MINI-SENTINEL PROSPECTIVE ROUTINE OBSERVATIONAL MONITORING PROGRAM TOOL:

CREATION OF THE ANALYTIC DATASET AND CONDUCTING ANALYSIS FOR THE GROUP SEQUENTIAL GENERALIZED ESTIMATING EQUATION (GS GEE) TOOL

Technical Users' Guide version: 1.0

Prepared by: Andrea Cook, PhD,[1] Robert Wellman, MS,[1]
Jennifer Nelson, PhD[1]

Author Affiliations: 1. Group Health Research Institute, Seattle, WA

July 23, 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 HHSF22320091000.

Table of Contents

I. OVERVIEW

The main purpose of this tool is to identify and extract a cohort of interest and use regression analysis to estimate the association between a medical product exposure and health outcome of interest that is adjusted for pre-specified confounders. This analysis can also be conducted repeatedly over time using group sequential methods. Specifically, a group sequential regression using generalized estimating equations (GS GEE) is implemented. The analytic dataset creation program is written in SAS and can be customized using various parameter settings that define exposures, outcomes, date ranges, age ranges, and other implementation details. Sections III through VIII describe the key program specifications and main assumptions underlying each of the query parameters for GS GEE Data Creation Program version 1.0. Program specification requirements, formats, and default values of all parameters are defined. The second part of the documentation details the analysis parameters and discusses an example report from applying the GS GEE method.

II. GS GEE METHOD OVERVIEW

Group Sequential Generalized Estimating Equation (GS GEE) method is an approach that can flexibly and robustly accommodate a variety of different exposure and outcome types. Confounding is taken into account through regression adjustment for characteristics measured at baseline and thought to be associated with both the exposure of interest (e.g., medical product of interest) and the health outcome of interest (HOI). Specifically, like the propensity score matching tool, this method is suitable for use in a cohort design where users of an exposure of interest are compared to those who receive an alternative exposure, e.g. an active-comparator new user cohort design, or are unexposed during the same period of time. Instead of exposure matching on a propensity score, however, regression estimation is used to control for baseline confounders. While the GS GEE methodology can accommodate continuous confounders, data sharing and privacy issues with data partners may require the use of aggregated data, and thus the use of categorical confounding variables. This could involve the inclusion of several categorized individual confounders like age, gender, and comorbidity status or it could adjust for a categorized summary score like a propensity score. Since the method uses a robust GEE framework, it can accommodate a variety of safety outcomes, including outcomes that are binary (i.e., yes or no), count (i.e., number of events), or continuous in nature (e.g., level of a laboratory value). Below we describe two common types of exposure-HOI scenarios that are likely to arise in Mini-Sentinel and discuss how the GS GEE method can be utilized in each instance.

For short term exposures (e.g., a one-time injection or an antibiotic that is used for short period of time), the method involves defining a binary indicator of being exposed or unexposed and a binary occurrence of an HOI within a pre-specified risk window following product initiation. Subjects are not included in a given sequential analysis until the completion of their HOI risk window (i.e., until that subject's risk window has been fully observed) so that all patients have the same follow-up time. A binomial error structure is assumed and combined with a logit link function to estimate an adjusted odds ratio (OR) as the measure of risk in the exposed group relative to the comparator group.

For longer term exposures (e.g., a chronically used drug taken over several months or years), the method defines binary exposure status based on the first record of receipt of either the exposure of interest or the comparator. The observation time, or time at-risk, is calculated as the time from exposure initiation to the first of either exposure discontinuation (with a lag if desired), occurrence of the outcome of interest, disenrollment, or death. This definition yields a binary outcome indicating whether an HOI occurred during the at-risk period, i.e., the period in which the person was observed to be exposed. For an outcome to be included in a given sequential analysis, it must have occurred before the date of that analysis. Poisson regression is used to estimate an adjusted relative rate (RR) that takes into account duration of time at-risk through an offset term included in the model.

GS GEE can be implemented as a one-time analysis or in a formal sequential monitoring framework with multiple testing and early stopping that is accounted for statistically. If implemented sequentially, the user can specify the total number and frequency of sequential tests as well as the desired level of the signaling threshold over time and the desired overall false positive error rate according to scientific, practical, and statistical preferences. The hypothesis being tested depends upon the type of exposure (i.e., medical product) and HOI and the corresponding regression model used. For example, for a binary outcome and logistic regression model, the null hypothesis (H_0) based on the estimated OR comparing the independent exposure and comparator groups is H_0: OR=1. Regardless of the outcome type, exposure type, or sequential design specifications, at each analysis the null hypothesis is evaluated via the computation of a standardized score test statistic. This statistic is then compared to a pre-specified threshold to determine whether there is a signal for elevated risk, or whether monitoring should continue.

To incorporate group sequential monitoring, GS GEE uses a non-parametric permutation approach that is particularly suited for rare outcomes (i.e., does not rely on large sample theoretical assumptions). Specifically, it flexibly simulates data under the null hypothesis of no difference between exposure groups (e.g., H_0: OR=1 for logistic regression, H_0: RR=1 for Poisson regression). It uses the unifying boundary approach and defines the boundary based on the permuted data, incorporating the probability of stopping at earlier analysis times and type I error inflation due to repeated testing. The user can select a pre-specified number of analysis times, timing of analyses (based on observed, or expected sample size, at each analysis time), and a total expected maximum sample size by the end of the assessment based on scientific, practical, and statistical preferences. Boundary shape is also user-specified and so can flexibly handle a number of scenarios. For example, a flatter boundary (e.g., a boundary with a constant signaling threshold over time) will, on average, signal earlier for lower elevated risk than a boundary that is more conservative; that is, it requires a stronger effect to signal at earlier analysis times. However, given the same sample size, a flatter boundary will have less power to signal later on during the surveillance period compared with boundaries that employ early conservatism by having a higher signaling threshold at earlier tests. The boundary values for GS GEE are based on the standardized test statistic (as opposed to an error-spending or alpha-spending scale). Therefore, signal decision rules can be planned directly on the standardized scale of the risk quantity of interest rather than the alpha scale, and thus readily facilitate straightforward sequential design decision-making. In all, the following sequential analysis parameters must be specified: shape of the boundary (Pocock, O'Brien

Flemming), planned testing frequency (e.g., 12 looks with the first look after 10,000 observations and then evenly spaced looks after that point or 12 looks with the first look after 1 year and quarterly looks after that), and total maximum sample size at end of surveillance. Once specified, the signaling boundary at each analysis time point can be computed based both on these input parameters as well as the permuted score test statistic under the null.

III. TERMINOLOGY

In the interest of simplicity, the term “scenario” is used throughout this document to refer to a set of parameters and criteria used to define an execution of the dataset creation query The “requester” refers to an individual (or group of individuals) who initiates the query request and defines the scenarios. The term “analyst refers to an individual who creates request Input Files and distributes the query to the Data Partners.

The terms "exposure", “exposure of interest”, and “comparison exposure” are used to represent exposure to a medical product or procedure as defined by the query requester. The exposure of interest will generally be a product/procedure newly on the market (the product/procedure under surveillance) and the comparison exposure a product with a known safety profile (that serves to establish baseline risk). An exposure can be defined using any set of pharmacy and/or procedure codes found in the Mini-Sentinel Common Data Model (MSCDM). For example, exposure to a drug product dispensed in the outpatient setting can be defined as observation of one or more National Drug Codes (NDCs) in the pharmacy dispensing file, whereas exposure to a vaccine can be defined based on observation of specific procedure codes in the procedure file.

The terms “outcome”, "event", and “event of interest” are used to represent the occurrence of a diagnosis as defined by the query analyst An event can therefore be defined using any set of diagnosis and/or procedure codes found in the MSCDM.

The term “claim” is used to represent an outpatient pharmacy dispensing or medical encounter/record with any of the codes for the exposure(s), event(s) or condition(s) of interest.

The term “member” is used to represent an individual with relevant criteria for enrollment, exposure(s), event(s) and condition(s) (as specified by the query parameters). A member can be further defined as a “user” if evidence of use of exposure(s) of interest is observed. Whenever a user is identified, the service date on the claim of the exposure of interest observed during the relevant period of interest is labeled the “index date”.

The term “risk window” is used to represent the range of days, as measured from the index date of a particular member, in which outcomes or events of interest are required to occur in order to be included in the analysis. Similarly, the term “exclusion window” is used to represent the range of days, as measured from the index date of a particular member, in which occurrence of outcomes or events of interest disqualify the member from inclusion in analysis.

The term "study start" refers to the calendar day on which safety surveillance begins. Exposure data starting on this day are potentially eligible for inclusion in the analytic data set.

The term "look" refers to a single analysis within a sequential analysis framework. A look corresponds to a point in time at which data will be pulled from the MSCDM and analyzed for evidence of a safety signal. For a planned sequential analysis, the number and timing of looks will be planned in advance. The current analysis day is the last day for which exposure data is potentially eligible for inclusion in the analytic data set. It is expected that the current analysis day corresponds to a particular planned look and that data through the latest possible outcome risk window is complete.

IV. DATASET CREATION QUERY TOOL SUMMARY

Dataset creation query tool is used to create an analytic data set compatible with the specifications required by the GS GEE regression analysis. The analytic data set created by tool contains grouped data among a cohort of members treated with exposure(s) of interest or a comparator exposure during a period defined by a start date and an analysis day (*i.e.*, the query period where study day 1 corresponds to the start date). Data are aggregated by exposure status and confounder strata, whereby each row of the data set includes information about the frequency of select event(s) and the number of members (or person-time at risk) in each exposure-confounder stratum.

One run of tool generates one aggregated analytic data set (SAS table) and an optional individual level analytic data set. For more details on output tables, please see Section VII Each run of tool performs a complete refresh of the data from the study start to the current analysis time.

The tool requires the specification of several parameters to define a scenario. These include program parameters to specify a request identifier, scenario label, query period, and age range(s). The names of six input files (built as SAS datasets) containing several parameters must also be specified.

The first input file is the Analysis Plan File which defines the timing of the sequential analyses. The second input file is the Exposure Codes File which defines categories of exposures. The third file is the Exposure Categories File which identifies and codes the exposures of interest. The fourth file is the Outcome Codes File which defines events of interest. The fifth file is the Outcome Windows File which is used to define the risk and exclusion windows for events of interest. The sixth file is the Age Groups File; it lists the age categories to be retained in the final analysis set.

All parameters and input file specifications are described in Section V.

The default behavior of tool is that exposures of interest correspond directly to exposure categories. An optional SAS macro can be used to enable exposure of interest definitions that rely on multiple separate exposures. For more details on this option, please see Section V.C.

V. DATASET CREATION PROGRAM PARAMETER AND INPUT FILE SPECIFICATIONS

A. PROGRAM PARAMETER SPECIFICATIONS

There are 12 main program parameters that must be specified. These include a request identifier, scenario label, study start date and current analysis day for the query period, age stratifications, an indicator for production of individual-level data set, and six input files (Analysis Plan , Exposure Codes, Exposure Categories, Outcome Codes, Outcome Windows, and Age Groups). Three of these parameters are specified by the requester; nine parameters are specified by the request programmer, based on information provided by the requester.

Table 1 contains detailed specifications for each of these required parameters.

Table 1. Main Program Parameter Specification

Parameter	Field Name	Description
Request Identifier	REQUESTID	**Details**: a prefix added to output log and list files to track the various executions of the program. **Defined by:** Request programmer **Input type:** Required **Format:** Alphanumeric **Example**: REQUESTID =to08_seqmeth_wp1_b2
Scenario Label	ANASET_PREFIX	**Details**: a prefix added to output SAS data sets to track the various executions of the program. **Note:** Cannot exceed 7 characters. **Defined by:** Request programmer **Input type:** Required **Format:** Alphanumeric **Example**: ANASET_PREFIX =mmrv_
Query Start Date	ANASTART	**Details:** date for the start of the query identification period. If ANASTART =06Sep2005, only treatment episodes initiated on or after this date will be considered. **Defined by:** Requester **Input type:** Required **Format:** ddmmmyyyy **Example**: ANASTART =06Sep2005

Parameter	Field Name	Description
Analysis Day	ANADAY	**Details:** last day of the query identification period, measured in study days. If ANADAY=14, only treatment episodes initiated in the 14 days beginning with ANASTART will be considered. **Note:** ANASTART corresponds to study day 1. It is expected, but not enforced, that ANADAY corresponds to the study day of a planned look in the Analysis Plan File. **Defined by:** Requester **Input type:** Required **Format:** numeric **Example:** ANADAY =2681
Individual Level Creation	KEEP_IND_LEVEL	**Details:** used to specify the optional storage of an individual level data set. If KEEP_IND_LEVEL =Y then the individual level data set is saved locally at the DP site, otherwise it is erased upon program completion. **Note:** the interim level data set is required by the distributed portion of the risk difference (RD) regression analysis code. **Named by:** Request Programmer **Input type:** Optional **Format: .**sas7bdat **Example:** KEEP_IND_LEVEL =Y

Parameter	Field Name	Description
Age Group Definitions	AGE_STRAT	**Details:** age group categories for reporting. Specifying this parameter will (1) restrict to certain age groups and (2) specify how age groups will be stratified in the result tables. For example, to have results stratified by 20 year increments for members 40-99 years of age, AGESTRAT=40-59 60-79 80-99 would be entered. **Note 1:** age is determined at the index date. **Note 2:** various units of time can be used. Valid values are: • **D**: days • **W**: weeks • **Q**: quarters • **M**: months • **Y**: years (default value) **Note 3:** lower value is binding. If AGESTRAT=0-5 5-10, then all 5 year olds will be placed in the second age group. If AGESTRAT=0-5 6-10, then all 5 year olds will be placed in the first age group. For example, to have results stratified by 6 month increments for the first two years of life and then by 2 year increments until the age of 6, AGESTRAT = 00M-05M 06M-11M 12M-17M 18M-23M 02Y-03Y 04Y-05Y needs to be entered. **Defined by:** Requester **Input type:** Required **Format:** AA-AA BB-BB ZZ-ZZ **Example:** AGE_STRAT=%str(**00m-10m 11m-12m 13m-14m 15m-16m 17m-19m 20m-23m 24m+)**
Analysis Plan File	LOOK_PLAN	**Details:** name of the SAS dataset defining the look times of previous through current analyses. For specific details on the content of this file, see Section IV.B.1. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** LOOK_PLAN=mmrv_anaplan_2681

Parameter	Field Name	Description
Exposure Codes File	EXPOSURE_CODES	**Details:** name of the SAS dataset defining the query exposures of interest. It lists the codes of interest and their corresponding code type (e.g. NDC or PX-C4). For specific details on the content of this file, see Section IV.B.2. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** EXPOSURE_CODES=mmrv_exposures
Exposure Category File	EXPOSURE_CATS	**Details:** name of the SAS dataset mapping exposure categories to a numeric value. For simple analyses the exposure class of interest will be equated with 1 and the comparison group with 0. For specific details on the content of this file, see Section IV.B.3. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** EXPOSURE_CATS=mmrv_exposureCats
Outcome Codes File	OUTCOME_CODES	**Details:** name of the SAS dataset defining the outcomes of interest. It lists the codes of interest, corresponding code type (e.g. DX-09) and an outcome number to enable pulling data for the analysis of multiple outcomes at once. For specific details on the content of this file, see Section IV.B.4. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** OUTCOME_CODES=mmrv_outcomes
Outcome Windows File	OUTCOME_WINDOWS	**Details:** name of the SAS dataset defining, for each outcome of interest, the risk and exclusion windows as well as care settings for each outcome definition and exclusion criteria. For specific details on the content of this file, see Section IV.B.5. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** OUTCOME_WINDOWS=mmrv_windows

Parameter	Field Name	Description
Age Groups of Interest	AGEGROUPS	**Details:** name of the SAS dataset listing the age strata (defined via the Age Group Definitions parameter) of interest, all strata not listed in this table will be excluded from the final analytic tables. For specific details on the content of this file, see Section IV.B.6. **Named by:** Request programmer **Input type:** Required **Format: .**sas7bdat **Example:** AGEGROUPS =mmrv_ageGroups

B. INPUT FILE SPECIFICATIONS

In addition to the 12 main program parameters, several required parameters must be specified in the Analysis Plan, Exposure Codes, Exposure Categories, Outcome Codes, Outcome Windows, and Age Groups files.

1. Analysis Plan File

The Analysis Plan File is required. The analysis plan is required at the time of querying the MSCDM because analysis times define a level of stratification that data aggregation must reflect. For flexibility to conduct analyses we would specify the stratification to be either at weekly or monthly intervals. We will have a separate look plan file to conduct the formal analyses specified in the analysis program for GS GEE. There are two required parameters that must be specified; both must be specified by the analyst.

Table 2. Analysis Plan File Specification

Parameter	Field Name	Description
Look Number	LOOK	**Details:** index of a particular analysis within a sequential analysis plan. Table should have rows corresponding to looks 1…n. The current analysis for which the data is being pulled should correspond to one of these looks (not necessarily the last one). **Defined by:** Requester **Input type:** Required **Format:** Numeric **Example:** 3
Look Study Day	STOPDAY	**Details:** last study day contributing data to the analysis for the given look. **Defined by:** Requester **Input type:** Required **Format:** Numeric **Example:** 21

2. Exposure Codes File

The Exposure Codes File is required. It contains the comprehensive set of codes used to define the exposure classes which are used to define exposures of interest. National Drug Codes (NDCs), ICD-9-CM procedure codes, or Healthcare Common Procedure Coding System (HCPCS) codes can be used to define exposure(s) of interest. Exposure(s) can be defined using any mix of allowed code types.

The structure of the Exposure Codes File must reflect how codes should be queried to define a unique exposure. The GROUP field is used to group all codes pertaining to a given exposure of interest. For example, a group for "Exposure1" could be defined by all NDCs for any oral forms of anti-diabetic medications, a group for "Exposure2" by a mix of NDC and HCPCS codes for certain insulin products, and another group for "Exposure3" by only those NDCs for a recently approved oral form of anti-diabetic medication.

There are four required parameters that must be specified; all four must be specified by the requester.

Table 3 below describes specifications for the Exposure Codes File.

Table 3. Exposure Codes File Specification

Parameter	Field Name	Description
Query Code Category	CODECAT	**Details:** data type for code system being used, corresponding to data tables within the MS-CDM. Valid values are: • **NDC**: NDC codes contain in dispensings table • **PX**: Procedure codes contained in procedure table **Defined by:** Requester, with support from the Mini-Sentinel Operations Center (MSOC) as needed **Input type:** Required **Format:** Alphanumeric; SAS character $3 **Example:** PX

Parameter	Field Name	Description
Query Code Type	CODETYPE	**Details:** type of each procedure and/or drug code value included in the CODE field (below) of this file. Valid values are: **NDC** • **NDC**: length inferred from code field **PX** • **09**: ICD-9-CM procedure • **10**: ICD-10-CM procedure • **11**: ICD-11-CM procedure • **C4**: CPT-4 procedure (*i.e.*, HCPCS Level I) • **HC**: HCPCS procedure (*i.e.*, HCPCS Level II) • **H3**: HCPCS Level III procedure • **C2**: CPT Category II procedure • **C3**: CPT Category III procedure **Defined by:** Requester, with support from the Mini-Sentinel Operations Center (MSOC) as needed **Input type:** Required **Format:** Alphanumeric; SAS character $3 **Example:** C4
Code	CODE	**Details:** NDC and/or procedure code values to be used to define the exposure(s) of interest. **Note 1:** remove decimal points from the code value. **Note 2**: CODETYPE must be consistent with the expected format of the CODE value (*e.g.*, the program will not find any valid matches in the data for CODETYPE=PXC4 and a 3-digit code value). **Named by:** Requester **Input type:** Required **Format:** Alphanumeric; SAS character $11 **Example:** 90710

Parameter	Field Name	Description
Exposure Class	CLASS	**Details:** label for the exposure class to which the code corresponds. Class values may correspond directly to exposures of interest (e.g. MMRV) or to exposure classes that are used together to define a compound exposure of interest (e.g. MMR and V are two separate class values that could be used together to define a comparison exposure group of MMR+V). **Note 1:** remove decimal points from the code value. **Note 2**: CODETYPE must be consistent with the expected format of the CODE value (*e.g.*, the program will not find any valid matches in the data for CODETYPE=PXC4 and a 3-digit code value). **Note 3**: see Section IV.C. For more details on defining compound exposure classes (such as MMR+V). **Named by:** Requester **Input type:** Required **Format:** Alphanumeric; SAS character $10 **Example:** MMRV

3. Exposure Category File

The Exposure Category File is required. It contains the comprehensive list of exposures of interest and comparison exposures and assigns each a numeric code for the purpose of analysis. In most simple analyses, the exposure of interest should be assigned as 1 and the comparison exposure as 0. This assignment can then be used directly by the analysis programs without further modification.

The freedom to assign any codes is to facilitate pulling data for multiple analyses at once. For example, using a sensitivity analyses with alternative exposure definitions. At this time, analysis code requires a binary exposure variable and data pulls with multiple exposure codes will need to be further processed prior to running the analysis.

There are two required parameters that must be specified; both must be specified by the requester.

Table 4 describes the specification for the Exposure Category File.

Table 4. Exposure Category File Specification

Parameter	Variable Name	Description
Class	CLASS	**Details:** standardized name used to refer to a query GROUP for event(s) of interest to be queried. **Note 1**: must match a CLASS value from the Exposure Codes File or a variable name defined in the optional assign_compound_exposure() program described in Section IV.C. **Named by:** Requester **Input type:** Required **Format:** Alphanumeric; SAS character $10; no special characters (*e.g.*, commas, periods, hyphens, etc) allowed, and underscores must be used to mark spaces. **Example:** MMR_V
Name of Event Subgroup	EXPANACAT	**Details:** contains names of subgroups of event codes in this file. This variable is not used by the MP algorithm but it is useful for tracking purposes. **Named by:** Requester **Input type:** Required **Format:** Numeric **Example:** 0

4. Outcome Codes File

The Outcome Codes File is required. It contains the comprehensive set of codes used to refine the definition of the event(s) of interest. Event(s) of interest can be defined using any mix of ICD-9-CM diagnosis codes.

There are five required parameters that must be specified; four must be specified by the requester.

Table 5 contains detailed specifications for the Outcomes File.

Table 5. Outcome Codes File Specification

Parameter	Field Name	Description
Outcome Number	Num	**Details:** numbering system for outcome(s). **Note:** the final analytic data set will contain an outcome variable named Y_n for every outcome number designated. **Named by:** Requesting programmer **Input type:** Required **Format:** Numeric **Example:** 2
Query Code Category	CODECAT	**Details:** data type for code system being used, corresponding to data tables within the MS-CDM. Valid values are: • **DX**: Diagnosis codes contained in the diagnosis table **Defined by:** Requester, with support from the Mini-Sentinel Operations Center (MSOC) as needed **Input type:** Required **Format:** Alphanumeric; SAS character $2 **Example:** DX
Query Code Type	CODETYPE	**Details:** type of each procedure and/or drug code value included in the CODE field (below) of this file. Valid values are: **DX** • **09**: ICD-9-CM diagnosis • **10**: ICD-10-CM diagnosis • **11**: ICD-11-CM diagnosis **Defined by:** Requester, with support from the Mini-Sentinel Operations Center (MSOC) as needed **Input type:** Required **Format:** Alphanumeric; SAS character $2 **Example:** C4

Parameter	Field Name	Description
Code	CODE	**Details:** Diagnosis code values to be used to define the outcome(s) of interest. **Note 1:** remove decimal points from the code value. **Note 2**: CODETYPE must be consistent with the expected format of the CODE value (*e.g.*, the program will not find any valid matches in the data for CODETYPE=PXC4 and a 3-digit code value). **Named by:** Requester **Input type:** Required **Format:** Alphanumeric; SAS character $11 **Example:** 90710
Outcome Class	CLASS	**Details:** a label to identify the outcome of interest to which the code corresponds. **Named by:** Request Programmer **Input type:** Required **Format:** Alphanumeric; SAS character $1 **Example:** F (represents fever)

5. Outcome Windows File

The Outcome Windows File is required. It is used to specify the risk windows, as measured in days from the member's index date, for outcome events and exclusion events. Outcome events are the events of interest under surveillance. Exclusion events are the same outcome (e.g. fever), but occurring in a time period that disqualifies the exposure from being included in the analysis. For example, a child with a history of seizures prior to vaccination could be excluded from an analysis on the rate of seizures after vaccination through setting the exclusion window.

Additionally, the care setting of outcomes can be specified for events and exclusions separately. For example, the outcome of interest could be seizures occurring in an inpatient setting (settings: inpatient or emergency room) while history of seizures could be determined from any general care setting (settings: inpatient, emergency room, or ambulatory)

There are seven required parameters that must be specified; all of which must be specified by the requester.

Table 6 contains detailed specifications for the Outcome Windows File.

Table 6. Outcome Windows File Specification

Parameter	Variable Name	Description
Outcome Number	NUM	**Details:** standardized name used to refer to a query GROUP for exposure(s) of interest to be queried. **Note**: must match NUM values from the Outcomes File. **Named by:** Request programmer **Input type:** Required **Format:** Numeric **Example:** 1
Exclusion Window Begin	EXCL_L	**Details:** left endpoint of exclusion window, measured in days from the index date. **Note 1:** endpoints are included the risk window. **Defined by:** Requester **Input type:** Required **Format:** Numeric **Example:** -180 (Exclusion window begins 180 days prior to index date.)
Exclusion Window End	EXCL_U	**Details:** right endpoint of exclusion window, measured in days from the index date. **Named by:** Requester **Input type:** Required **Format:** Numeric **Example:** 0 (The most recent date in the exclusion window is the index date.)
Exclusion Settings	EXCL_SET	**Details:** contains the care settings considered for the exclusion outcomes. The following are valid entries; all entries must be quoted, separated by a space, and contained within parentheses: • **IP:** inpatient hospital stays • **IS:** non-acute institutional stays • **ED:** emergency department visits • **AV:** ambulatory visits • **OA:** other ambulatory visits **Defined by:** Requester **Input type:** Optional **Format:** Alphanumeric; SAS character $26. **Example:** ('AV','OA','IP','ED')

Parameter	Variable Name	Description
Event Window Begin	EVT_L	**Details:** left endpoint of event risk window, measured in days from the index date. **Note 1:** endpoints are included the risk window. **Defined by:** Requester **Input type:** Required **Format:** Numeric **Example:** 7 (Events occurring on the 7th through EVT_U day after the index date will be identified.)
Event Window End	EVT_U	**Details:** right endpoint of event risk window, measured in days from the index date. **Named by:** Requester **Input type:** Required **Format:** Numeric **Example:** 10 (Events occurring on EVT_L through the 10th day after the index date will be identified.)
Event Settings	EVT_SET	**Details:** contains the care settings considered for the outcome events. The following are valid entries; all entries must be quoted, separated by a space, and contained within parentheses: • **IP:** inpatient hospital stays • **IS:** non-acute institutional stays • **ED:** emergency department visits • **AV:** ambulatory visits • **OA:** other ambulatory visits **Defined by:** Requester **Input type:** Optional **Format:** Alphanumeric; SAS character $26. **Example:** ('IP','ED')

6. Age Groups File

The Age Groups File is required. This table lists all of the age categories, as determined by the main program Age Stratification parameter, to be kept in the final analytic table.

There is one required parameters that must be specified; it must be specified by the requester.

Table 7 contains detailed specifications for the Age Group File.

Table 7. Age Groups File Specification

Parameter	Variable Name	Description
Name of Age Group	AGECAT	**Details:** standardized name used to refer to a query GROUP for exposure(s) of interest to be queried. **Note**: must match values created by MS_AGESTRAT macro in accordance with AGESTRAT value from the main program. **Named by:** Requester **Input type:** Required **Format:** Alphanumeric; SAS character $10; no special characters (*e.g.*, commas, periods, hyphens, etc) allowed, and underscores must be used to mark spaces. **Example:**17m-19m

C. COMPOUND EXPOSURES

The exposure codes file is used to indicate the exposure status of each individual to a product class. In many cases a product class will directly correspond to an exposure of interest (e.g. receipt of MMRV vaccine). However, there are times when a combination of exposures will define a natural exposure group of interest (e.g. a natural comparison to receipt of MMRV vaccine is concurrent receipt of MMR and V vaccines). In this case we call the exposure a compound exposure.

To facilitate compound exposures a SAS macro entitled assign_compound_exposure is used within the program. The use of this macro is optional. In cases in which the requestor has indicated a compound exposure is of interest or needed for comparison, the requesting programmer can program a macro by this name and define compound exposures in terms of the individual exposure classes defined by the exposure codes file. The requesting programmer should include the macro definition in SAS after the main program has been defined, so that the default macro of the same name is overloaded.

See Section VIII for an example of the assign_compound_exposure macro.

VI. DATASET CREATION PROGRAM STEPS

The general program steps are:

1. Process the six input files
2. Exposures: Extract medical claims from the diagnosis and procedure files
3. Exposures: Extract drug claims from the outpatient pharmacy file
4. Exposures: Combine exposure records into one table and add subject demographics (age at index date and sex)
5. Cohort: limit exposure records to age of interest, roll-up to one record per exposure index-date, define compound exposures and code final exposures of interest
6. Outcomes: Extract medical claims from the diagnosis and procedure files
7. Outcomes: Flag exclusion and events for each exposure

8. Stage the records into the final analytic dataset format
9. Aggregate data by strata (age, sex, and outcome exclusion/event status)

VII. DATASET CREATION PROGRAM EXECUTION

When implementing query programs within the MSDD, the Mini-Sentinel Operations Center (MSOC) uses a uniform folder structure across Data Partners to facilitate communications between MSOC and Data Partners and to streamline file management.

Each request distributed by MSOC is assigned a unique Request ID. Upon receipt of the request, Data Partners create a folder named after the Request ID and several subfolders to organize program inputs and outputs. One of the folders contains output to be sent to MSOC and another contains intermediate files that remain with the Data Partner, but could be used to facilitate follow-up queries if necessary. Appropriate retention policies apply.

Table 8 defines the local environment variables that must be initialized by the user to execute the program (*i.e.*, defined by the Data Partner prior to execution of the program). Please note that these values cannot be left blank. Each Data Partner is required to enter user inputs at the beginning of the SAS Program sent with each request. These inputs are unique to each Data Partner.

Table 8. Environment Variable Definitions

Label	Field Name	Description
Data Partner ID	DPID	Enter the two character partner ID.
Site ID of Data Partner	SITEID	Enter the two character Site ID.
Demographics Table Name	DEMTABLE	Enter the name of the MSCDM Demographics table.
Dispensing Table Name	DISTABLE	Enter the name of the MSCDM Dispensing table.
Diagnosis Table Name	DIATABLE	Enter the name of the MSCDM Diagnosis table.
Procedures Table Name	PROCTABLE	Enter the name of the MSCDM Procedures table.
Label	**Field Name**	**Description**
Input file folder	INFOLDER	Enter the path where the input files will be saved.
Output file folder	MSOC	Enter the path where the shared output tables will be saved.
Dataset file folder	DPLOCAL	Enter the path where the local SAS datasets will be saved.
Libname of the MSCDM	INDATA	Enter the path where the MSCDM data is saved.

VIII. DATASET CREATION OUTPUT TABLES

One aggregate output table is always created by the modular program and output in .sas7bdat format to the output file folder. A second individual level output table is created optionally, if the individual creation level argument is set to Y, and output in .sas7bdat format to the dataset file folder. Tables are named based on environment variables and program arguments. In terms of SAS environment variables, the aggregate output table is named: &DPID.&SITEID._&ANASET_PREFIX._m3anaset_&ANADAY._ag and the optional individual level table is named similarly but without the _ag suffix.

Aggregate Output Table: per specification required of the Analysis Module 3 (Regression) Sequential Analysis Program (required for relative difference analyses):

DPID	SITEID	AgeGroup	Sex	X	S	Y_1_Excl	…	Y_n_Excl	N_Obs	Obs_t	Y_1	…	Y_n
HM	GHC	11m-12m	F	0	7	0	…	0	100	100	1	…	0
HM	GHC	11m-12m	F	0	7	1	…	0	5	5	0	…	0

Interpretation of the aggregate output table: This table is the stratified aggregation of the individual level output table (across PatID-IndexDate pairs). N_obs and obs_t represent the total number of observations and exposure days within each stratum. In this example the exposure is a vaccine (one time exposure) and thus each observation contributes one exposure day (N_obs=obs_t).

The stratification variables are: AgeGroup, Sex, X (Exposure Code), S, and Y_1_Excl … Y_n_Excl (an exclusion flag for each event). The per strata summary variables are: N_obs (number of observations), Obs_t (total days exposed), and Y_1 … Y_n (total number of events).

The first line of the above table indicates that among 11-12 month old girls vaccinated with the comparison vaccines (X=0 indicates MMR+V) at HM-GHC, with no events during the exclusion window for any outcome of interest, and entering the analysis on the 7th study day there was one outcome event of type 1 (Seizure) and no outcome events of any other type (Y_2 … Y_n) out of the 100 one-time exposures. The second line indicates that among a similar set of children (age-sex-exposure) there were 5 vaccinations for which the recipient had a history of seizure (Y_Excl_1=1) and no outcome events of any type (Y_1 … Y_n = 0).

Individual Level Output Table: per specification required of the Analysis Module 3 (Regression) Sequential Analysis Program (required for risk difference analysis)

DPID	SITEID	AgeGroup	Sex	X	S	PatID	IndexDate	startDay	startWeek	Obs_t	Y_1_Excl	Y_1	…	Y_n_Excl	Y_n
HM	GHC	11m-12m	F	0	7	XXX1	10/19/2001	5	1	1	0	0	…	0	0
HM	GHC	11m-12m	F	0	7	XXX2	10/20/2001	6	1	1	1	0	…	0	1

Interpretation of the individual level output table: This table contains one record per PatID-IndexDate pair identified for the cohort. The variable startDay is the indexDate translated to the day of study timescale (i.e. where the anaStart parameter corresponds to day 1). The startWeek variable has been added to facilitate uptake diagnostics at a scale likely to be more refined than typical analysis plans. For each of the outcome event types (numbered in the Outcome Code and Outcome Windows Files) there are two indicator variables, one if an outcome was observed in the exclusion window, relative to the indexDate, (Y_Excl_#) and one if an outcome was observed in the risk window, relative to the indexDate (Y_#). Records with an exclusion flagged will not be included in the sequential analysis of the corresponding outcome. In this example the exposure is a vaccine (one time exposure) and thus obs_t equals 1 for each observation.

For the purposes of aggregation, the start time (S) is reported as the last day within the analysis period the observation entered the data set. (For example, startdays 5 and 6 are recorded as 7 when the sequential analysis has been occurring weekly.) The values of the S variable correspond to study days of looks stated in the Analysis Plan File and are reported in the aggregated output table that will be returned to the MSOC.

IX. DATASET CREATION PROGRAM EXAMPLE

Tables 9-14 below show partially-populated examples of the Analysis Plan, Exposure Codes, Exposure Category, Outcome Codes, Outcome Windows, and Age Groups Files used to create the aggregate output table described in Section VII:

Table 9. Examples of Analysis Plan File

Look	StopDay
1	7
2	14
…	…
52	364

This example of an analysis plan corresponds to pulling the 52nd weekly analysis (looks 1 through 52 with stopDay increasing by 7 between each look).

Table 10. Example of Exposure Codes File

Code	CodeCat	CodeType	Class
90710	PX	C4	MMRV
9948	PX	07	MMR
00064826	NDC	NDC	V

Table 11. Example of Exposure Category File

Class	ExpAnaCat
MMRV	1
MMR_V	0

Table 12. Example of Outcome Codes File

Code	CodeCat	CodeType	Class	Num
780.6	DX	09	F	2
345	DX	09	S	1
780.3	DX	09	S	1

Table 13. Example of Outcome Windows File

Num	Excl_l	Excl_U	Excl_Set	Evt_l	Evt_U	Evt_Set
1	-180	0	('AV','OA','IP','ED')	7	10	('IP','ED')
2	-5	0	('AV','OA','IP','ED')	1	5	('AV','OA','IP','ED')

Table 14. Example of Age Groups File

AgeCat
11m-12m
13m-14m
15m-16m
17m-19m

In this example, the requester additionally specified the following parameters to be used:

- The individual level output table should be saved at the local DP site (e.g. risk difference sequential analyses will be performed subsequently)
- The sequential analysis period begins on 06 September 2005
- The current analysis queries the first 2,685 days of the ongoing surveillance period
- The results should be stratified according to the following age groups: 11-12,13-14,15-16,17-19 months
- The exposure classes MMR and V should be combined into one exposure group of interest labeled MMRV.

For this request, the compound exposure definition requires an additional piece of code to be written by the request programmer, as follows:

```
%macro assign_compound_exposure();
   *define new exposure class;
        MMR_V=0;
        if (MMR=1 & V=1) then do;
                MMR_V = 1;
                MMR = 0;
                V=0;
        end;
%mend;
```

For this request, the program could be executed using the following SAS macro call:

```
%PULL_ANASET_M3(REQUESTID=to08_seqmeth_wp1_b2,
                ANASET_PREFIX=mmrv,
                LOOK_PLAN=infolder.anaPlan,
                ANASTART=06Sep2005,
                ANADAY=2681,
                AGESTRAT=%str(00m-10m 11m-12m 13m-14m 15m-16m 17m-19m 20m+)
                AGEGROUPS=infolder.ageGroups,
                EXPOSURE_CODES= infolder.exposures,
                EXPOSURE_CATS= infolder.exposureCats,
                OUTCOME_CODES= infolder.outcomes,
                OUTCOME_WINDOWS= infolder.outcome_windows,
                KEEP_IND_LEVEL=Y);
```

X. GS GEE ANALYSIS PARAMETERS

Given the aggregated analysis dataset from the analytic dataset query pull program that has been outlined in the previous sections the next step is to conduct the formal GS GEE regression analysis. To conduct the analysis and automate the report we have created R code that uses the aggregated analysis datasets from all of the sites. The following analysis parameters are required to be specified by the user.

Parameter	Variable Name	Description
Analytic Dataset	Dat	**Details:** One dataset that combines all aggregated datasets from the data pull program across sites and then de-aggregates the data to create an individual level dataset. **Named by:** Request programmer **Input type:** Required **Format:** csv
Confounder Locations in Dataset	Zi	**Details:** which columns of the dataset are the confounders to be included in the model **Defined by:** Request Programmer **Input type:** Required **Format:** Numeric Vector **Example:** Zi=c(2:3)
Confounder Labels for report	Zlab	**Details:** Names of the confounders and categories to be used to make the report more readable. **Defined by:** Request Programmer **Input type:** Optional **Format:** Character List **Example:** `Zlab=list(Age=rbind(c("11m-12m","11m-12m"),c("13m-14m","13m-14m"),c("15m-16m","15m-16m"),c("17m-19m","17m-19m"),c("20m-23m","20m-23m")),Sex=rbind(c("M","Male"),c("F","Female")))`
Exposure labels	xlab	**Details:** Names of the exposure groups to be used to make the report more readable **Defined by:** Request Programmer **Input type:** Optional **Format:** Character vector **Example:** `xlab<-c("MMR+V","MMRV")`
Outcome labels	outlab	**Details:** Name of the outcome to be used to make the report more readable **Defined by:** Request Programmer **Input type:** Optional **Format:** Character **Example:** `outlab<-"Seizure”`
Analysis Times	LTime	**Details:** Either name of dataset or specified in code defining the look times of previous through current analyses. **Named by:** Request programmer **Input type:** Required **Format:** csv or R code **Example:** LTime=mmrv_anaplan_2681.csv or `LTime=c(364,seq(364+91,1274,91))`

Parameter	Variable Name	Description
Current Analysis Time label	lablook	**Details:** Name of the look plan to be descriptive for the report **Defined by:** Request Programmer **Input type:** Optional **Format:** Character **Example:** `lablook<-"Quarterly Looks with One Year Lag"`
Final planned study sample size	Nend	**Details:** Sample size that the study will be finished if no signal occurs **Defined by:** Requester **Input type:** Required **Format:** Numeric **Example:** `Nend<-80000`
Current analysis number	CurLook	**Details:** The current analysis that we are looking at the data. **Defined by:** Request Programmer **Input type:** Required **Format:** Numeric **Example:** `CurLook<-3`
Shape of the sequential monitoring boundary	boundshape	**Details:** Shape of the sequential monitoring boundary **Defined by:** Request Programmer **Input type:** Required **Format:** character (“Fleming” or “Pocock” **Example:** `boundshape<-“Pocock”`
Previous analysis time sequential boundaries	PrevBounds	**Details:** After the first analysis time we will need to input a file with the values of the previous sequential boundaries used from previous analysis times. **Defined by:** Request Programmer **Input type:** Required **Format:** numeric csv **Example:** `outprevbounds.look3.csv`
Indicator of analysis time adjustment	TimeAdj	**Details:** Whether the analyses should include analysis time as an additional confounder **Defined by:** Request Programmer **Input type:** Required **Format:** Boolean **Example:** `TimeAdj=TRUE`

XI. EXAMPLE ANALYSIS REPORT

The analysis report for the GS GEE method includes a main section with the key results and appendices with additional detail. The main section is comprised of a methods summary face page, demographics table, exposure uptake figure, and a primary surveillance results table. The appendix contains results by data partner site, analysis time, and demographics. A full discussion of an example report can be found on the Mini-Sentinel website.