting Presentations</i>	51
VIII. MSDD QUERY REQUEST SUMMARY	53
A. MODULAR PROGRAMS	53
B. SUMMARY TABLES AND QUERY TOOL	56
C. AD HOC REQUESTS	58
D. POSTINGS TO MINI-SENTINEL WEBSITE.....	59
1. <i>Reports</i>	59
2. <i>Other Postings</i>	60
E. LESSONS LEARNED	62
1. <i>Modular Programs</i>	62
2. <i>Summary Tables and Query Tool</i>	62
IX. CONCLUSION	62
X. REFERENCES	63

I. INTRODUCTION

This report describes the activities of the Mini-Sentinel Coordinating Center's Scientific Operations Center during Year Four - September 23, 2012 to September 22, 2013. Ongoing activities are included in the report, as well as one-time activities that were undertaken during the project year.

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 Mini-Sentinel Operations Center (MSOC) is part of the Mini-Sentinel Coordinating Center (see Figure 1). The MSOC leads Mini-Sentinel's scientific and management operations, via the Scientific Operations Center (SOC) and Management Operations Center (MOC), respectively. The SOC is organized into three groups: Data Infrastructure, Production, and MSDD and Communications. Through these groups, the SOC: 1) creates the infrastructure, tools, and processes required to implement and use the Mini-Sentinel data resources; 2) supports the scientific work of the Data, Methods, and Protocol Data Cores as well as all Mini-Sentinel project workgroups; and 3) provides technical support, guidance, and consulting on appropriate uses of the Mini-Sentinel data resources. This report focuses on the activities and responsibilities of the Scientific Operations Center, but by necessity includes cross-functional activities of the Management Operations Center.

1. Responsibilities of the Data Infrastructure Group

The **Data Infrastructure Group's** primary responsibilities are to build and manage the data infrastructure and tools required to enable rapid and efficient querying of the Mini-Sentinel Distributed Database (MSDD). The Data Infrastructure Group has developed [Standard Operating Procedures](#) for creating SAS programming code for use in the network, manages and supports all programming activities, develops and manages the Mini-Sentinel public website and private secure communications systems, and oversees implementation and use of the Mini-Sentinel Distributed Query Tool.

2. Responsibilities of the Production Group

The **Production Group** has primary responsibility for efficiently using the MSDD to answer data requests. Data requests can be initiated by FDA or Mini-Sentinel workgroups and typically involve the use of existing Mini-Sentinel querying tools such as modular program and summary tables. The Production Group establishes Standard Operating Procedures for query request fulfillment, including activities related to initiating the request and parameter settings, testing the request parameters, executing the request, and providing a report to the requester (see Figure 2).

3. Responsibilities of the MSDD and Communications Group

The **MSDD and Communications Group** has primary responsibility for supporting Data Partners' development of their Mini-Sentinel Distributed Database (MSDD). This involves developing, updating, and managing the Mini-Sentinel Common Data Model (MSCDM), managing the data refresh and approval process that includes data quality checking 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 MSDD and Communications Group develops and implements data quality checking and characterization metrics and works with Data Partners to improve use of the MSDD for FDA activities.

Together, the SOC's **Data Infrastructure Group, Production Group, and MSDD and Communications Group** enable efficient and appropriate use of the Mini-Sentinel data resources. The groups work closely on a daily basis to improve functioning of the network and to develop new tools; most Scientific Operations Center analysts work across these three groups to ensure effective communication. SOC staff members are members of the Mini-Sentinel Data Core and support and work closely with the FDA, Data Partners, and Collaborating Institutions on all scientific Mini-Sentinel activities.

Figure 1. Mini-Sentinel Coordinating Center

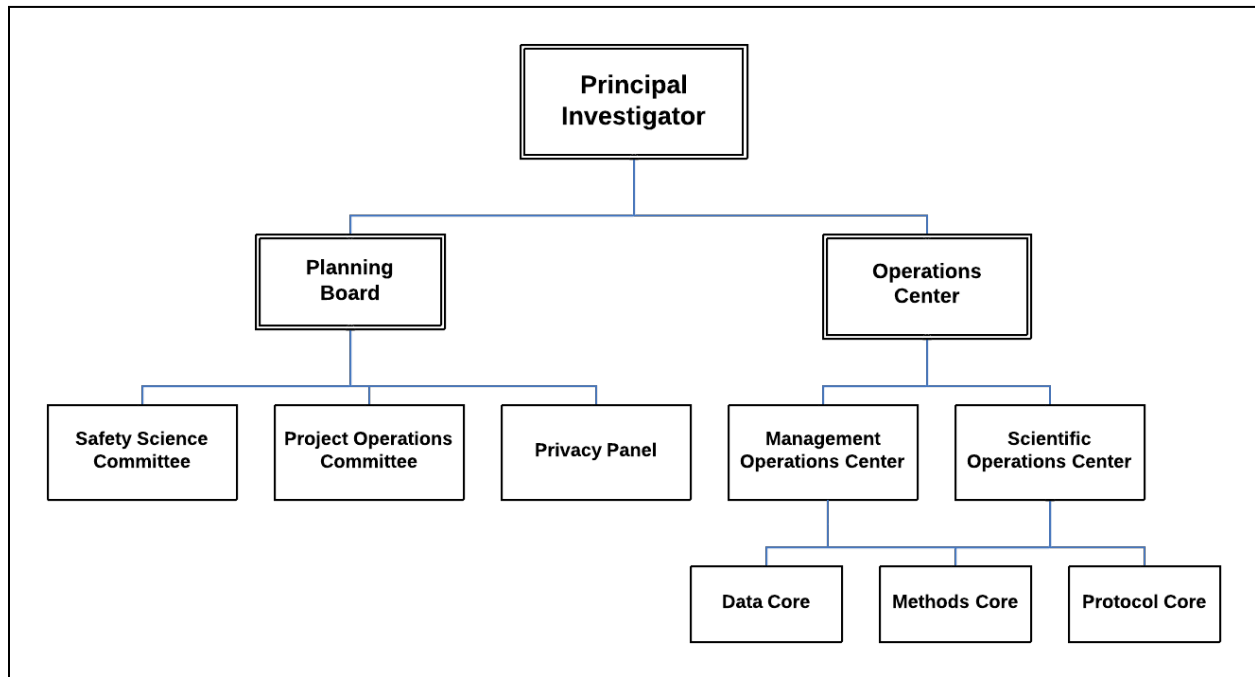
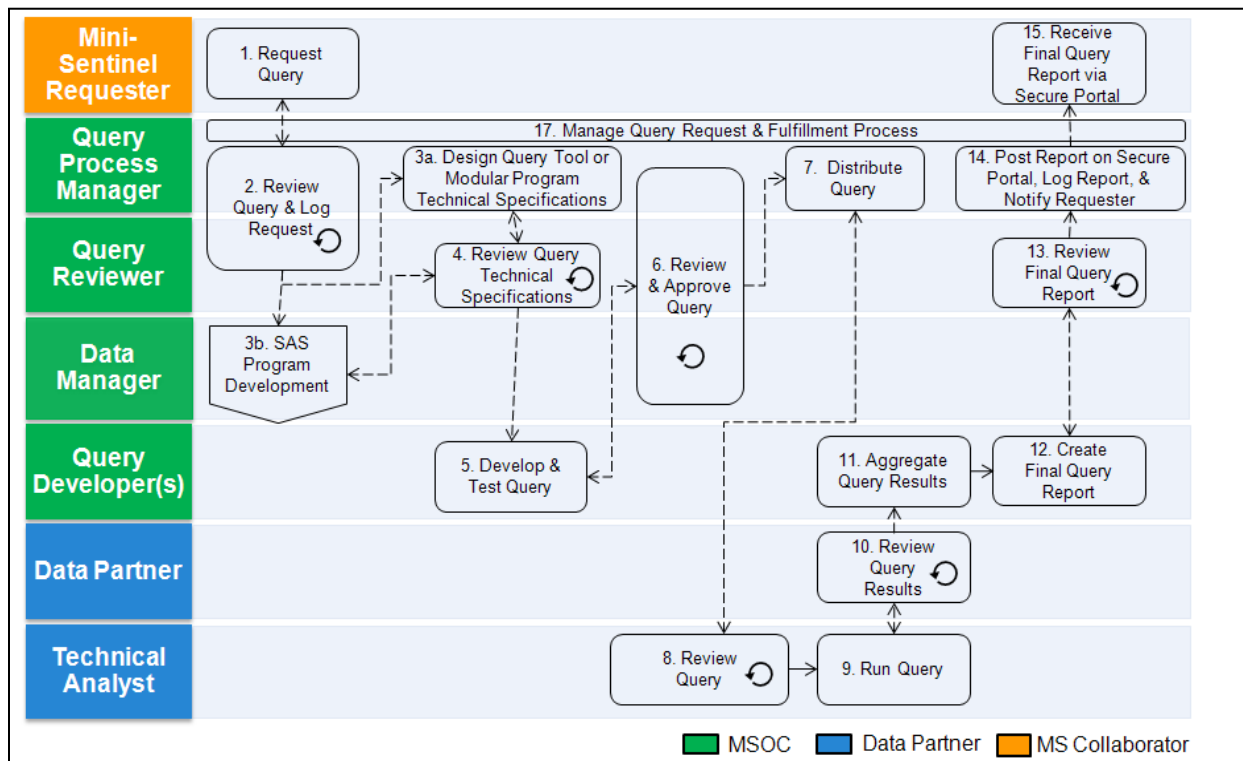


Figure 2. Mini-Sentinel Query Fulfillment Process



C. MINI-SENTINEL DATA CORE

1. Overview

The Mini-Sentinel Data Core is an oversight committee that sits inside the Mini-Sentinel Coordinating Center (see Figure 1) and is led by investigators from the Mini-Sentinel Collaborating Institutions. The Data Core provides guidance to the FDA and the Mini-Sentinel Scientific Operations Center on activities related to development and appropriate use of the Mini-Sentinel data resources. The Data Core works closely with the two other Mini-Sentinel Coordinating Center cores, the Methods Core and the Protocol Core. As directed by FDA, the Data Core Leader(s) assist the Mini-Sentinel Scientific Operations Center with external communications, including presentation of the ongoing Mini-Sentinel activities at scientific meetings and related venues.

2. Members of the Data Core

- Data Core Leaders
- Mini-Sentinel Scientific Operations Center staff (i.e., Data Infrastructure, Production, and MSDD & Communication Groups)
- 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, those with health plan administrative and claims data in the MSCDM format include Aetna, HealthCore, Inc. (working with WellPoint data), the HMO Research Network, Humana, Kaiser Permanente Center for Effectiveness and Safety Research (KP CESR), Lovelace Clinic Foundation, OptumInsight, and Vanderbilt University (working with 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. Efforts to incorporate these data areas into the MSCDM are ongoing and will continue to be the focus of activities in subsequent years.

D. DISTRIBUTED DATA APPROACH

Mini-Sentinel uses a distributed data approach in which Data Partners maintain physical and operational control over electronic data in their existing environments.¹⁻⁷ 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 Mini-Sentinel Common Data Model standardized format, employing identical names and formatting for each data element across all Data Partners. Data Partners execute standardized programs provided by the Operations Center or project workgroups and return

the output of the programs to the MSOC or project workgroups. Typically, the output of these programs 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)ⁱ. 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 aggregate 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 MSCDM v3.0 is comprised of 11 data tables with person-level medical care and administrative data. One data table, the State Vaccine table, is new in Year 4. This section describes the 11 data tables. Twelve summary tables, derived from these data tables are described in [Section V.B.](#) below.

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 added in Year 4 are available in [Overview and Description of the Common Data Model](#).ⁱⁱ

A common unique person identifier is included in each table to allow linkage across the tables and comprehensive view of patient care during an enrollment period. The unique person identifier is not a true identifier (e.g., Social Security Number), but rather a health-plan generated, alpha-numeric string unique to each person in the data files. Each health plan maintains a link between the unique person identifier and the true identifier, which is retained by the Data Partner. The person identifier is unique

ⁱ <http://www.hhs.gov/ocr/privacy/>

ⁱⁱ MSCDM v3.0 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.

within a health plan and is not shared outside the health plan with other Data Partners, the MSOC or the FDA.

Each table is briefly described below.

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)⁸ enrollment table structure and similar in structure to the other common data models evaluated.

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

Dispensing: The dispensing table represents outpatient pharmacy dispensing captured by the Data Partners 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. This typically requires summation of dispensing information by unique person identifier, dispensing date, and dispensed NDC. No negative days supplied or amounts dispensed appear in the table and no corrections are made for values that are “out of range,” such as 900 days supplied.

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

Medications dispensed at discount pharmacies (e.g., Walmart, Target) 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. An analysis of pharmacy dispensing data for 11 HMORN health plans found that OTC medications accounts for 2% to 9% of all outpatient dispensings between 2000 and 2007, although this rate of capture is likely to be a small portion of all OTC use.⁹ Infused medications, vaccinations, and other medications (e.g., injections) not dispensed through a pharmacy (e.g., provided directly by medical providers) are captured in the procedure table because those administrations are considered “procedures” within the existing medical coding nomenclature and are captured by the Data Partners in a separate data stream. A very small percentage (less than 0.1%) of outpatient dispensings represent NDCs for procedures.⁹ 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 and is 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. Additional information in this table 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 one-to-many relationship so that all the associated diagnoses are recorded in the diagnosis table. The diagnosis table includes one row for each unique diagnosis recorded during an encounter. The table also includes a flag for whether the diagnosis was recorded in the primary 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. In Year Four, the length of the diagnosis code variable was expanded to eight characters to accommodate ICD-10 diagnosis codes.

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 one-to-many relationship so that all the associated procedures are recorded in the procedure table. The procedure table includes one row for each unique procedure recorded during an encounter. The table includes the unique person identifier, the procedure code, an indicator for the care setting in which the procedure was recorded, and the specific type of procedure recorded. 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. The table allows capture of any existing or future coding standards. In Year Four, the length of the procedure code variable was expanded to eight characters to accommodate ICD-10 procedure codes.

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 because the death registry also has a data lag. Within the death table, the death date is recorded, along with imputation method if the exact date is not known. The table also includes a source for the 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 separate cause of death table that records diagnosis codes reflecting the underlying condition, along with coding dictionary used, type of contribution to the death, and the source of the information. 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. In Year Four, the length of the cause of death code variable was expanded to eight characters to accommodate ICD-10 diagnosis codes.

Laboratory Result: The laboratory result table structure includes multiple tables that together facilitate capture and use of laboratory result information. The table represents results and information 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. Two variables were removed from the model in Year Four including a local code for ordering provider and a flag indicating whether the LOINC code was imputed. There are no imputed LOINC codes in the MSCDM.

In Year Four, the laboratory result data model was updated to include a list of currently known LOINC and CPT-4 (Current Procedural Terminology) 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.

In Year Four, 12 Data Partners have implemented the following laboratory results (test results are from blood, serum, or plasma unless noted): alkaline phosphatase (ALP), alanine aminotransferase (ALT), absolute neutrophil count (ANC), total bilirubin, creatine kinase total, creatine kinase MB fraction,

creatinine kinase MB relative index (creatinine kinase MB fraction divided by creatinine 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 Stat Vaccine Table to capture state vaccine registry information collected by the 4 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 codes using a range of terminologies, including CPT-4 and 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 Four, the Mini-Sentinel Clinical Data Elements Workgroup led a number of expansion activities. One key activity was incorporating the laboratory results and vital signs data into the regular MSDD updates and quality checks from all Data Partners with laboratory results (12 Data Partners) and vital signs (9 Data Partners) data. The MSCDM laboratory results data dictionary was finalized and approved and now only requires routine updates along with the rest of the MSCDM. The workgroup guided the addition of eight laboratory test result types into the MSCDM at four additional KP sites. Detailed laboratory results data characterization was undertaken and completed for six types of laboratory test results, described below. Finally, the workgroup led development and implementation of two Modular Program enhancements; Modular Program 3 was enhanced to allow stratifications of change in BMI for the pediatric population and Modular Program 6 was enhanced to allow use of laboratory test results as an index event or post-diagnosis event.

2. Roles and Responsibilities

The Clinical Data Elements Workgroup lead team, comprised of the Data Core co-leads, and some of the MSOC and FDA staff on the workgroup, led and managed all aspects of the workgroup. Activities completed include:

- Weekly conference calls to address all deliverables and ensure adherence to timelines
- Monthly Data Partner webinar and conference calls
- Communicated with Data Partners and with the FDA
- Supported, guided, and assisted Data Partners with incorporating laboratory results
- Made changes requested as a result of data characterization
- Provided reports and updates to the Data Core
- Wrote and revised programming needed to characterize the laboratory results data
- Implemented and validated the modular program revisions for clinical data elements

3. Selection of Data Elements

In Year Two, the initial set of laboratory tests included in the MSCDM included:

- glucose (random and fasting)
- hemoglobin
- HbA1c
- creatinine
- ALT
- alkaline phosphatase
- total bilirubin
- INR
- D-dimer
- lipase
- absolute neutrophil count (ANC)

Each of these laboratory tests, except ANC, were incorporated by all participating Data Partners (ANC was incorporated only by HealthCore).

In Year Three, eight new laboratory test results were incorporated into the MSCDM, including:

- troponin-T
- troponin-I
- platelets
- total CK
- CK-MB fraction
- pregnancy
- influenza testing
- ANC

This represented the first time that non-blood tests were incorporated into the MSCDM, as pregnancy tests were comprised of urine tests as well as blood tests and influenza testing included specimens from several different sources (e.g., nasal swab or wash, oropharyngeal swab, and bronchoalveolar lavage).

Of the 12 Data Partners providing laboratory data, all have updated these data through at least 2011. Most now include 2012 data and some a portion of 2013 data.

Vital signs incorporated into the MSCDM during Year Two included height, weight, systolic blood pressure, diastolic blood pressure, and tobacco status. In Year Two, several Kaiser Permanente regions and three HMORN sites incorporated vital signs data. In Year Three, no new/additional vital sign data elements were added, but existing vital sign data elements were updated by all nine Data Partners contributing vital sign data. In Year Four, all Data Partners contributing vital sign data updated data elements through at least 2011. Most Data Partners updated through calendar year 2012, and some with a portion of 2013 data.

Further information regarding building the clinical components' data model can be found in the Mini-Sentinel [Year 2 CDM and Data Core Activities Report](#).

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

While Years Two and Three activities focused on incorporating new types of laboratory test results into the MSCDM, Year Four activities focused on expansion of laboratory test results data table capabilities and use. In Year Four, an extensive development and implementation activity was undertaken to incorporate the laboratory results data tables into regular MSDD updates and quality checks at all Data Partners with laboratory results data for a select set of tests.

The MSCDM laboratory results data dictionary was finalized and approved and the laboratory data table was renamed the "Laboratory Results Data Table." The lab data dictionary is now updated as new information becomes available and as data characterization activities reveal refinement is necessary for completeness and/or advisable for clarity. The revised data model for the laboratory results table is included in MSCDM v3.0 available on the Mini-Sentinel website.

In Year Four, detailed laboratory results data exploration and characterization was undertaken to implement a valid, harmonized, "common" laboratory data model that incorporates standardized results units to enable use in routine MS data activities. Year Three work demonstrated variability and inconsistency in the structure of laboratory test results reporting across Data Partners and within Data Partners over time/across facilities, complicating uniformity in lab data mapping. This lack of consistency was observed in virtually every data element of the laboratory results table (e.g., result units, normal ranges, LOINC codes, specimen sources, patient location, result location, and test immediacy). A systematic process was used to determine which laboratory test types were priorities for characterization and the workgroup then led the characterization process. First, the laboratory tests in the MSCDM were ranked based on anticipated level of difficulty in characterizing and harmonizing. Second, in collaboration with the FDA, six laboratory test types were selected for Year Four characterization. These test types included:

- Alanine aminotransferase
- Creatine kinase, total
- Creatinine
- Glucose, fasting and random
- Hemoglobin A1c
- INR

In Year Four, after test types for characterization were selected, we developed program code to assess the numbers, types, and variations of test subcategories, result values, result units, LOINCs, patient

location, result location, specimen sources, immediacy, modifiers, abnormal indicators in the laboratory results table both within and across Data Partners. The Data Partners ran the characterization programs against their laboratory results data 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. For example, the workgroup identified a wide range of recorded result units for hemoglobin A1c, including the following units, as being used by a single Data Partner:

- Blank
- %
- % A1C
- % A1c
- % NGSP
- % OF TOTAL
- % TOTAL HGB
- % of Hgb
- % of total
- %A1C
- %AIC
- %Hb
- %HbA1c
- %NGSP
- %T.Hgb
- %THb
- HbA1c%
- MG/DL
- NULL
- PERCENT
- Percent
- g/dL
- mmol/mol

This Data Partner was instructed to:

- “upcase” all lowercase result units for consistency
- set all variations of “% A1c” to the standard hemoglobin A1c result unit, and
- remove records with invalid result units (e.g., mg/dL)

As “mmol/mol” is a valid original result unit (standard in Europe), the Data Partner was provided the mathematical formula to convert this result unit to the MSCDM standard and asked to apply this formula to any original result units of mmol/mol before submitting future laboratory results data table refreshes. The data model captures the original and the converted values.

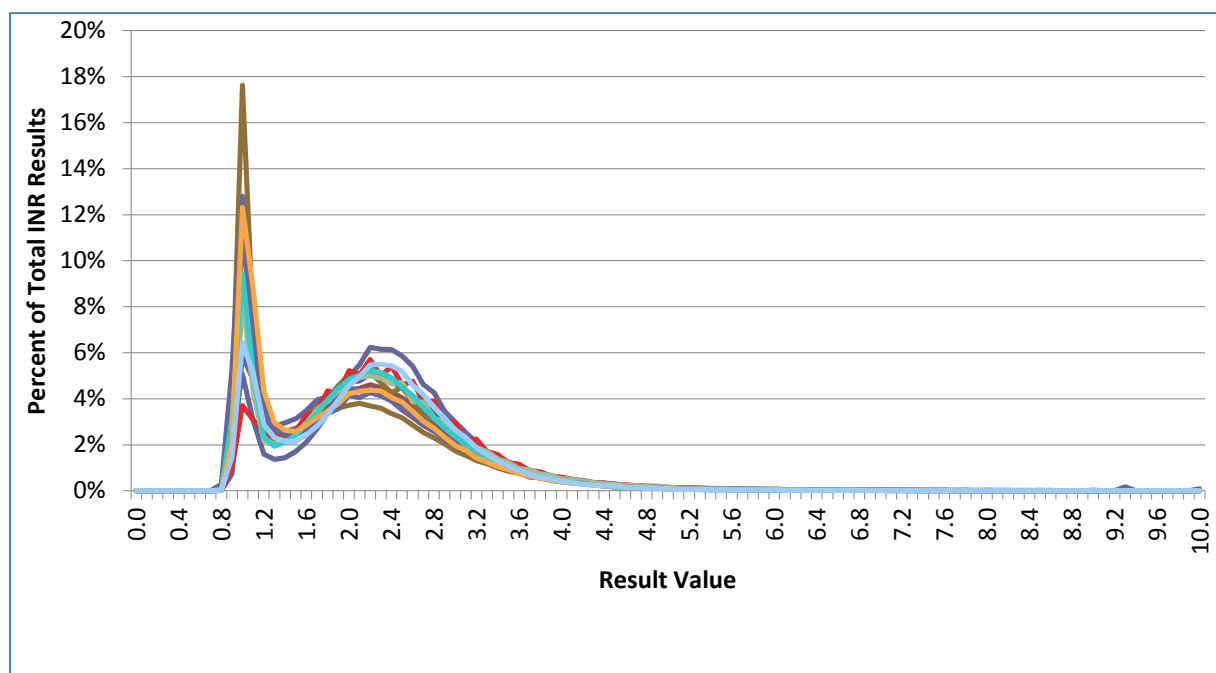
In the second example, counts of “Original Result Unit” were identified from one Data Partner in the total creatine kinase data characterization. Investigation found that the “blank” result units have the same results distribution as the original result units labeled “U/L” As U/L (units/Liter) is the generally accepted standard result unit for total creatine kinase, this Data Partner was instructed to keep those

test results in the MSCDM (i.e., not delete these test results with missing result units) and to set the results with “blank” units to “UNK” (Unknown) units in the standard result unit data field while keeping the original result unit data field as “blank”. Also, total creatine kinase results labeled “Units”, “Units/L”, etc., were set to U/L at all Data Partners. This intensive scrutiny and subsequent guidance is the norm rather than the exception in characterizing and harmonizing the laboratory results table. While the Clinical Data Elements Workgroup anticipated many of the inconsistencies observed in the data (some non-numeric results for quantitative tests), some findings were not anticipated. Unanticipated inconsistencies are being dealt with as they arise.

Result values represent a fundamental difference between the MSCDM clinical data elements tables’ content and the administrative data tables. That is, the clinical laboratory results table contains actual clinical results; it does not only indicate the presence or absence of the test. Therefore, the Year Four data characterization work included characterizing the results content and values observed in the data.

One example of the results characterization work is presented below. Figure 3 is a graph of the percent of INR results from each participating Data Partner that are specific result values. As the graph confirms, the INR result values contained in the MSCDM are both appropriate and valid and indicate the Data Partners extracted, transformed, and loaded the correct laboratory data into the MSCDM. This interpretation of the results data is based on the distribution of the INR results. In particular, the spike at INR = 1 is indicative of normal INR and is observed across all Data Partners. The secondary peak is indicative of the usual therapeutic anticoagulation targeted values. The long right tail of the distribution is also consistent with clinical expectations for population INR distribution.

Figure 3: Percent of INR Results by Data Partner (one line per Data Partner)



Using the characterization approach detailed above, the Clinical Data Elements Workgroup and participating Data Partners completed data characterization and provided guidance for harmonizing all

six laboratory results test types targeted for Year Four. Although the focus in Year Four was on clinical laboratory data, data quality checks were 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. Queries could be initiated via de novo programming or a modular program, and could be done on any test type included in the MSCDM. While requesters could query whether a valid laboratory result was present, laboratory result values could not be queried.

Following the characterization and standardization of ANC, creatine kinase total, creatinine, glucose, HgA1c, and INR lab result values in Year Four, investigators can now query result values for these six test types via de novo programming.

In Year Five, the MSOC will enhance modular programs to allow 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 will allow investigators to query the results of any laboratory test type, specific laboratory test results will only be able to be queried after they have been characterized and deemed ready for use by the Clinical Data Elements Workgroup.

6. Use of Standards and Controlled Terminologies

Mini-Sentinel 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 Mini-Sentinel laboratory result test type nomenclature. 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 will continue to work with FDA and the Data Partners to assess more robust application of LOINC (or potentially other standards) as possible.

Laboratory test results are qualitative (e.g., urine pregnancy) or quantitative (e.g., blood glucose) and must be standardized to uniform units. For example, “positive,” “negative,” “borderline,” and “un determined” are pregnancy result units found in source data that must be standardized. As previously discussed, numeric result units are frequently reported in different units (e.g., Units, U, IU) and must also be standardized. 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.

7. Potential Next Steps for Clinical Additions

There are multiple potential areas of future work related to the characterization of and additions to the Laboratory Results Table and the Vital Signs Table in the MSCDM. Some of these include:

- Characterization and harmonization of other MSCDM clinical laboratory test results.
- Exploration of data elements to enhance understanding of patterns, frequency, usefulness, and logic inconsistencies within the existing data (e.g., implausible changes in height).

- Assessment of the sensitivity and positive predictive value of selected coded diagnoses based on laboratory results data availability (e.g., CKD diagnosis).
- Investigation of the value added by including laboratory test results into medical product safety surveillance (e.g., CK and statin exposure, glucose and hypoglycemia with antidiabetic medication exposure).
- Capturing test results from additional data sources such as inpatient facilities and specialty clinical centers (as few Data Partners currently contributing laboratory results data have inpatient laboratory test results).
- Refinement of modular programs to enable more robust feasibility assessments of laboratory tests results and vital signs.
- Development of “user’s guides” about the strengths and weaknesses of the clinical laboratory and vital signs data captured, the contents of the tables, what harmonization has been done, which laboratory test types are available from all (or only a subset of) participating Data Partners, which Data Partners have more comprehensive laboratory data capture, and sub-populations with clinical data available for use in Mini-Sentinel.

B. OTHER REVISIONS TO THE MSCDM

During Year Four, the following modifications were made to the MSCDM: 1) clarifications to table descriptions and variable guidelines; 2) expansion of the length of diagnosis and procedure code variables to accommodate ICD-10 codes; 3) addition of a comprehensive laboratory result guideline table; and 4) addition of Incident Summary Tables to the MSCDM. All revisions are described below and included in the updated MSCDM available on the Mini-Sentinel website.

1. MSCDM Tables: Text Revisions

The MSDD and Communications Group continued to revise and clarify the MSCDM based on lessons learned and feedback from Data Partners as well as other stakeholders. As Mini-Sentinel investigators used the model more frequently and as new collaborators and programmers began to work with the CDM, more descriptions and variable guidelines needed to be clarified or amended. Definitions of some of the fields and tables (e.g., meaning of uniqueness of a row in each table) as well as the examples provided were refined to improve understanding of the model.

A new footnote was added to the MSCDM Enrollment table to guide Data Partners in setting enrollment start and end dates. Previously when an enrollment record ended on or after January 2, 2000, the enrollment start date could be any date on or before January 1, 2000. In Year Four, the MSOC began enforcing a hard start date of January 1, 2000. For records with an enrollment start date before this date, Data Partners were advised to set the enrollment start date equal to January 1, 2000 (artificially move it forward).

2. MSCDM Variables: Expansion of Code Lengths

To prepare for ICD-10-CM implementation, the lengths of the diagnosis and procedure code variables were expanded from 6 characters to 8 characters.

3. MSCDM Tables: Addition of Laboratory Result Guidelines

During Year Four, twelve of the Data Partners began providing regularly updated laboratory result data. To assist their efforts, MSOC added comprehensive laboratory result guidelines to the MSCDM Overview and Description. For each of the twenty-five laboratory tests in the model, this table shows the acceptable values for test sub-category, specimen source(s), standardized result units and guidance on standardized results. This table will be updated as MSOC and the Data Partners work to characterize and standardize the lab test results in Years 4 and 5.

As mentioned above in [Section II](#), Overview of the Common Data Model, in Year Four the model was also updated to include a list of currently known LOINC codes and common CPT-4 codes. The model also includes a table of standard abbreviations for common laboratory units.

4. MSCDM Tables: Addition of Incident Summary Tables

In Year Four, descriptions of the three new [Incident Summary Tables](#) were added to the MSCDM for completeness and transparency (Incident ICD-9-CM Diagnosis Summary Table (3-Digit), Incident Generic Name Summary Table, and Incident Drug Category Summary Table). These tables are used to enable rapid querying through the Mini-Sentinel Query Tool. A description of the new Age Groups Summary Table was also added. This is a static table in the MSDD, which provides a key for the age group stratifications within each summary table.

C. EXPANSION REPORT

To ensure that the MSCDM evolves in a way that anticipates the Agency's future needs, FDA charged MSOC with developing a plan for expansion of the MSCDM. The MSCDM Expansion Workgroup, led by the Data Core co-Leads, included the MSOC Data Core, Darren Toh (MSOC), David Cole (MSOC), Meghan Baker (MSOC), Patrick Archdeacon (FDA-CDER), Marsha Reichman (FDA-CDER), and Michael Nguyen (FDA-CBER). The workgroup solicited specific priorities for exposures of interest and health outcomes of interest from the Center for Drug Evaluation and Research (CDER), the Center for Devices and Radiological Health (CDRH), and the Center for Biologics Evaluation and Research (CBER), and gathered information about the exposure setting, timing of outcomes with respect to exposures of interest, and capture in the current MSCDM. The exposures and health outcomes of interest and their related characteristics were used to guide structured discussions with Mini-Sentinel Data Partners.

1. Priority Exposures and Health Outcomes of Interest

A majority of the 160 exposures of interest identified by CDER and CBER occur predominantly in outpatient settings. Whereas many exposures of interest to CDER are administered orally in outpatient settings, many exposures of interest for CBER are administered by injection or IV infusion in inpatient and outpatient settings. Specific clinical data needs were identified for a minority of the 89 health outcomes of interest reviewed. Supplemental death information (cause of death), necessary to accurately establish disease-specific mortality, was identified as a priority area for expansion as well.

2. Capture of Priority Exposures and Health Outcomes of Interest in the Current MSCDM

Medications dispensed on an outpatient basis are relatively completely captured in the MSDD. By contrast, capture of inpatient administrations is variable and incomplete. Although many exposures of interest to CDER occur predominantly in outpatient settings, inpatient administrations are not uncommon. Additionally, exposures of interest to CBER include blood products and products that are administered, in part, in inpatient settings. CDER and CBER have strong interests in developing the ability to assess the safety of medications given primarily in the inpatient setting. Of note, inpatient laboratory data related to inpatient diagnoses of several health outcomes of interest are not included in the current MSCDM from most Data Partners. Finally, cause of death is not available for two-thirds of deceased members. The Data Partners have various mechanisms for acquiring information about an enrollee's death. To confirm a member's death, half of the Data Partners link to local (state) death registries to update the death status and identify cause of death of their members.

In identifying recommendations for addressing gaps in the MSCDM, the workgroup undertook a preliminary assessment of risks and likelihood of success for each identified gap and identified the high priority expansion activities below. All expansion activities are detailed in the [MSCDM Expansion Plan](#).

Access to inpatient data streams: Encounter-level inpatient claims received by many of the large Data Partners do not enable us to identify specific inpatient administrations. Moreover, inpatient data streams are available for a minority of Data Partners accounting for less than 10% of enrollees in the MSDD. Rather than invest the significant time and resources to develop internal inpatient data sources, an alternative approach is to partner with a national hospital-based organization and develop inpatient table(s) for the MSCDM from their data sources. Initial assessments would focus on exposure-outcome pairs likely to occur during the same inpatient stay. As cross-institutional linkages are developed, exposure-outcome pairs could cross settings.

Enhance supplemental death data through linkages with the National Death Index: Although not initially recommended by the workgroup given the 2-year lag in release of National Death Index data, recent improvements in the timeliness of release have elevated the priority of this expansion activity.

Mother-infant linkage: Evaluating the safety of medications and vaccines in pregnant women requires linkage between mothers and infants. The recommendation leverages an ongoing collaborative pilot research program, the Medication Use in Pregnancy and Birth Outcomes Program (MEPREP), between the FDA and researchers at the HMO Research Network Center for Education and Research in Therapeutics (CERT), Kaiser Permanente Northern and Southern California, and Vanderbilt University. An ongoing Mini-Sentinel workgroup activity is underway to address the feasibility of obtaining information from select states for Post-Licensure Rapid Immunization Safety Monitoring (PRISM) Data Partners.

Refinements to Current MSCDM Tables: A flag that indicates whether data are available for chart abstraction will improve the efficiency of validation studies, patient 5-digit ZIP code will facilitate the incorporation of ZIP code-based measures of sociodemographic factors as confounders, and the UDI will facilitate device-based evaluations. Data Partners advised against the incorporation of plan indicators or primary/secondary insurance indicators. Standardizing plan definitions would be time-consuming and definitions change over time. Similarly, the determination of primary vs. secondary insurance is often

made on a patient-by-patient basis. Availability and reliability of physician specialty across Data Partners should be established before its inclusion in the MSCDM.

D. LESSONS LEARNED AND SUGGESTIONS FOR FUTURE WORK

Year Four witnessed substantial progress in the expansion of MSCDM to include clinical data elements and the identification of future expansion priorities. A summary of lessons learned related to MSCDM expansion and suggestions for future work follows.

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 for analysis. To that end, MSOC will need to continue to develop “user’s guides” and provide education about the strengths and weaknesses of the clinical laboratory and vital signs data captured, the contents of the tables, what harmonization has been done, which laboratory test types are available from all (or only a subset of) participating Data Partners, which Data Partners have more comprehensive laboratory data capture, and sub-populations with clinical data available for use in Mini-Sentinel. Such information will enable other Mini-Sentinel teams and workgroups to make informed use of these clinical data tables.

2. Expansion of the MSCDM

Articulating specific exposures and health outcomes of interest proved to be a useful approach to identifying expansion priorities for the MSCDM. Important gaps in inpatient data capture emerged as did opportunities to leverage existing internal and external data sources through linkage. As noted earlier, recommendations for future work include the following:

- Partner with a national hospital-based organization to expand access to inpatient data streams.
- Enhance supplemental death data through linkages with the National Death Index.
- Leverage an ongoing collaborative pilot research program, the Medication Use in Pregnancy and Birth Outcomes Program (MEPREP), to establish data linkages between mothers and infants.

IV. MINI-SENTINEL DISTRIBUTED DATABASE

A. DATA QUALITY ASSURANCE REVIEW AND CHARACTERIZATION

1. Overview

All data transformed by the Data Partners into the MSCDM are checked via standard data quality assurance (QA) review and characterization programs developed by the MS Scientific Operations Center and refined through feedback from the Data Partners. These programs are often referred to as the “data checking” programs. Data Partners run the programs on their local implementation of the MSCDM after each data “refresh” and provide MSOC with the summary data checking output. For each data refresh the Data Partner performs an Extract-Transform-Load (ETL) process to update their implementation of

the MSDD. The ETL process is described in detail in Section III of our [Year One Common Data Model Report](#).

2. Roles and Responsibilities

The MSDD and Communications Group lead the data review process. The MSDD and Communications Group reviews the data checking output, documents the findings from the data, identifies data issues that require discussion or documentation, and communicates with the Data Partner to determine next steps. The next steps could include approval of the refresh, approval of the refresh with specification for corrections to be made during the next refresh, or rejection of the refresh which would require a revised ETL and complete re-review. The specific steps included in the refresh process are described in [Mini-Sentinel Standard Operating Procedure Data Quality Checking and Profiling](#) and have not changed in Year Four. The SOP includes the following high-level steps:

- Data Partner implements their local ETL process
- Data Partner executes data QA review and characterization programs
- Data Partner reviews data QA review and characterization output, revises ETL as necessary, and re-runs data QA review and characterization programs
- MSDD and Communications Group reviews data QA and characterization output, within and across ETLs for the Data Partner
- MSDD and Communications Group provides data QA and characterization report to Data Partner for review and comment
- MSDD and Communications Group and Data Partner review and discuss data QA review and characterization report, agreeing to any necessary changes and their timeline
- MSDD and Communications Group approves the ETL

Once the ETL is approved, the Data Partner executes the Summary Tables program and updates their Summary Tables to enable use by the Mini-Sentinel Query Tool. The Data Partner is required to run a “metadata refresh dates” query on the Query Tool to inform MSOC that the data are ready for querying. This is a query that the Data Partner submits and runs on its own data. It provides information to the MSOC on the dates available for each query type from 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 MSDD and Communications 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 to characterize the breadth and depth of the data available for querying. These measures address areas such as missing data, invalid values, invalid date ranges, and internal inconsistencies. The design and the scope of the data QA review and characterization programs must balance expected variability across Data Partners, based on the way partners access and use administrative and claims data and electronic health record data, with the need to ensure that the MSDD tables match the defined Mini-Sentinel requirements.

The data QA review and characterization programs are run after each data refresh. The data quality activities are organized into four levels of data characterization, based on the type of checks being

performed. A description of the data characterization approach and the findings accompanies this report and can be found under the Data tab of the Mini-Sentinel website in a separate document titled [Mini-Sentinel Year 1 Data Quality and Characterization Procedures and Findings](#).

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 the required variables contain data and conform to the formats specified by the MSCDM data dictionary. For each MSCDM variable, data QA review 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. Additionally, the date of birth 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 data dictionary. Table 1 illustrates several of the Level 1 data QA review and characterization items for the dispensing table.

Table 1. 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 unenrolled members. Table 2 illustrates several of the Level 2 data QA review characterization items for the enrollment table.

Table 2. 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. All 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 have an expected 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. Periods of sharp increases or decreases are also unexpected. This profiling characterizes specific data variables for each Data Partner and aggregates information for cross-institutional comparisons. The Level 3 data

characterizations also evaluate trends to help identify data gaps and unusual patterns both within an ETL and across Data Partners' ETLs. Examples of trends within a single ETL include:

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

Examples of trends across ETLs include number of members and number of records—both of which are expected to increase with each ETL and with the addition of new data. Other Level 3 data characterization topics include counts of procedures per encounter by encounter type and year and diagnoses per encounter by encounter type and year. This approach has been used successfully by the HMO Research Network, the Vaccine Safety Datalink, and other distributed networks to identify issues within their distributed databases.

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

- Overall table statistics
 - Number of records in the table
 - Number of unique PatIDs
- 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 addressing unexplained variability. We note that this level of data check is not intended to find all data anomalies, but rather to assess metrics that can be readily checked and flagged for explanation. Detailed, topic-specific data checking is required for every Mini-Sentinel query as review of specific data areas or patient cohorts may uncover anomalies not identified in the initial data checking activities.

d. Level 4 Data QA Review and Characterization

In Year Four, MS Scientific Operations Center developed five new Level 4 data checks to 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 new checks inspect:

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

These Level 4 checks are included in the v3.1 release of the data QA review and characterization programs.

4. Data QA Review and Characterization Revisions

MSOC released three new versions of the data QA review and characterization programs during Year Four, to incorporate feedback from Data Partners and expand our knowledge of the MSDD. Version 3.0 was released in November 2012 and included the initial data quality review programs for the MSCDM laboratory and vitals tables. This release also gave MSOC visibility into the diagnosis and procedure code lists for each Data Partner, expanded the PatID and EncounterID matching data checks, added several Level 2 date comparison checks, and fixed several reported bugs in the programs. Version 3.0.1 was released in March 2013 and included minor changes for UNIX operating system compatibility. Version 3.1 was released in July 2013 and included new Level 4 data checks, improvements to summarizing the Level 1 and Level 2 data check results especially for the laboratory data, information about the distribution of discharge date in the Encounter table, new dataset names for the laboratory and vitals data checking output, and other minor changes and bug fixes. 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 the 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](#) and the [Mini-Sentinel Distributed Database Summary Report](#)—provide details of the data QA review and characterization activities and results across all Data Partners.

6. Data Completeness and Availability

In Year Four, data QA review and characterization activities were expanded to generate a set of data availability and data completeness reports after every approved data refresh. These reports were initially created as part of the Task Order 7 Drug Utilization project. 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 that Data Partner. Updated reports not posted on the public website, but are shared with the FDA on a regular basis.

7. Principal Diagnosis Flag (PDX) Variable Investigation

In response to several queries by 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 will be reported in the Fall of 2013 – Year Five. The investigation will help guide use of the PDX variable and may lead to changes in how the Data Partners populate the variable in future data refreshes.

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 two 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)ⁱⁱⁱ 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 structure data model allows for inclusion of ICD-10 or any other diagnosis coding terminology.

Procedures: Medical procedures are captured using ICD-9 procedure codes and *Healthcare Common Procedure Coding System* (HCPCS)^{iv} codes, including Current Procedural Terminology-4 (CPT-4)^v 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 structure 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.

In addition, 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 have been adopted for vaccine-specific work involving immunization registries. 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.^{vi}

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^{vii} diagnoses codes. These are the codes available through the source of the information, typically State death registries.

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,

ⁱⁱⁱ <http://www.cdc.gov/nchs/icd/icd9cm.htm>

^{iv} <http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html>

^v <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.page?>

^{vi} <http://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx>

^{vii} International Classification of Diseases, 10th Revision; <http://www.cdc.gov/nchs/icd/icd10.htm>

'+', '++', '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.

Some commonly referenced controlled terminologies such as RxNorm, and the Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT) are not currently included in the MSCDM. Although these and several other potential relevant controlled terminologies are increasingly being adopted by electronic health record systems and some health plans providers for transmission of clinical information, the Mini-Sentinel Data Partners do not uniformly capture information using those 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^{viii} 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^{ix} as a key binding point for engagement related to Mini-Sentinel data standards, specifically the **ONC Query Health Initiative**. Several members of the MSOC staff, and associated vendors, are actively engaged with the S&I Framework activities, especially the S&I Framework Query Health Technical and Clinical Workgroups, and will remain engaged with those activities. In addition, MSOC and FDA recently completed participation in 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

^{viii} <http://www.healthit.gov/policy-researchers-implementers/meaningful-use>

^{ix} <http://www.siframework.org/>

Format (HQMF). The goals of the pilot were to 1) assess, adopt and implement the ONC Query Health meta-data standards for the Mini-Sentinel Query Envelope used by the Mini-Sentinel Distributed Query Tool, 2) beta-test an upgrade of the Mini-Sentinel Distributed Query Tool with PopMedNet Version 3.0 which is consistent with current Query Health standards for distributed querying, 3) incorporate the i2b2 HQMF query adapter into the PopMedNet architecture, and 4) work with Beth Israel Deaconess Medical Center (BIDMC) in Boston to pilot end-to-end querying using the i2b2 HQMF adapter with the existing BIDMC i2b2 installation. In addition to the Query Health pilot project, MSOC involvement has included face-to-face meetings with S&I Framework staff, webinars, and participation on several working groups.

The Query Health pilot was successfully completed during the year. The pilot illustrated a full integration of i2b2 into the PopMedNet infrastructure to enable a secure distributed query to a live i2b2 node at BIDMC. A video of the integration^x and a related poster presentation^{xi} are available online. As part of the pilot project, we contributed 2 new “cells” on the publically available i2b2 hive.

Finally, as part of the effort to adopt and promulgate the ONC Query Health standards, we successfully transitioned the Mini-Sentinel Distributed Query Tool to PopMedNet version 3.2. This complex transition from the prior PopMedNet platform involved upgrading the PopMedNet software to adhere to the ONC Query Health distributed querying standards and upgrading the entire network with the new software platform. The Mini-Sentinel Distributed Querying Tool adheres to the ONC standards and guidelines for secure distributed querying.

3. Impact of Transition to ICD-10

As mentioned above, the existing MSCDM and the existing modular programs can accommodate ICD-10 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, ICD-10 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 will require work on developing new HOI algorithms or validating mappings between ICD-9 and ICD-10 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.

^x Available at: <http://www.youtube.com/watch?v=sqDAo6E-b1o> [July 17, 2013]

^{xi} Available at: http://www.popmednet.org/?page_id=39 [July 17, 2013]

C. LESSONS LEARNED

MSOC continues to increase the scale, effectiveness, and timing of our data core activities. In Year Four, we completed 50 data refreshes and added the review of laboratory data to this process. Selected lessons learned in Year Four are detailed below.

Data Changes and Quality Improvements: Two Data Partners truncated their MSCDM data in Year Four. Although the changes led to a net decrease in the overall number of unique patient identifiers in the MSDD, they also led to a net increase in the quality and breadth of that data. One large Data Partner transitioned to a new data warehouse and moved from a member-centric structure to a new patient-centric structure. Members who enroll in different health plans at different times will now have the same patient identifier. This increases our ability to track patients over time. A small Data Partner truncated 1.5 years of legacy data that fluctuated wildly in quality, was most likely incomplete, and was inaccessible for follow-up investigation. These kinds of local changes in source data, data warehouse upgrades, and data platform improvements are expected throughout the Mini-Sentinel project. Early and effective communication between FDA, MSOC, the Data Partners, and other Mini-Sentinel data requesters allows for a smooth transition when changes occur.

Database Changes and Stability Improvements: At the beginning of Year Four, there were still a few Data Partners who were using “views” (i.e., virtual tables) of their source data in MSCDM format instead of physical tables (i.e., datasets). Using views saves storage space because it obviates the need to create separate physical files in MSCDM format for the entire membership. Views include the most recent data in the source database, including data that have not been checked and approved by MSOC. The data in a view is dynamic while the data in a dataset (permanent table) is static. As a result when a Data Partner uses a view to respond to a request, especially a re-run of an earlier request, it can be difficult to understand the results because the source data have changed. During Year Four, MSOC strongly encouraged all Data Partners to transition to using datasets instead of views for their MSCDM data. By mid-2013 the three Data Partners who were still using views successfully modified their ETL processes to switch to MSCDM datasets. This change improves the stability of the MSDD and ensures that no unapproved data are used to respond to Mini-Sentinel requests.

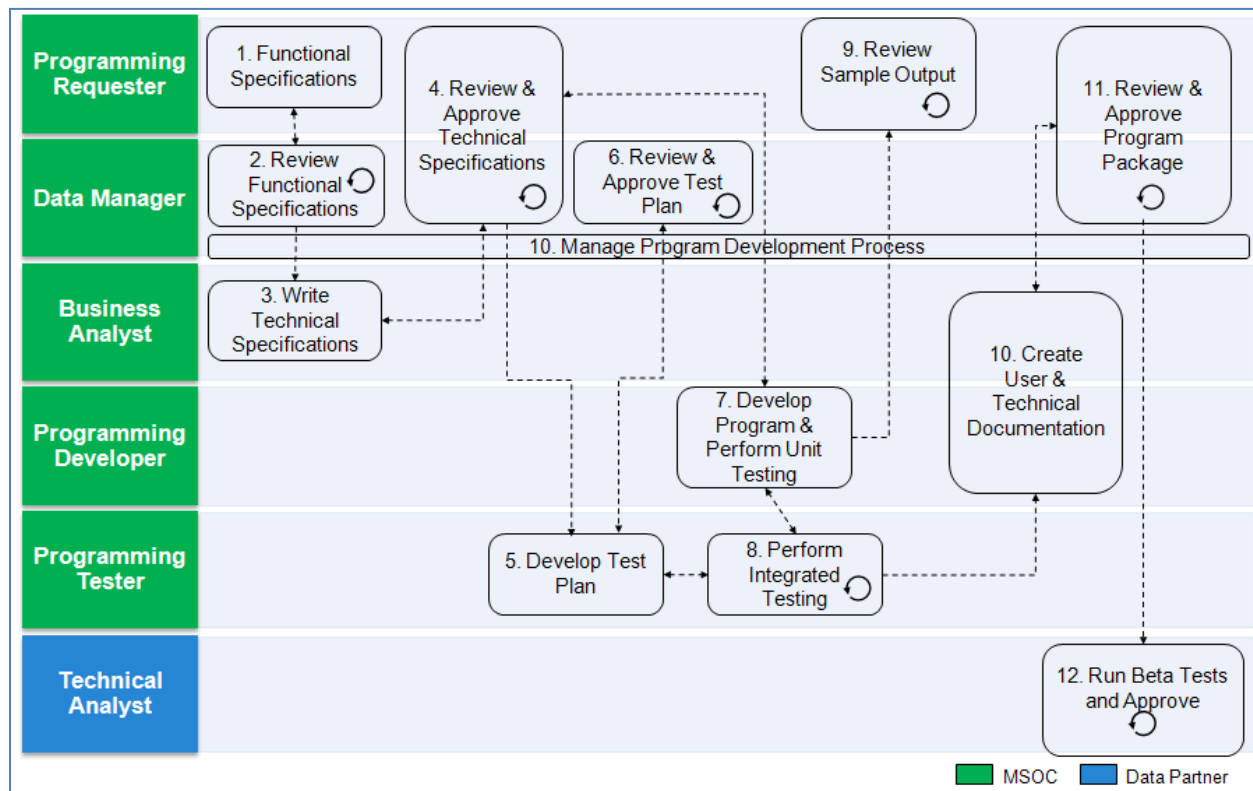
V. MINI-SENTINEL ANALYTIC TOOLS

A. MODULAR PROGRAMS

1. Overview

As part of the Year Four activities, the Data Infrastructure Group implemented several enhancements and additions to the pool of existing modular programs (MPs). By end of Year Four, a total of six MPs were available to facilitate rapid response to common FDA queries by Data Partners. All MPs, whether they are enhancements from previous versions or new additions, went through the Mini-Sentinel Query Fulfillment process (see Figure 4). Plans for future enhancements to MPs will be determined jointly FDA.

Figure 4: SAS Program Development Process Flow



Each of these programs has several required input parameters (e.g., exposures or outcomes), and the output contains summary-level counts (e.g., number of members with an incident exposure to a drug, number of members with a specific diagnosis/condition, at-risk populations) stratified by various parameters (e.g., age group, sex, year). All programs and documentation are posted on the [Data Activities section](#) of the Mini-Sentinel website, when available. Revised input forms are shared with FDA for production data queries.

[Section V.A.4](#) below (Modular Program Revisions) summarizes the various enhancements and additions implemented during Year Four and includes a short description of all six modular programs available.

2. Roles and Responsibilities

The development and revision of Mini-Sentinel MPs require careful planning for use of available internal and external resources (i.e., timing and effort). The Data Infrastructure Group is solely responsible for maintaining the MP development process in good standing. During Year Four, both internal and external resources were used for SAS programming, testing, and Quality Compliance (QC) efforts. Data Partners supported the Data Infrastructure Group with the validation of all enhanced and new MPs. The roles and responsibilities of each group are described below.

Data Infrastructure Group:

- Prepare MP development plan (what feature and module to add and when)

- Identify new features of potential interest to FDA and various workgroups
- Assess feasibility of new features, modules and programs requested by FDA or workgroups
- Prepare specification and QC documents
- Ensure compliance with the Mini-Sentinel Program Development SOP
- Coordinate exchange of information between FDA, MS Lead Team, Data Core, Data Partners, and External Programmers
- Update Data Core and Data Partner on status of MP development
- Hold training webinars for FDA, Data Core, and Data Partners
- Keep documentation and input forms up-to-date; share with FDA as needed

External Programmers:

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

Data Partners:

- Test and validate each new release of MPs
- Provide feedback on efficiency and functionality of the MPs

3. Modular Program Revisions

Year Four modular program revisions included: 1) enhancements of the existing programs with new features and capabilities; and 2) modularizing existing code to add to the pool of modular programs. All revisions and additions were implemented in three waves of development. Each wave included between five and ten items to implement, and represented one complete cycle through the Mini-Sentinel Program Development SOP process (i.e., from specification to final release of the code).

a. Enhancements to Modular Programs

Enhancements included: 1) features requested or approved by FDA; and 2) several revisions to solve issues and inconsistencies reported by Data Partners, FDA, the MSOC Production team, or external programmers. Summary descriptions of each enhancement are listed below.

Attrition table: For **MP3** and **MP9** only, a new table summarizing how many observations (e.g., members, treatment episodes, dispensings) were excluded during various stages of the MP cohort selection algorithm, allowing the user to see how a cohort evolves after successively applying these MP filters.

Wildcard and exact medical code match: Previous versions of all MPs identified HOIs in the MSDD by selecting records that included a medical code (i.e., diagnosis or procedure) that started with the code specified by the user in the input files -- a method known to potentially include undesired cases. For example, querying ICD-9-CM diagnosis code 250.0 (diabetes mellitus without mention of complication) would allow inclusion of records with: 1) exactly code 250.0 specified; 2) any of the four valid sub-codes such as 250.00 (type II, not stated as uncontrolled), 250.01 (type I, not stated as uncontrolled), etc; or 3) invalid sub-codes due to data entry errors. Selection of the correct cohort thus required the user to pass

the MP a comprehensive and accurate set of valid codes. With the addition of the wildcard and exact code match, the user gains more control over the specific record selected by code. For example, if records with exactly ICD-9-CM code 250.0 can be selected, the user can specify so using the exact code match feature.

User-defined coverage type: Previous versions of all MPs were inconsistent in the way type of coverage (i.e., medical coverage only, pharmacy only, or both medical and pharmacy) was enforced. For some MPs, the default was to require that valid members be enrolled to both medical and pharmacy coverage, whereas with other MPs coverage was determined based on type of input codes were involved in the query. For instance, a query on outpatient pharmacy dispensing specified with only National Drug Codes would only require pharmacy coverage and not medical coverage. Due to varying baseline and demographic characteristics of certain populations affected by this type of coverage (e.g., Medicare Part D eligible members), query results could be biased and therefore not relevant to requester. With the addition of this new feature, users can specify what coverage types are required irrespective of the type of input codes involved in the query.

Allow all MPs to only consider first valid instance or episode: This feature, that was previously available in some MPs, is now available in all. It allows for assessment of risk of events or characterization of use and reports metrics on the very first valid incident exposure or treatment episode identified by the MP algorithm. This feature is optional and needs to be turned on by the user.

Flexible interaction between MSCDM Principal Diagnosis Flag and Care Setting fields: Prior versions of MPs, requiring the use of MSCDM principal diagnosis flag (i.e., PDX), did not allow selection of records with PDX values other than Primary vs. any other value, nor did it allow the selection of records with encounter types other than hospital inpatient or emergency department. With the addition of this new feature, the user has total flexibility in terms of what PDX value(s) and care setting(s) can be selected by the MP algorithm. For instance, if a user wishes to select all records with PDX set to primary, irrespective of the encounter type, it is possible to do so.

Add Amount Supplied to MP Output Tables: Up until the end of Year Three, all modular programs reporting on use of outpatient pharmacy dispensings only included metrics on days of supply as a measure of how many days members were exposed to drugs of interest. With the addition of this new default feature, actual amount of supply (e.g., “the number of pills”) is also reported.

Allow MP3 to consider multiple events per treatment episode: Previous versions of MP3 allowed identification of only one adverse event per treatment episode, whether multiple events could be found in the data or not (i.e., the MP algorithm would stop searching for additional events after one valid event was found). This enhancement enables the user to either only count the first event or to count multiple.

Change in Body Mass Index (BMI) Calculation Tool: For MP3, new functionality was added to calculate the change in BMI z-score for cohort members in a user-defined baseline and follow-up period. An extra table outputs cohort metrics (e.g., new users, treatment episodes and events) stratified by range of change in BMI z-score. This module can only be used for cohorts where members are aged 2-19 years.

Laboratory Querying Feature: New functionality was added to MP6 to enable querying of the Laboratory Results Table. The occurrence of a laboratory test can be used as: 1) the index event; or 2)

post- event treatment. An additional output table, which stratifies the cohort by the number of resulted lab tests performed per member during the lookup period, is generated if laboratory results are used to define post-event treatment. Users can additionally determine if labs occurring more than once on the same day are counted more than once.

b. Conversion of Existing Code into Modular Programs (New Modular Programs)

Modular Program 8: As part of CDER Task Order #7 (*Mini-Sentinel Operations Center Response to Potential Exposure/Outcome Associations*), MSOC had developed a program for the **Drug Use Studies project (Comparison to Nationally Projected Databases)**. This program aims to assess uptake and persistence patterns for New Molecular Entities (NMEs). This program has been modularized and an input form for it was created to allow FDA requesters to use it for routine production queries.

Modular Program 9: Since MP1 (i.e., medication/procedure use), MP2 (i.e., medication/procedure use among those with a specific condition), and MP5 (i.e., background rate of health outcomes of interest) shared so many common characteristics and were similar in nature, the Data Infrastructure Group decided to bundle all three into one program, thus saving on future maintenance work, documentation, and support. All features available in any one of the three programs were preserved.

c. Summary of the Six Available Modular Programs

Modular Program 3 (incident medication/procedure use and outcomes): Evaluates the rate of specified outcomes (defined by ICD-9-CM diagnosis codes, procedure codes, or medication dispensings) among those with incident exposure to medications, procedures or diagnoses, with or without a pre-existing condition defined by ICD-9-CM diagnosis codes or procedure codes (ICD-9-CM or HCPCS). **For example:** Rate of stroke during exposure to an antidiabetic medication among new users of the medication who also had a prior diabetes diagnosis.

Modular Program 4 (concomitant medication/procedure use): Characterizes concomitant use of specified products or groups of products (defined by National Drug Codes (NDC)) dispensed in the outpatient pharmacy setting or procedures/diagnoses recorded in any setting, among those with incident use of specified products or procedures/diagnoses with or without a pre-existing condition, defined by ICD-9-CM diagnosis codes or procedures codes (ICD-9-CM or HCPCS). **For example:** Characterization concomitant use of atypical antipsychotic drugs and selective serotonin reuptake inhibitors among those with a diagnosis of depression.

Modular Program 6 (medication/procedure use following a diagnosis): Provides rate of medication/procedure use among at-risk, diagnosed populations, as well as metrics on time to first medication/procedure use from diagnosis index date. Optional features include: ability to restrict to incident diagnosis and/or naïve-to-treatment (i.e., medication and/or procedure) patients, and ability to add pre-existing conditions. **For example:** rate of oral antidiabetic medication use following first diagnosis of diabetes; rate of hip replacement surgeries following a fall at home among female patients aged 65+ with osteoporosis.

Modular Program 7 (most frequently used codes prior & post index event): Characterization of 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 of interest 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 codes are provided. **For example:** Top 10 generic drug names observed in the 30 days before and after a heart transplant.

Modular Program 8 (drug utilization, uptake rate, and persistence): Characterization of the use of New Molecular Entities (NMEs) and assessment of uptake and persistence patterns. New use of each NME can be parameterized with user-defined 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, second, etc), gap (in days) between valid treatment episodes, survival analysis. **For example:** assess uptake rate of a newly approved antidiabetic medication.

Modular Program 9 (background rate and characterization of health outcomes of interest among individuals with or without conditions of interest): Characterizes the use (via prevalence and incidence rates) of specified diagnoses, procedures, or outpatient medication dispensing among at-risk populations. **For example:** Use of asthma medications among those with an asthma diagnosis by age group, sex, and year; use of anti-TNF agents among those with a psoriasis diagnosis; prevalence and incidence rates of type 2 diabetes stratified by age groups, sex, and year.

4. Other Modules

The Data Infrastructure Group developed several new programming capabilities during Year Four. These new features were selected based on feedback from FDA and the MSOC Production Group. Each new capability is listed below.

Daily Dosing Analysis: A simple, optional module for **MP9** to characterize daily dosing patterns of cohorts exposed to specified outpatient pharmacy dispensing. For each valid dispensing identified by the MP algorithm, the module utilizes the dose information specific to each NDC code and the days supply to calculate a daily dose. All dispensings are then characterized by range of daily dosing and treatment patterns for all exposed members are reported (i.e., what types of members are exposed to what daily doses).

Combination Tool: Defining health outcomes of interest and medical product exposures sometime require complex algorithms beyond the capabilities of the Modular Programs. For example, an exposure of interest could be defined as triple therapy (e.g., for treatment of Hepatitis C), as a medication dispensing preceded by at least two doctor visits or one hospitalization, or a combination of a treatment and a surgical procedure. To address this limitation, the Data Infrastructure Group is developing a “combination tool” (to be released in Fall 2013 – Year Five) that will allow any modular programs or workgroup/evaluation projects to study combinations of events by combining multiple items or MSCDM variables (e.g., NDC, ICD-9-CM, HCPCS, Encounter Type, Number of Visits, etc) into HOI concepts or exposures.

Incidence Risk Ratio (IRR): A standalone tool for use with **MP3** to automate the comparison of two cohorts and their incidence rates. Cohorts are identified by the user specified MP3 input codes. The tool utilizes the output from the execution of MP3 and generates both the crude and adjusted incidence rate ratios for the two cohorts by producing the incidence rate ratio estimates and their corresponding 95% confidence intervals. The user also has the capability to control for age, sex, year and Data Partner

within the adjusted rate ratio calculations. The IRR tool utilizes a Poisson regression and a large sample approximation for calculation of the IRR, and thus may not be robust against samples with small event rates. The tool is being extended to control for additional MP3 stratifiers such as comorbidity score.

Event identification module added to MP4: The original version of **MP4** was used to characterize the 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. FDA requested an enhanced version of MP4 that included the option to characterize the frequency of select events(s) during episodes of concomitant use (similar to MP3). To achieve this, the event functionality of MP3 was enhanced and added to MP4. In addition, the way primary exposure, secondary exposure, and concomitant exposure are defined was also enhanced. The enhanced MP4 now outputs metrics for three cohorts in each execution of MP4: 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. Several additional parameters were also added to allow increased flexibility in the definition of concomitant exposure.

5. Beta-testing

Each revision of modular program or summary table code follows the Mini-Sentinel Program Development SOP, and is therefore beta-tested by at least two Data Partners before being distributed to every other Data Partners. Moreover, for each major release of a new or enhanced MP, all Data Partners are required to validate the new features by running a generic request using known and non-controversial scenarios. Doing so allows the Data Infrastructure Group to ensure that newly developed and released MPs can be run efficiently at all Data Partner sites, thus speeding up the query process for FDA requesters.

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 queries 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 Mini-Sentinel distributed database. 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 nine 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), drug exposures (ingredient name and drug category), and enrollment. The three incidence summary tables represent incidence counts of diagnoses (3-digit ICD-9-CM) and drug exposures (ingredient name and drug category). 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 OptumInsight, 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 (added in Year Four):

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, and 3-digit ICD-9-CM code.

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 Data 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 programmers are also responsible for keeping all lookup tables up to date. These are lists of all NDCs, diagnosis codes, procedure codes, and HCPCS (provided by Ingenix, Inc.) that include a text description of each code. Most recent lookup tables are sent to Data Partners with the package to generate summary tables. They provide a crosswalk between code (which appears in each Data Partner's MSDD) and description so that descriptions appear in the Query Tool.

MSOC staff are responsible for sending the SAS program and lookup tables as part of a package to each Data Partner after each new data refresh has been approved. 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 on all 12 summary tables in some

way. Data Partners are responsible for running the 16 test queries, reviewing the output, and uploading results. Finally, MSOC staff examine test query results, follow-up with the Data Partner about any unexpected results, and approve when appropriate so that 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.

FDA regularly submits summary table requests. The Production Group manager logs the request in the request tracker and assigns it an identification number and an analyst responsible for completing the request. This analyst works with the requester as needed to address any potential issues and finalize specifications for the request. Queries are then 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 production manager and often an epidemiologist, before being sent to the requester. MSOC staff are available to answer any questions about the report.

3. Summary Table Revisions

During Year Three, the Data Infrastructure Group implemented major revisions to simplify creation of the summary tables that are used for rapid querying via the Mini-Sentinel Query Tool. There is now a single distributed program with nested macros that improve efficiency through re-use of intermediate files for multiple purposes. MSOC also developed in Year Three and implemented in Year Four a new SAS program for the creation of summary tables for incident counts (events and members) for three different types of outcomes: (1) incident outcome by 3-digit ICD-9-CM diagnosis code; (2) incident exposure by generic name; and (3) incident exposure by drug category.

During Year Four, the summary table changes made in Year Three were implemented. The Data Infrastructure Group finalized this program and sent it to all Data Partners, and worked with Data Partners to ensure that Summary Tables were created correctly. Currently, all but one Mini-Sentinel Data Partners have the capability of responding to prevalence, incidence, and most frequent utilization queries. The Data Infrastructure Group also made some efficiency-related improvements, based on Data Partner feedback. This version is currently being quality-tested and is expected to be rolled out to Data Partners in September, 2013.

C. MINI-SENTINEL DISTRIBUTED QUERY TOOL

1. Overview of Query Tool

The FDA Mini-Sentinel Distributed Query Tool allows the Production Group staff to create and securely distribute “queries” to Data Partners and enables Data Partners to review, execute, and securely return the results of those queries. The system allows different levels of query automation that can be set at the discretion of the Data Partners. The network is hosted in a private cloud environment in a Federal Information Security Management Act of 2002 (FISMA)^{xii} compliant TIER III data center. The Mini-Sentinel Query Tool is based on the PopMedNet™ software platform. The implementation design and architecture are detailed in the Mini-Sentinel [Distributed Query Tool: Overview and Administrators Guide](#).

^{xii} <http://csrc.nist.gov/groups/SMA/fisma/index.html>

The Mini-Sentinel Distributed Query Tool (see screenshot of the login screen in Figure 5) currently allows rapid distributed querying of preprocessed summary tables. Using preprocessed summary tables speeds the querying process because it:

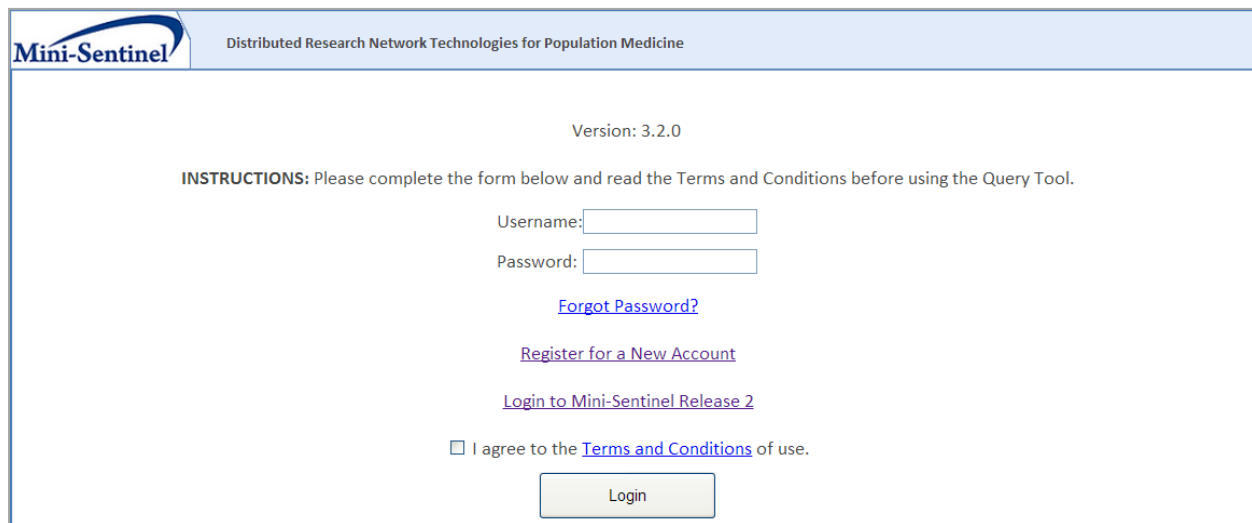
- Obviates the need to access person-level data, thereby avoiding local privacy and patient–confidential, data-release authorization procedures
- Allows use of a simple menu-driven querying tool interface
- Allows nontechnical Data Partner staff to execute and return results
- Avoids the need to specify, create, and validate new SAS programming codes to answer simple questions

The expected response time for these queries is two business days. The system includes three broad query types: prevalent queries, incident queries, and most frequent utilization queries. The nine prevalence queries represent prevalence counts of diagnoses (3-, 4-, and 5-digit ICD9-CM), procedures (3- and 4-digit ICD-9 and HCPCS), drug exposures (ingredient name and drug category), and enrollment. The incident queries represent diagnoses (3-digit ICD-9-CM) and drug exposures (ingredient name and drug category). For diagnoses and procedures, the system generates rates per 1000 enrollees, events per 1000 enrollees, and the number of events per person. For drug queries, the system generates users per 1000 enrollees, dispensings per 1000 enrollees, days supplied per dispensing, and dispensings per user. The tables also include the number of enrolled days per year by age group and sex to enable more precise calculation of prevalent rates. The most frequent utilization queries return the most frequently observed utilization (drug exposures, diagnoses, or procedures) defined by events or number of users by age group, sex, and year. The Mini-Sentinel [Distributed Query Tool Investigator Manual](#), a description of the [Mini-Sentinel Summary Tables](#), and additional documentation are available on the Mini-Sentinel website and have additional details on the summary tables and a description of how to create and distribute queries.

The Mini-Sentinel Distributed Query Tool architecture is consistent with the standards promulgated by the Standards and Interoperability (S&I) Framework supported by ONC. Mini-Sentinel staff is working 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 Mini-Sentinel query tool system. The selected data source was i2b2 (Informatics for Integrating Biology and the Bedside) – a widely used data repository and analysis platform. We 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.

Through this engagement we continue to communicate the lessons learned from implementation and operation of the Mini-Sentinel distributed querying system. These lessons include the need for detailed technical documentation and user training material, the need for security documentation and clearance by each Data Partner, and barriers faced related to installation of external software on local computers.

Figure 5. Distributed Query Tool Login Page



2. Network Implementation

The distributed querying network was established in partnership with MSOC, Mini-Sentinel information technology vendors, and the Data Partners. The implementation process involved establishment of a “staging” network that allowed testing of governance, security, and querying capabilities of the software platform, development of a series of user manuals, and implementation of a production site to allow secure distribution of queries. Use of the system has led to several revisions and enhancements. During Year Four, query tool enhancements focused on improving system architecture and functionality, better alignment with national querying standards as developed by ONC, and improving workflow for MSOC and Data Partners.

3. Platform Enhancements for Mini-Sentinel Query Tool Version 3.2

The Mini-Sentinel Query Tool software platform undergoes ongoing annual maintenance to improve the software platform 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 also were made to better align our infrastructure with national querying standards described by the ONC S&I Framework Query Health Initiative. The specific annual enhancements to the technical architecture have been implemented to allow for a broader and more efficient use of the Query Tool software to improve:

Maintainability: Ongoing platform upgrades are necessary to maintain the Query Tool and improve its efficiency and sustainability as the system activity grows and it is used to distribute more queries.

Enhancements: The upgrades allow modularization of enhancements using a plug-in design.

Scalability: The Query Tool can cultivate and support new networks, projects and users.

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

The specific platform re-architecture enhancements adopted by the Mini-Sentinel Query Tool are outlined below:

.NET 4 Framework: An application platform that is comprised of common language runtime and class library features, providing higher efficiency for overall code management and updates.

Entity Framework: A type of object-relational mapping used as part of the new Query Tool platform. As part of this framework, an Entity Manager view, Data Access Layer, and Structure Business Layer Class were developed.

Common Controls: Information that is presented on the Query Tool portal, including functions that present the user with information in grids and lists.

Complex Controls: Controls used to perform functions based on a specialized set of data, such as the controls for roles. Specifically, the Query Tool is able to understand all the information associated with the identified defined roles.

PopMedNet Library: A library within the software platform that contains a set of helper and utility functions and services that are used across the entire Query Tool application.

Hub Background Service: Services that are outside the application and are essential to keep the application running.

User Interface (UI): The entire UI for the Query Tool has been enhanced as part of the platform work. The UI contains all the graphical and textual information that the Query Tool presents to the user. The main function of the UI is to translate tasks and results into a format that the user can understand as they navigate through the system. Examples of new UI improvements include:

- **New Menu Layout:** Menu names have changed to streamline navigation throughout the Query Tool. For example, menu tabs include pages for Home, Requests, Profile, Resources, Reports, and Network.
- **New Code Selector Controls:** A pop-up window was added for code selection and improved search functionality with codes and wildcards.
- **Newly Designed UI Buttons** throughout the Query Tool
- **New Master Page Template and Home Page Layout**
- **New Request (Query) Summary Page:** New streamlined format for submitting query requests.
- **New Request (Query) Result Detail Page:** Data Partners have a new enhanced view of their query request results.
- **New Request (Query) Status Page:** Data Partners can view a list of their query request statuses on the Query Tool Portal.
- **New Response Page:** Updated DataMart Client for Data Partners to view their workflow of outstanding or completed query requests.
- **New Model Administration Page:** Data Partners may have access to multiple types of data models and query types.

- New Profile Page: Easily accessible user settings pages for Data Partners to administer their user profile.
- New DataMart Administration Page: New Administration page for Data Partners to manage access and rights and download the latest version of the DataMart Client software.

Business Layer: Incorporates and implements the business logic, located between the data access layer and the user interface which coordinates the application, processes commands, makes logical decisions, and performs calculations.

- Enhanced Access Controls
- Event Manager
- Authentication
- Code Event Logger
- Business Rules
- Notification Manager
- Request Scheduling Manager and Request State
- Web Services
- File Distribution or File Transfer feature

Model Adapter Construction: The platform upgrade includes the newly designed concept for Model Adapters. This feature abstracts the request implementation from the system platform into a “Model Plug-in”. This new type of architecture separates the concerns of the network platform from the details of the requests (i.e., queries) that travel through it. The result is a network that forms a tunnel through which requests and responses travel.

- **Summary Table Model Adapter:** The Mini-Sentinel summary query functionality has been adapted to the model plug-in software for the 3.0 platform for all three query types: Prevalent, Incident, and Most Frequent Utilization Queries.
- **i2b2 Model Adapter:** Creation of an i2b2/PopMedNet plug-in adapter that allows for use of the i2b2 Query Composer to construct queries that can be executed against i2b2 data sources within the Mini-Sentinel platform. The design and development of a PopMedNet/i2b2 Model adapter currently fulfills the standards for the FDA to participate in the ONC Pilot Program. A video details the ONC pilot (<http://www.youtube.com/watch?v=sqDAo6E-b1o&feature=youtu.be>) and a poster also was presented (http://www.popmednet.org/wp-content/themes/twentyeleven/documents/jklann-AMIA_2012_Poster_submit_pdf.pdf).
- **Modular Program Adapter:** (In development) This adapter aims to improve the workflow process by running multiple Mini-Sentinel projects simultaneously through the Query Tool. This query type is based on the File Distribution/SAS query types where an investigator has the ability to enter zero, one, or more individual modular programs, documents, and/or text strings representing Secure FTP Portal links.
- **Request Metadata Plug-in:** This new request type will give MSOC the ability to query the metadata of previously submitted Mini-Sentinel queries from within the Query Tool.

4. Deploying Platform Enhancements to the Data Partners

Formal testing of system upgrades involved extensive testing by MSOC and the Mini-Sentinel IT vendor on the Mini-Sentinel staging network. MSOC reviewed sample results to confirm proper system functionality. MSOC and the Mini-Sentinel IT vendor held weekly meetings to plan the transition and track testing of the query tool. The IT vendor deployed two parallel production environments (Mini-Sentinel 2.x and Mini-Sentinel 3.x) to ensure full functionality of the query tool throughout the transition process.

After MSOC approved all upgrades and enhancements, Data Partners were transitioned to the updated secure production server and Portal, and software upgrades were installed. MSOC provided Data Partners with updated role-based user manuals (e.g., *DataMart Administrator Manual*, *Investigator Manual*, *Overview and Administrators Guide*) and detailed setup instructions. Technical questions about the software and security architecture were answered by MSOC staff and the Mini-Sentinel IT vendor responsible for creating and operating the system. Once transitioned to the production server, MSOC issued test queries for each query type to ensure the upgraded system was functional and operating as expected.

MSOC and the software developer 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 the Data Partners of the changes implemented. There are currently 17 unique Data Partners using the Mini-Sentinel Distributed Query Tool Portal.

5. Portal Enhancements

Enhanced User Registration: To allow users to register for an account via a portal page. Workflow is used to track user registration requests, automate the processing of these requests wherever possible, notify all parties involved, and provide an audit trail of the activity. Enhanced User Registration reduces the effort required to administer the Mini-Sentinel Network.

Network Browser: Version 3.0 allows for flexible and rich relationships between networks, projects, organizations, groups, and users. The Network Browser feature allows users to see these relationships and allows administrators to manage these relationships using a control similar to the Windows File Explorer.

Sticky User Settings: PopMedNet users set user interface “settings” in various parts of the query tool to optimize the user interface for the tasks to be performed. Examples include display filters on display grids, number of results returned from results grids, and panels that should remain open and closed on pages which implement complex activities.

Single Sign On: To improve the management of queries and to allow for a single point of entry for multiple Mini-Sentinel applications, a single sign on landing page is being created for the upgraded platform. The single sign on for Mini-Sentinel web-based applications will allow a user to sign in through a secure landing page and gain access to all available applications due to a Lightweight Directory Access Protocol (LDAP) or Active Directory (AD) based single sign on service. MSOC is piloting this feature with the Query Tool, Mini-Sentinel Public Website administration page, and the Mini-Sentinel Data Catalog.

Next, MSOC will transition all Mini-Sentinel Data Partners with access to the Query Tool and the secure file transfer portal to using the single sign on feature for access to both applications.

D. WEB-BASED LIBRARY AND TOOLKIT

1. Overview of Web-Based Library

All modular programs, data checking algorithms, and related tools are posted to the Mini-Sentinel website upon completion. These programs and the related documentation are updated as needed. The Mini-Sentinel tools also include SAS ‘macros’ that can be re-used by programmers. Programmers writing SAS programs running against the MSDD are able to use these macros to speed their development work. The advantage of doing so is twofold. First, it reduces programming effort and cost to design, write, and test these commonly used procedures, thus speeding the development phase, reducing the potential for programming errors, and minimizing quality check time. Second, it implicitly allows Mini-Sentinel investigators to use algorithms (e.g., the Mini-Sentinel stockpiling algorithm) already validated by other investigators or FDA requesters and executed by all Data Partners, thus ensuring consistency across different projects.

As part of Year Four activities, the Data Infrastructure Group kept building on the web-based library of tools previously established in Year Three. In addition to enhanced and new modular programs posted to the [Modular Program section of the website](#), four new tools were added to the [Toolkit Section of the website](#): the Charlson Comorbidity Index tool, the Incidence Risk Ratio tool, the Daily Dosing Analysis module, and the MSOC Log Checker tool program. These new tools are used by some modular programs and other distributed programs.

2. Description of Currently Available Tools

All SAS code posted to the Mini-Sentinel library includes a user guide and documentation. In addition each standalone macro comes with examples and test datasets to be used as test scenarios to speed development work. Table 3 contains a list of all programs and macros posted as Year Four activities. More macros will be posted as they become available, and all new modular programs will be posted once finalized.

Table 3. Description of Year Four Web-Based Library

Program Name	Short Description
Modular Program 1 v2.0	Year Two version of MP1: medication/procedure use
Modular Program 3 v7.0	Incident medication/procedure use and outcomes
Modular Program 4 v6.0	Concomitant medication/procedure use
Modular Program 6 v7.0	Medication/procedure use following a diagnosis
Modular Program 7 v5.0	Most frequently used codes prior & post index event
Modular Program 8 v4.0	Drug utilization, uptake rate, and persistence
Modular Program 9 v3.0	Background rate and characterization of health outcomes of interest among individuals with or without conditions of interest
MS_AgeStrat v1.0	Age & time stratification tool
MS_CreateEpisodes v1.0	Creation of continuous treatment episodes with maximum allowable treatment gap
MS_Denominator v1.0	Reconciliation (i.e., bridging) of enrollment episodes with maximum allowable gap
MS_Envelope v1.0	Reclassification of Encounter Type value to IP for non-IP encounters identified during actual IP stays
MS_GetPharmacy v1.0	Extraction of outpatient pharmacy records with drug codes of interest
MS_GetMedical v1.0	Extraction of medical records with diagnosis and/or procedure codes of interest
MS_FreezeData v1.0	Creation of snapshot/frozen MSDD datasets for cohort of patients of interest
MS_Stockpiling v1.0	When an outpatient pharmacy dispensing is filled in early, make the next dispensing start at the end of the previous
MS_ConfirmElig v1.0	Confirm that medical and pharmacy records of interest or episodes within eligibility periods
MS_CCI v1.0	Standalone module to stratify cohort of interest into groups based on their medical complexity using the Deyo adaptation of the Charlson Comorbidity Index
MS_IRR v1.0	Standalone tool to be used with risk assessment conducted with MP3 to automate the comparison of two cohorts and their incidence rates
MS_DosingAnalysis v1.0	Standalone module to characterize daily dosing patterns of cohorts exposed to outpatient pharmacy dispensing
MS_LogChecker v1.0	Standalone module to be used with any MS distributed program to analyze and report the content of log file(s) generated by program execution

E. ELECTRONIC SUPPORT FOR PUBLIC HEALTH (ESP)

MSOC is using an open source electronic medical record public health surveillance platform known as Electronic Support for Public Health (ESP) as a tool for creation of synthetic clinical data (i.e., vital sign and laboratory results) in the MSCDM format.

1. Enhancing ESP's Driver to Create Fake Clinical Data

Building on the work of Year Three, enhancements were made on the ESP platform. The processes generating synthetic data were updated to reflect the new laboratory and vital data tables of the MSCDM. More specifically, new test types and names were added, and some formats were revised (e.g.,

names, codes, lows, highs, and units). As a result, simulated data to develop programs using the current structure of the two clinical data tables of the MSCDM are available to the MSDD community.

2. Installation and Documentation

This customized version of ESP is currently installed on one workstation owned by MSOC. It is only available for use by the MSOC staff. Documentation and training material is available to allow MSOC staff to be able to independently create synthetic datasets for Mini-Sentinel use.

F. LESSONS LEARNED

During Year Four the Infrastructure Data Group successfully implemented a series of updates and enhancements to the pool of analytic tools. These tools are actively used to respond to Mini-Sentinel queries (see Section VIII), and through this use have identified opportunities to improve their use and efficiency. The procedures developed by MSOC to implement changes to the analytic tools proved valuable in helping to ensure reliable transitions to new and updated tools. A summary of lessons learned related to the key Mini-Sentinel analytic tools is below.

1. Modular Programs

During the first three years of the MS pilot MP development was focused on building new programs and fixing minor issues. After two years of intense use by FDA and the workgroups, important feedback on new features and bug fixes from FDA, Data Partners and MSOC staff was collected. Based on this feedback, the focus of Year Four was incorporation of major enhancements to the existing pool of MPs, adding multiple new features and modules, and fixing several issues.

The Data Infrastructure Group installed and began using “bug tracker” software (MantisBT) to document, track, and assign suggested enhancements, bugs, and feature request. This issue tracking system is proving invaluable for management of MP enhancements and other programming tasks, and has become the main communication tool between the Data Infrastructure Group and programmers and software developers.

To better accommodate the volume and scope of requested MP revisions, the programming tasks were divided into three short development cycles, coupled with the new in-house version control system. This approach of rapid-cycle development and improved management systems has proven effective in maintaining timelines and has resulted in several new MP “releases”. In Year Five, MSOC will spend more time on the enhancement design phase, creating clearer specification documents and more comprehensive Quality Compliance plans to avoid unnecessary back and forth communications between FDA, MSOC, and programmers. Additionally, we will explore new ways of providing modular program education.

2. Summary Tables and Distributed Query Tool Software

The Mini-Sentinel Distributed Query Tool is the most actively used tool within Mini-Sentinel and has proved very useful in quickly generating high-level information regarding exposures, diagnoses, procedures, and enrollment. To date, the Query Tool has been used to issue over 250 summary table queries that generated information on over 1,000 drug exposures, diagnoses, and procedures.

The increasing importance of the tool has highlighted our need to tightly manage software upgrades to ensure that the tool is available for use. The Query Tool is a complex software application that now involves dedicated software management and support by MSOC. The Version 3 platform enhancements increase the scalability and ease of maintenance of the tool. The plug-in architecture allows more information to flow through the Query Tool and improves the workflow of the Data Partners and MSOC.

During Year Four, MSOC worked closely with Data Partners to streamline the summary table creation process. This process involves: 1) sending each Data Partner a package to generate summary tables as soon as each data refresh is approved; 2) checking the logs returned from Data Partners for any issues from the program runs; 3) sending instructions on getting the newly-generated summary tables connected to the Query Tool; 4) sending test queries; and 5) reviewing test query data to make sure there are no issues. The process has been streamlined in such a way that allows each Data Partner's most recent data to be available for querying as quickly as possible.

VI. MINI-SENTINEL INFRASTRUCTURE

A. MINI-SENTINEL DATA CATALOG

The Mini-Sentinel Data Catalog (MSDC) is a software tool that tracks all data requests within the Mini-Sentinel distributed data network. The MSDC was updated substantially during Year Four to improve functionality, reporting, and tracking capabilities. New search and reporting features also were developed during the year, as well as the architecture to import and track all Mini-Sentinel Summary Table requests.

1. Function of the Mini-Sentinel Data Catalog

The MSDC tracks information about all queries distributed within the network, including the query description, query type (e.g., modular program, beta testing, workgroup), project and query unique identifiers, relevant budget item, query distribution and response dates, and the participating data partners.

The MSDC incorporates Data Partner query response information by parsing file names of data received and automatically emails MSOC team members whenever a Data partner uploads a query response to the Mini-Sentinel secure portal. The auto-generated email includes information such as Data Partner name, file name, file location, and upload date.

The MSDC report function produces essential metrics and allows users to select from a variety of filters to customize reports for different audiences. It is hosted within the Mini-Sentinel secure private cloud environment in a Federal Information Security Management Act of 2002 (FISMA)^{xiii} compliant TIER III data center.

^{xiii} <http://csrc.nist.gov/groups/SMA/fisma/index.html>

2. Expansion of the Mini-Sentinel Data Catalog during Year Four

In Year Four, the MSDC was used to track more than 150 data requests and has proven to be an important tool for tracking and managing data requests and evaluating overall network performance. Several technical improvements were made to the MSDC. Working with our IT partners, we designed and implemented an architecture that allows communication between the Mini-Sentinel Distributed Query Tool and the MSDC. This enables the MSDC to import Summary Table query metadata and use those data for tracking and reporting functions. .

Text search functionality was added to allow users to search for workplans (i.e., data requests) containing any text string. This feature can be used to search for prior queries on a specific topic. Audit functionality was expanded to include information on workplan and project creation, update, deletion and locking. A “send” function was added that allows users to send workplans directly to Data Partners from within the MSDC.

Year Four work on the MSDC also involved ensuring data completeness and accuracy by creating data integrity controls on single data points and writing queries to check data that is not controlled at the time of entry. MSDC user documentation was written.

3. Future Work

The MSDC has proven to be a valuable tool in tracking MS projects. However, further enhancements can be made to improve usability, tracking details, and reporting functionality:

- Integrate additional request types into a single tracking system.
- Enhance text search capability
- Increase data integrity by building edit checks
- Improve request entry functions and usability
- Improved search and reporting functionality
- Enhance access controls

B. IMPLEMENTATION OF MANTIS ISSUE TRACKING SYSTEM

During Year Four, MSOC implemented MantisBT, a free popular web-based issue tracking system, released under the terms of the GNU General Public License (GPL). It helps MSOC track and organize the multiple programming and technical projects being developed and implemented by local MSOC staff and collaborators and vendors. Since the tracking system is web-based, virtually any web browser can access the secure issue tracking system. [MantisBT is hosted](#) at the same private cloud environment hosting the secure and distributed query tool portals ([please see Section V.C.1 Overview of the Mini-Sentinel Distributed Query Tool](#)).

MantisBT is a collaborative environment, and can be used as a communication tool with which the MSOC admin staff and various programmers and collaborators can share information, comment on and resolve issues, and exchange documents. Using the issue tracking system as both an organizational and communication tools avoids painful and hard to track electronic mail exchanges and manual progress tracking.

By the end of Year Four, most major MS projects with a development component (e.g., core modular and ad hoc program development processes, workgroups) or sequential nature (e.g., data quality assurance review and characterization) were using the issue tracking system to track issues and progress on programming, quality assurance, and beta testing.

C. AUTOMATED REPORTING TOOL

1. Function of the Automated Reporting Tool

The creation of modular program and summary table reports requires a series of manual processes. To improve efficiency and reduce manual processes, the Data Infrastructure and Production Groups co-led development of the Mini-Sentinel Automated Reporting Tool. The tool is a set of SAS programs and Excel VBA code that create reports from modular program output that greatly reduces the time needed to produce reports and minimizes manual processing. In Year Four, the first automated report tool was created for use with output generated by Modular Program 3. The tool is able to import the output of a simple MP3 run and generate report tables that are then reviewed by the Production Group before submitting to FDA.

2. Future Work

- Increase flexibility to incorporate output from additional Modular programs and more complex Modular Program 3 output
- Automatically generate a specification sheet and lists of codes used

D. MINI-SENTINEL SECURE PORTAL

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. Any approved members of the Mini-Sentinel community can transfer documents in a section specifically assigned to them (or their group/organization).

During Year Four, MSOC implemented several upgrades to the secure portal system to enhance the security features and administration. The maintenance and administration of the system was made easier and can now be performed directly 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).

E. TESTING ENVIRONMENT AND SYNTHETIC DATA

In Year Four, the Data Infrastructure Group implemented the Mini-Sentinel testing environment in which modular programs and workgroup/evaluation programs are developed, tested, checked for quality compliance, and validated. It consists of: 1) a pool of high-performance workstations installed with programming and editing applications (e.g., SAS, program editor, graphic analytics, data processing and formatting); and 2) a synthetic version of the MSDD with data for 5 million fictitious members

spanning six calendar years. The workstations are only available to internal MSOC staff, whereas random samples of synthetic data can be shared with MS programmers and collaborators upon request.

During Year Four, MSOC enhanced its internal testing environment with additional computing resources, and the enhancement of the available synthetic database.

Computing Resources:

Eight new high-performance workstations were added to the testing environment, for a total of thirteen. All workstations can be shared between all MSOC programmers and data analysts, and are accessible via remote desktop applications.

In addition to performing test runs and quality compliance checks described earlier, three of those workstations were set up with additional capabilities of virtual environments allowing the MSOC to replicate the Data Partner environment in terms of different operating systems (e.g., linux, UNIX), SAS versions, and volume of data (both in number of members and years spanned). Doing so allowed the MSOC to improve testing of its various modular programs by reducing the need to require multiple Data Partners to go through several rounds of beta-testing.

Synthetic Data:

Up until Year Three, the clinical data elements were not included in the MSOC synthetic data. In Year Four, a special program was designed to read simulated laboratory and vital data generated by the ESP software and output these data in the MSCDM format. These data were then merged with the core MSCDM synthetic tables and are now used as test data sets. This program is easily customizable and can be revised to accommodate future revisions and additions to the MSCDM clinical data.

F. LESSONS LEARNED

Mini-Sentinel Data Catalog (MSDC):

The MSDC updates enable MSOC to accurately track and manage Mini-Sentinel data requests. Data request tracking still requires several steps that could be streamlined by with additional features and integration with the Mini-Sentinel Distributed Query Tool.

Automated Reporting Tool:

Creating a modular program or summary table report manually takes one to two days of analyst time and one to several hours of programmer time. The Automated Reporting Tool created this year has cut the report generation time for MP3 to a few hours of analyst time with no programmer involvement. Automating as much of the reporting process as possible will make it substantially more efficient; work will continue to expand use and functionality of the Automated Report Tool.

Secure Portal:

During the upgrade process, multiple users experienced issues using the secure portal with some types of documents (e.g., spreadsheets and zip archives). Since these types of documents are crucial, the upgraded system had to be reverted back to its original system in order to fix the issues. Going forward

any upgrades to the system will need to undergo more thorough testing. To ensure that all users accounts are appropriate and that users are removed as necessary (e.g., due to changes in employment), MSOC will be generating periodic reports with lists of portal accounts for each organization (e.g., FDA, Data Partners, MSOC) and reviewing logs of access for audit trail purposes. Each organization will review the list of users to verify the continued need for access to the secure portal.

Testing Environment and Synthetic Data:

With the volume of data requests increasing, automating pre-distribution program package testing and validation will become increasingly important. Automation will obviate several manual steps that must be undertaken by MSOC staff to relieve some of the beta testing burden placed on Data Partners.

For testing purposes MSOC must continue to work on building internal data resources that reflect the introduction of new medical products. Synthetic data and lookup tables that reflect new products must be built to enable efficient testing of distributed code.

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. Midway through Year Four, the format of our regularly scheduled meeting changed from a weekly teleconference to monthly web conference. The expanded format improves communication and enables more substantive presentations. Examples of the Year Four presentation topics include a discussion of chart validation approaches for a workgroup, walk-through of a new workplan, review of the MSOC organization chart and key contacts, guidance on data retention policies and procedures, and a review of modular program features and planned enhancements. The new format has been well-received. In addition to these regularly scheduled meetings, the MSOC regularly communicated 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 various 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 also available by email and phone if needed. In Year Four, the MSOC expanded its role in the workgroups by advising on the use of Modular Programs and Summary Table queries in feasibility studies. In Year Four, the MSOC completed ten Modular Program requests and two Summary Table Requests in support of workgroup activities.

Additionally, 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 all Mini-Sentinel Collaborators. This system has been designed to be compatible with all 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 Four.

Table 4. Meetings and Presentations

Date	Presenter(s)	Venue	Presentation Title
11/5/12	Jennifer Popovic, Nicolas Beaulieu	Active Surveillance Framework Workgroup (webinar for external programmers)	Mini-Sentinel Programming: Guidelines and Resources
11/5/12	Jeff Brown	American Medical Informatics Association annual symposium, Chicago, IL	Late Breaking Session - Realizing a National Learning Health System
11/15/12	Jeff Brown	FDA Webinar	Overview of Mini-Sentinel Analytic Tools
12/18/12	Jennifer Popovic, Nicolas Beaulieu	Active Surveillance Framework Workgroup (webinar for external programmers)	Mini-Sentinel Programming: Overview of ToolKit Macro Programs and Modular Program 3
1/31/13	Jeff Brown	Brookings Sentinel Initiative Public Workshop, Washington, DC	Opportunities to Expand the Public Health Impact of the Sentinel Initiative: FDA Mini-Sentinel as a National Resource
1/31/13	Sebastian Schneeweiss, Jennifer Nelson	Brookings Sentinel Initiative Public Workshop, Washington, DC	Modular Programs
2/1/13	Lesley Curtis	Mini-Sentinel Investigators'	Mini-Sentinel Data Resources

Date	Presenter(s)	Venue	Presentation Title
		Meeting:, FDA	
2/1/13	Marsha Raebel	Mini-Sentinel Investigators' Meeting, FDA	Data Resources: Mini-Sentinel Clinical Data Content and Capabilities (laboratory results and vital signs)
2/1/13	Susan E. Andrade	Mini-Sentinel Investigators' Meeting, FDA	Birth Certificate Data Matching for the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) Program
2/1/2013	Jeff Brown	Mini-Sentinel Investigators' Meeting, FDA	Overview of Mini-Sentinel Analytic Tools
2/1/2013	Joshua J Gagne, Sebastian Schneeweiss	Mini-Sentinel Investigators' Meeting, FDA	Hd-PS capability for MPs
3/12/13	Jeff Brown	Public Health and the Learning Health System: A National Meeting, The Network for Public Health Law, Ann Arbor, MI	Lessons from Two Distributed Networks for Public Health
3/13/13	Robert Rosofsky	Boston Area SAS Users Group, Boston, MA	The Mini-Sentinel Distributed Database Project
3/20/13	Jeff Brown	American Medical Informatics Association annual Summit on Clinical Research Informatics, San Francisco, CA	Pains and Palliation in Distributed Research Networks: Lessons from the Field
4/8/13	Jeff Brown	Medical Informatics World Conference, Boston, MA	Provider-Payer-Pharma Cross-Industry Data Collaboration: Overview of the Mini-Sentinel program
4/16/13	Nicolas Beaulieu, April Duddy	HMORN Conference, San Francisco, CA	Mini-Sentinel Modular Programs: Overview of Add-On Tools and Enhanced Features
5/19/13	Kevin Haynes	ISPOR Annual meeting, New Orleans, LA	Distributed research networks and applications in safety and out
6/4/13	Jeff Brown, Tiffany Woodworth	FDA Data Core Site Visit, Silver Springs, MD	Querying the Mini-Sentinel Distributed Database
6/12/13 6/13/13	Jeff Brown	Drug Safety Research Unit, 7th Biennial Conference On Signal Detection and Interpretation In	Data Mining Signal Detection in Longitudinal Databases DEBATE: What's the difference between signal generation, signal refinement and signal evaluation?

Date	Presenter(s)	Venue	Presentation Title
		Pharmacovigilance London, England	
8/25/13	Kevin Haynes	ISPE workshop	TBD
TBD	Kevin Haynes	Pre-conference Symposium	TBD

VIII. MSDD QUERY REQUEST SUMMARY

A. MODULAR PROGRAMS

A total of 58 Modular Program requests were initiated in Year Four. Of these, 52 were completed as of September 22, 2013. Of these 52 completed requests: CDER was responsible for 25 requests; CBER 14 requests; CDRH one request; workgroups 10 requests; and MSOC initiated two requests (Table 5). MP1 was used in four requests, MP2 in four requests, MP3 in 28 requests, MP5 in three requests, MP6 in four requests, MP7 in four requests, and MP9 in four requests (Table 6). The 52 completed requests involved between one and 96 modular program scenarios each for a total of 1135. A scenario is defined as a unique set of input 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 52 completed requests generated 75 reports.

The requests had varying levels of complexity, ranging from a straightforward MP1 request with one run, analyzing prevalent and incident drug use, to a complex request consisting of a MP3 with pre-existing conditions and exposure/event incidence input files. For example, one dabigatran request consisted of 32 scenarios to assess AMI events among warfarin and/or dabigatran users overall as well as users with a pre-existing condition of atrial fibrillation. This request used differing incident drug criteria, washout periods, and primary diagnosis criteria. In another example, an IVIG request using MP7 required two reports to present results for the occurrence of over 50 specified procedures and diagnoses before and after incident and prevalent IVIG use.

Table 5. Number of Modular Program Requests, Scenarios, and Reports by Requester in Year Four (September 23, 2012 to September 22, 2013)

Center/ Requester	Number of Requests Initiated	Number of Requests Completed	Number of Scenarios Completed	Number of Reports Completed
CDER	28	25	536	36
CBER	14	14	229	18
CDRH	2	1	37	3
Workgroups	12	10	246	15
MSOC	2	2	87	3
Total	58	52	1135	75

Table 6. Number of Completed Modular Program Requests and Scenarios by Modular Program in Year Four (September 23, 2012 to September 22, 2013)

Modular Program (MP)	Number of Requests	Number of Scenarios
MP1	4	17
MP2	4	80
MP3	28	723
MP4	0	0
MP5	3	62
MP6	4	51
MP7	4	44
MP9	4	85
Other*	4	73
TOTAL	55	1135

* These include ad-hoc requests that used newly-developed non Modular Program code but were distributed and executed as part of the Modular Program set of activities.

Note: The number of Requests in Table 6 does not match Table 5 because one request used multiple modular programs.

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 52 requests, 37 were completed on time by all Data Partners. Of the 15 remaining requests, the average number of days to completion past the due date was 4.9 and the median was 4 (including weekends and holidays). Overall, response time by Data Partners was well 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 the report contents, 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 10.6 days and the median time was 7 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 82 summary table queries were performed to respond to 28 requests during Year Four (Table 7). Multiple queries are sent per request 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. 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. The 82 queries performed included 204 drug products, 54 diagnosis-setting combinations, 45 procedure-setting combinations, and 187 HCPCS-setting combinations, each stratified by age group, sex, and year. CDER was responsible for 18 requests, while CDRH submitted two and CBER and the FDA leadership team submitted one request each. Two workgroups (the Intravenous Iron Workgroup and the 15 Cohorts Workgroup) also submitted one request each. Finally, MSOC initiated three requests to investigate counts as background information for modular program requests.

Data Partners were typically given two business days to complete each query, and all responded within the allotted time. Occasionally, MSOC would wait for a Data Partner to update its data before distributing a particular request, especially if the request was for more recent data.

For the 22 requests that generated summary table reports, 54 summary table reports completed (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 in the report.

Table 7. Number of Summary Table Query Requests in Year Four (September 23, 2012, to September 22, 2013), 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 Drugs Completed	Number of Completed Requests Involving Reports	Number of Reports Completed
CDER	18	18	38	199	16	32
CBER	2	2	7	12	2	2
CDRH	2	2	18	102	2	17
FDA Leadership	1	1	4	4	1	2
IV Iron WG	1	1	3	45	0	0
15 Cohorts WG	1	1	4	60	1	1
MSOC	3	3	8	68	0	0
TOTAL	28	28	82	490	22	54

Table 8 displays the number of queries completed during Year Four stratified by requester and query type. Generic name queries (33) and HCPCS queries (24) accounted for the bulk of activity.

Table 8. Number of Summary Table Queries Completed in Year Four (September 23, 2012, to September 22, 2013), 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	---	29	---	---	4	1	---	---	4	38
CBER	---	1	---	---	3	3	---	---	---	7
CDRH	---	---	---	1	1	1	1	5	9	18
FDA Leadership	---	1	---	---	---	---	---	---	3	4
IV Iron WG	---	---	---	---	---	---	---	---	3	3
15 Cohorts WG	---	---	---	---	---	---	---	---	4	4
MSOC	---	2	---	1	1	1	---	2	1	8
TOTAL	0	33	0	2	9	6	1	7	24	82

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 then needed to fulfill the requirements of such requests. De novo programming much adhere to the Mini-Sentinel 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 MSOC for use.

The flexible PDX-care setting interaction feature, daily dosing module, and MP4 event identification module described in [Section V.A.3](#) and [Section V.A.4](#) are the three de novo programming activities that led to Year Four ad hoc data requests. The new programming code was incorporated into existing modular programs (as opposed to creating new ones). Once approved for use, ad hoc requests followed the same Data Partner and report timing as existing Modular Programs.

D. POSTINGS TO MINI-SENTINEL WEBSITE

During Year Four, MSOC continued posting reports generated from summary table and modular program requests to the Mini-Sentinel website. These reports, completed during Years Two, Three, and Four, were approved for posting by FDA. No Data Partner-specific results are included in posted reports. In Year Four, a total of 38 reports were posted. All 38 reports appear in the “Assessments” tab on the website: 28 under the sub-tab “Exposures to Medical Products”, 4 under the sub-tab “Diagnoses and Medical Procedures”, and 6 under the sub-tab “Health Outcomes Among Individuals Exposed to Medical Products”. The titles of the reports are shown below.

1. Reports

a. Summary Table Reports Under “Assessments: Exposures to Medical Products”:

- Amphotericin use
- Analgesic use 2
- Anti-infective agents use
- Boceprevir and telaprevir use
- Cardiovascular therapy agents use
- Gastrointestinal therapy agents use
- Golimumab, ustekinumab, and dronedarone hydrochloride use (by quarter)
- Golimumab, ustekinumab, and dronedarone hydrochloride use (by year)
- Isoniazid use
- Natalizumab and efalizumab use
- Nucleoside reverse transcriptase inhibitor procedures
- Occurrence of selected generic drugs 3
- Occurrence of selected generic drugs 4
- Occurrence of selected generic drugs 5
- Occurrence of selected generic drugs 6
- Occurrence of selected generic drugs 7
- Occurrence of selected generic drugs 8

b. Modular Program Reports Under “Assessments: Exposures to Medical Products”:

- Bupropion and naltrexone use 1
- Bupropion and naltrexone use 2
- Clopidogrel and prasugrel use
- Dabigatran, rivaroxaban, and warfarin use
- Duloxetine, pregabalin, and milnacipran use 1
- Duloxetine, pregabalin, and milnacipran use 2
- Lindane use

- Occurrence of selected pediatric drugs 1
- Parkinson's disease medication use
- Selgiline use
- Vascular endothelial growth factor (VEGF) inhibitor use (by generic drug name)

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

- Hip implant procedures and diagnoses 2
- Injection amphotericin procedures
- Injection natalizumab procedures
- Injection ustekinumab and injection denosumab procedures

d. Modular Program Reports Under "Assessments: Health Outcomes Among Individuals Exposed to Medical Products":

- Angiotensin Receptor Blockers (ARBs), hydrochlorothiazide, atenolol, amlodipine use & celiac disease
- Dabigatran (Pradaxa), warfarin & GI bleed, intracerebral hemorrhage
- Dabigatran, warfarin & GI bleed, intracerebral hemorrhage
- Natalizumab, efalizumab, & progressive multifocal leukoencephalopathy (PML)
- Oxicam NSAIDs, modafinil/armodafinil, sulfamothoxazole & severe cutaneous adverse reaction (SCAR) events
- Warfarin & GI bleed, intracerebral hemorrhage

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.

2. Other Postings

a. Mini-Sentinel Data Core Modular Programs

During Year Four, the MSOC posted to the Mini-Sentinel website revised versions of documentation and code for Modular Programs 1-7, as well as documentation and code for the new Modular Program 9 (Table 9). Modular Program 9 combines the features of Modular Programs 1, 2, and 5 into a single tool, so those programs were subsequently archived. Mini-Sentinel's modular programs facilitate rapid querying of the Mini-Sentinel Distributed Database. Each program focuses on a specific type of question and executes against the Mini-Sentinel Common Data Model. Each MP has a specific set of required input parameters; the standardized output contains summary-level counts by Data Partners and overall (e.g., number of members exposed to a medical product, number of members with a specific diagnosis/condition) stratified by various parameters (e.g., age group, sex, year). 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 Four

Date	Document Title
11/29/12	Modular Program 1: Characterization of Use of Medical Product Exposures (version 2.0)
3/12/13	Modular Program 2: Characterization of Use of Medical Product Exposures among Individuals with or without Condition(s) of Interest (version 3.1)
3/12/13	Modular Program 5: Background Rates for Health Outcomes of Interest (version 2.1)
5/22/13	Modular Program 3: Frequency of Select Events During Exposure to a Drug/Procedure Group of Interest (version 5.0)
5/23/13	Modular Program 4: Concomitant Drug and/or Procedure Use (version 4.0)
5/23/13	Modular Program 6: Frequency and Duration of Treatment Following an Event of Interest (version 5.0)
5/29/13	Mini-Sentinel Modular Program 7: Drug Use, Medical Diagnoses, and Medical Procedures Before and After an Exposure or Event of Interest (version 3.0)

b. Mini-Sentinel Toolkit Library

During Year Four, 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 (Table 10). Each tool is a self-contained SAS® macro. These tools are used in combination to facilitate development of the Mini-Sentinel Modular Programs. MSOC will continue to post new and revised programming tools as they are developed.

Table 10. Mini-Sentinel Toolkit Library Documentation Posted In Year Four

Date	Document Title
12/28/12	SAS Macro Toolkit: Age Stratification (version 1.0)
12/28/12	SAS Macro Toolkit: Treatment Episode Reconciliation (version 1.0)
12/28/12	SAS Macro Toolkit: Envelope Algorithm Execution Requirement (version 1.0)
12/28/12	SAS Macro Toolkit: Envelope Algorithm: Inpatient Claim Reclassification (version 1.0)
12/28/12	SAS Macro Toolkit: Eligibility Episode Reconciliation (version 1.0)
12/28/12	SAS Macro Toolkit: Data Subset Creation (version 1.0)
12/28/12	SAS Macro Toolkit: Pharmacy Claims Extraction (version 1.0)
12/28/12	SAS Macro Toolkit: Medical Claims Extraction (version 1.0)
12/28/12	SAS Macro Toolkit: Transfer Unique Values (version 1.0)
12/28/12	SAS Macro Toolkit: Claims/Episodes Within Enrollment/Eligibility (version 1.0)
12/28/12	SAS Macro Toolkit: Word Counter (version 1.0)
4/8/13	SAS Macro Toolkit: Stockpiling (version 1.0)
4/23/13	SAS Macro Toolkit: All Macros (version 1.0)

E. LESSONS LEARNED

1. Modular Programs

The modular program request fulfillment lifecycle has become much more routine and predictable during the year. The Production Group has completed FDA and workgroup modular program requests representing information for hundreds of drug products, diagnoses, and procedures. Based on feedback from FDA and others in Year Three, MSOC implemented two changes in Year Four: an updated review process throughout the modular program life cycle and a more standardized reporting structure.

After a year of using the new review process, by which input files, test runs and reports are reviewed multiple times by various Production Group members, we have noted a decrease in modular program errors in both execution and logic and thus an improvement in query response time. Additionally, as we have used a more standardized approach to reporting, reports are more quickly generated and less error-prone using the [Automated Reporting Tool](#).

The increasing use of modular programs by various new requesters from FDA and MS workgroups has highlighted the need to improve communication with requesters. In many cases, new requesters may not be familiar with program capabilities or administrative claims data. The Scientific Operations Center has implemented a new standard in which a teleconference is set up with all new requesters to review program specifications and definitions and to answer general questions. Implementing this step will help ensure that questions are answered efficiently and appropriately and facilitate use of Mini-Sentinel by new requesters.

2. Summary Tables and Query Tool

With each request, MSOC continues to improve the reports summarizing results both in terms of the information contained in the tables and figures that are displayed and in terms of formatting. Moreover, there are certain formatting conventions and notes that have been added to reports that have been posted to the Mini-Sentinel website. Every time a new report is created, these formats and notes are now applied so that preparing them for website posting will subsequently be more efficient.

The FDA will often submit a request for a new medical product. Therefore, it is important that the lookup tables be kept as up to date as possible. However, as time passes by, some codes will be discontinued but will still appear during earlier years of the MSDD. MSOC has thus learned of the need to keep older, available codes in the lookup tables whenever possible.

IX. CONCLUSION

This report described the Mini-Sentinel Data Core activities undertaken during Year Four of the Mini-Sentinel project. As evidenced in the report details, MSOC had a productive year with continued expansion in the efficiency of operations, improved quality of data and reports, improved technical and programming capabilities. For Year 5 we look forward to continuing the progress of this unprecedented and significant public health initiative.

X. REFERENCES

1. Brown JS, Lane K, Moore K, et al. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative; U.S. Food and Drug Administration: FDA-2009-N-0192-0005. 2009. Available at: <http://www.regulations.gov/#!documentDetail;D=FDA-2009-N-0192-0005>. Accessed 3/16/11.
2. Maro JC, Platt R, Holmes JH, et al. Design of a National Distributed Health Data Network. *Ann Intern Med.* 2009; 151: 341-344.
3. Velentgas P, Bohn R, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiology and Drug Safety.* 2008; 17: 1226-1234.
4. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care.* 2010; 48: S45-51.
5. Brown J, Holmes J, Maro J, et al. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. Effective Health Care Research Report No. 13. (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHS290200500331 T05.) Rockville, MD: Agency for Healthcare Research and Quality, July 2009. Available at: <http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=150>. Accessed 3/16/11.
6. Brown J, Holmes J, Syat B, et al. Proof-of-principle evaluation of a distributed research network. Effective Health Care Research Report No. 26. (Prepared by the DEcIDE Centers at the HMO Research Network and the University of Pennsylvania Under Contract No. HHS290200500331 T05.) Rockville, MD: Agency for Healthcare Research and Quality, June 2010. Available at: <http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=464>. Accessed 3/16/11.
7. Brown J, Syat B, Lane K, et al. Blueprint for a distributed research network to conduct population studies and safety surveillance. Effective Health Care Research Report No. 27. (Prepared by the DEcIDE Centers at the HMO Research Network and the University of Pennsylvania Under Contract No. HHS290200500331 T05.) Rockville, MD: Agency for Healthcare Research and Quality, June 2010. Available at: <http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=465>. Accessed 3/16/11.
8. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr.* 2005:12-25.
9. Electronic Primary Care Research Network (ePCRN). Available at: <http://www.epcrn.bham.ac.uk>. Accessed 3/16/11.