MINI-SENTINEL METHODS

FEASIBILITY OF NEW MINI-SENTINEL GROUP SEQUENTIAL MONITORING METHODS IN A DISTRIBUTED SETTING - IMPLEMENTATION IN PRACTICE

Prepared by: Andrea J Cook, PhD[1,2], Robert D Wellman, MS[1], Azadeh Shoaibi, MS, MHS[4], Ram C Tiwari, PhD[5], Denise Boudreau, PhD[3], Sunali Goonesekera, SM[6], Tracey L Marsh, MS[2,3], and Jennifer C Nelson, PhD[1,2]

Author Affiliations: 1. Biostatistics Unit, Group Health Research Institute, Seattle, WA 2. Department of Biostatistics, University of Washington, Seattle, WA 3. Group Health Research Institute, Seattle, WA 4. Office of Medical Policy, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 5. Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 6. Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA

October 21, 2013

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 Methods

Feasibility Of New Mini-Sentinel Group Sequential Monitoring Methods In A Distributed Setting - Implementation In Practice

Table of Contents

I. OBJECTIVES AND DELIVERABLES

The purpose of this workgroup was to demonstrate the feasibility of at least one new sequential method in the Mini-Sentinel Distributed Database (MSDD) setting. Specifically this project was to build on the work from two previous Mini-Sentinel workgroups (1) Year 1 'Sequential Testing Methods Development'[1] and (2) Year 2 'Enhancing current sequential analytic techniques to improve causal inference'[2]. This workgroup was to implement in the MSDD at least one or more of the statistical methods previously developed and further to create statistical code that can be used in a semi-automated fashion. Specifically we were to demonstrate the technical capability within the MSDD to automate both the distributed (run at Data Partner) portion of the code to obtain de-identified data, or summary information, that is returned to the coordinating center and then statistical analysis code that will be run by the coordinating center to conduct formal statistical analysis for at least one of the sequential methods developed in the previous workgroups. We were to test the code at three Data Partner sites within the MSDD. We were further asked to automate the code to be conducted for multiple interim analyses or looks at the data.

The workgroup was able to implement and apply the following two group sequential methods: Group Sequential Generalized Estimating Equation (GS GEE) regression approach and the Group Sequential Inverse Probability of Treatment Weighting (GS IPTW) regression approach. For each method the following two sets of code were developed (1) distributed data portion and (2) statistical code that takes the results from the distributed code and completes the full application of the specific method. The distributed portion of the code was particularly challenging due to the limitations of the available SAS procedures licensed across different data partners. Specifically, we initially programmed the GS IPTW method using the IML package in SAS which is a common program for statisticians who develop new methods, but one of the Data Partners did not have this part of the SAS license. We were able to reprogram using less efficient SAS procedures, but a lesson learned for future method development is that some of the standard statistical software that statisticians use to apply new methods may not be available. Further, the group developed statistical code to automate the report of the results which we will go over in detail later in this document. Initially we were developing the reports in SAS, but we moved this part of the code to R since we were able to create more flexible automated reports that were easier to read and interpret. Within this document we will first describe for each group sequential method the parameters that are required initially to run such a method and then we will go over the results evaluating if the combination vaccine MMRV has higher rates of febrile seizures compared to two separate injections of the MMR and V (MMR+V) vaccine. For the example we go back in time and mimic if we actually started studying the new combination vaccine MMRV at 4 sites when it first went on the market on September 5, 2005.

II. MS PROMPT: GS IPTW

A. SUMMARY OF THE METHOD

Group sequential regression using inverse probability of treatment weighting (GS IPTW) performs site stratified IPTW regression estimation in a group sequential testing framework. It first derives at each site a site-specific adjusted risk difference (RD) estimate based on IPTW with weights from propensity scores adjusted for categorical baseline confounders. It then uses these site-specific RDs to calculate an overall

stratified RD across all sites. The method is designed for new user cohort designs where a short-term exposure is of interest.

1. Design

The method assumes an active-comparator new user cohort design in which there is an exposure of interest and an alternative, concurrently used, control exposure. An unexposed control group could also be used, but this design is likely to be less common in Mini-Sentinel. The method is designed for short term exposures, including one-time exposures (e.g., an injection) or an exposure that occurs for a relatively short period of time (e.g., an antibiotic)). It assumes a binary indicator of either being exposed or comparator and a binary indicator of the occurrence of an adverse event outcome of interest within a pre-specified risk window following product initiation. A person is not included in an analysis until their outcome risk window has been fully observed so that all subjects have the same duration of follow-up time.

2. Statistical Analysis

GS IPTW is a flexible approach that uses IPTW regression to control for baseline categorical confounders and estimates a site-stratified adjusted risk difference (RD)[2-4]. Specifically, it fits a propensity score model at each site and then calculates a site-specific adjusted RD using IPTW with the site specific propensity scores as weights. It further calculates the variance of each site adjusted RD estimate incorporating the fact that the weights are estimated from a model and are not known. These site-specific adjusted RD estimates and site-specific variances of the RD are sent to a central location. Here, this information is combined to calculate an overall site stratified RD estimate and corresponding variance. One then calculates a standardized IPTW test statistic (IPTW test=RD/sqrt(var(RD))) and compares this statistic to a preset signaling threshold. If the statistic exceeds the threshold, then a signal is generated and surveillance stops. If the statistic does not exceed the threshold, then monitoring continues. An important advantage of this approach is that it strongly controls for site confounding and accounts for potential interactions by site and other confounders. It has also been shown to be as efficient as a non-stratified estimate when no site interaction with other baseline confounders exists. Further, the method only requires that a minimum of one event occur at a given site in order to be able to estimate a RD. In contrast, methods that involve ratio estimators (e.g., relative risks) require that an event occur in both the exposed and unexposed group before estimation can begin.

To conduct group sequential monitoring for rare outcomes, GS IPTW uses a non-parametric permutation approach that flexibly simulates data under the null hypothesis of no elevated risk in the exposure group (i.e. Ho: RD=0) as opposed to making large sample normal approximation assumptions. It uses a unifying boundary approach[5] to define the boundary based on the permuted data, thus incorporating the concepts of both stopping at earlier analysis times and repeated testing (See Appendix for Statistical Detail). It requires that the user specify the desired number of analysis times or 'looks,' the timing of each analysis (based on observed or expected sample size at each analysis time), and a total maximum sample size at which the end of surveillance will occur if no signal has been detected. Boundary shape can also be flexibly specified and the choice will be dependent on the desired level of the signaling thresholds earlier versus later in the surveillance period. A flatter boundary will signal earlier for lower elevated risk but will have less power to signal later relative to a boundary that employs early conservatism and has a higher threshold early on and more power later. Boundary shape is quantified on the scale of a standardized test statistic and is chosen based on what decisions rules are desired (e.g.,

low boundaries early are only desirable if a specific action is expected to be taken if the statistic exceeds that boundary) in combination with statistical criteria. Given the desired boundary shape function and the permuted IPTW test statistic under the null, an IPTW test statistic stopping boundary is computed that is used to assess whether there is an elevated risk at each analysis time point or if no meaningful difference exists and surveillance should continue. This boundary is designed to hold the overall false positive signaling rate at a pre-specific level (e.g., 0.05).

3. Specific Input Parameters That Need to Be Specified to Conduct Analyses

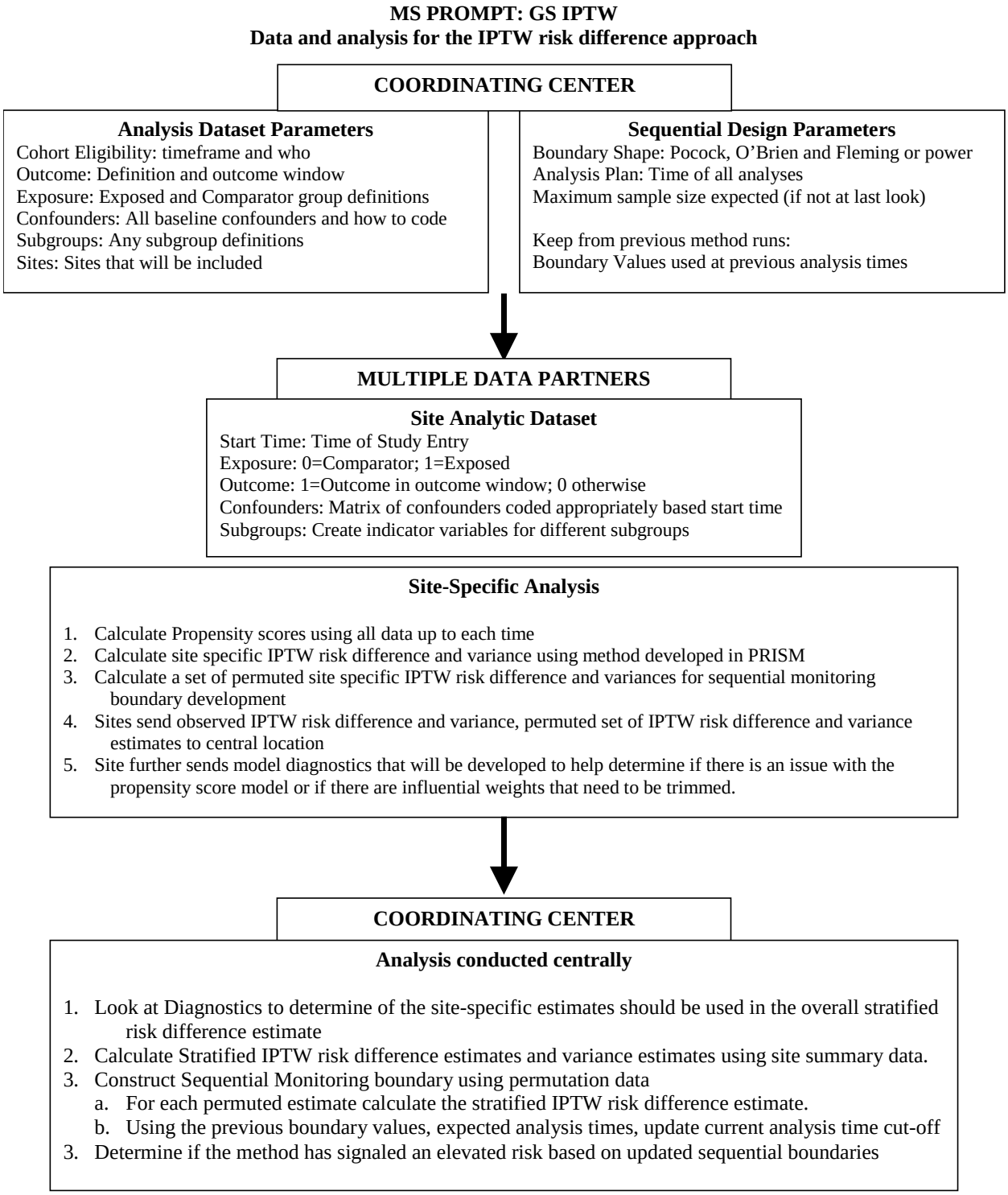
We assume that the user has already specified a standard set of dataset parameters (see Figure 1 Analysis Dataset Parameters) and used them to create a prospective new user cohort. Specifically, the exposure and comparator groups of interest have been specified, and the outcome of interest has been defined as being binary (i.e., occurring within a pre-specified outcome window after receipt of the exposure of interest or comparator product). Further, a set of relevant confounders have been defined and categorized (e.g., age has been categorized as 20-29yrs, 30-39yrs, and so on).

Once the analytic dataset is created based on the dataset parameters, a second set of parameters are required to specify the sequential monitoring and application of the GS IPTW method (see Figure 1 Sequential Design Parameters). Figure 1 further shows the flow of how the data and method moves from the central coordinating center and is distributed to sites for those interested in further details of the process. These additional parameters specify the details for how we will conduct prospective surveillance analyses with testing at multiple time points for evidence of an increased risk of a specific outcome in the exposed group of interest (typically a new medical product) compared to a comparator group. First, we must specify a 'look plan' that designates when each analysis will occur. Then the shape of the boundary must also be decided a priori. The existing code allows for flexibility in this choice, including specification of a flat boundary on the scale of the standardized test statistic, which has been used in previous safety surveillance evaluations. The user must also determine what sample size will be sufficient at the end of surveillance if no signal occurs (i.e., the maximum sample size). The maximum sample size is typically chosen to yield a certain level of power by the end of the surveillance to rule out excess risk that would concern the monitoring agency. This maximum sample size needed to achieve a specific level of power will vary depending on the look plan (more frequent monitoring requires a larger maximum sample size), shape of boundary (flatter boundary shape requires a larger maximum sample size), baseline rate of outcome (larger maximum sample for lower baseline outcome rates), proportion exposed (further away from 50% exposed the larger the maximum sample size), and confounder strength (stronger confounding larger maximum sample size). The complete list of sequential design parameters needed to run an analysis are summarized below:

Method and Sequential Parameters

- Outcome Type: Binary
- Adjust for Analysis Time: TRUE (only option)
- Maximum Sample Size: maximum sample size when surveillance will be complete given no signal
- Look Plan: plan of when the analysis will occur (e.g. first analysis after 30,000 observations and then 10 evenly spaced looks up to the Maximum sample size.)
- Boundary Shape: Pocock[6], O'Brien Fleming[7], or power function

Figure 1. Flow Diagram of the MS PROMPT GS IPTW method including parameter specification and in general how the method works in the distributed data setting



B. APPLICATION TO MMRV AND MMR+V

We will now walk through the semi-automated report of surveillance results for the vaccine example comparing risk of febrile seizure 7-10 days after receipt of MMRV (exposure of interest) versus MMR+V (comparator) among children 12-23 months of age. We will first review the results from Analysis 1 in detail. Then we will go over a subset of the results on Analysis 4, when a signal was detected. This review is intended to serve as a guide both to summarize the test case findings from the specific vaccine example and to showcase more generally what information can currently be provided when running the PROMPT: GS IPTW module. The following analysis and sequential parameters have been set for this example:

- Outcome Type: Binary
- Maximum Sample Size: 118,328
- Look Plan: Look 1 at 1 year (12 mths), then quarterly looks for an additional 2.5 years (30 mths) (i.e., at 364(12 mths), 455(15 mths), 546(18 mths), 637(21 mths), 728(24 mths), 819(27 mths), 910(30 mths), 1001(33 mths), 1092(36 mths), 1183(39 mths), and 1274(42 mths) days since start of surveillance)
- Boundary Shape: Pocock

1. Analysis 1

We will first go over Analysis 1. We will start with a summary of the main features and inputs and then discuss tables that provide further detail.

a. Primary Summary

Figure 2 summarizes the main features and inputs for the current analysis, including the comparison groups, list of confounders, indicator of adjustment for look, look plan, maximum sample size, and boundary shape. Table 1 summarizes the demographics of the current analysis dataset by exposure group, and Figure 3 describes the uptake of the exposure of interest and comparator (Figure 3).

Table 2 contains the main surveillance results, including data from all analyses up to the current look. Correspondingly, Analysis 1 Table 2 contains a single row with results from the first look. Analysis 1 occurred on day 364 from initial start date (9/6/2005). Among 12,652 new users of MMR+V and 2,726 new users of MMRV, 0.040% (5 events) and 0.179% (5 events) had febrile seizures, respectively. After adjusting for age and sex using IPTW within each site and combining the site-specific RD into a site stratified estimate across all sites, the adjusted seizure rate among those exposed to MMR+V was 0.043% compared to 0.074% for those exposed to MMRV, yielding a RD of 0.031%. The IPTW standardized RD test statistic was 0.794, which did not cross the critical boundary of 1.297 (based on unifying Pocock boundary calculations, see Appendix for details) and so there was no signal. The first look spent 0.025 of the available total cumulative alpha of 0.05.

In addition to these first three parts of the primary results, it may also be important to assess differences by site, analysis time, and demographics to see if one site or confounder subgroup is especially influential, or if there is an indication that the results vary by analysis time. The next section will describe the appendix tables that address these important issues.

Figure 2. GS IPTW parameters for MMRV and MMR+V example (taken directly from the Title Page of the output)

- Method: MS PROMPT GS IPTW Stratified Risk Difference
- Brief Description: Site Stratified Adjusted Risk Difference method applied using site Inverse probability of treatment weighting with sequential monitoring boundaries based on permutations.
- Estimate: Stratified adjusted Risk Difference
- Exposure of Interest: MMRV
- Control Comparator: MMR+V
- Outcome: Seizure
- Confounders: Age, Sex within each Site
- Adjusted for Look Time: Yes
- Look Times(Days): 364, 455, 546, 637, 728, 819, 910, 1001, 1092, 1183, 1274
- Maximum Sample Size: 269561
- Boundary: Pocock using Unifying Boundary approach with 10000 permutations to derive boundary

Table 1. GS IPTW for look 1 showing demographics of population by exposure group

	Total N (%)	MMR + V N (%)	MMRV N (%)
Total	15448 (100.0)	12652 (81.9)	2796 (18.1)
Age in months			
11-12	8276 (53.6)	7156 (56.6)	1120 (40.1)
13-14	3053 (19.8)	2360 (18.7)	693 (24.8)
15-16	2547 (16.5)	1905 (15.1)	642 (23.0)
17-19	1075 (7.0)	835 (6.6)	240 (8.6)
20-23	497 (3.2)	396 (3.1)	101 (3.6)
Sex			
Male	7909 (51.2)	6497 (51.4)	1412 (50.5)
Female	7539 (48.8)	6155 (48.6)	1384 (49.5)
Site			
4	2829 (18.3)	2806 (22.2)	23 (0.8)
15	6402 (41.4)	3686 (29.1)	2716 (97.1)
16	6217 (40.2)	6160 (48.7)	57 (2.0)

Figure 3. GS IPTW for look 1 showing uptake over time by exposure group

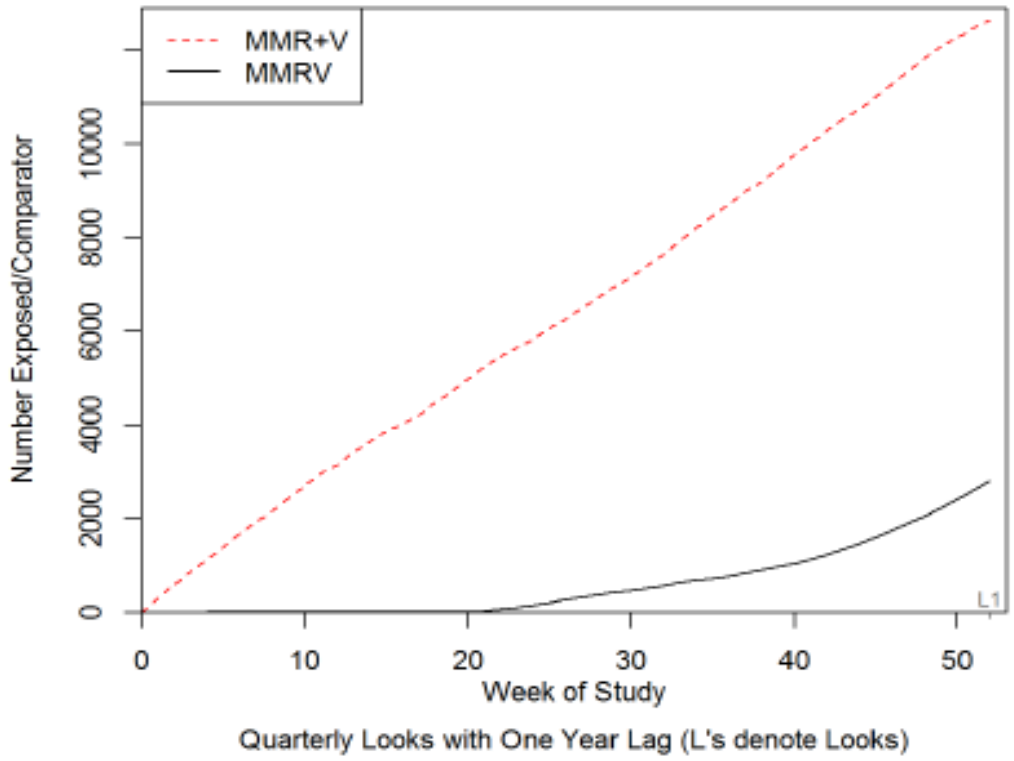


Figure 1: Total uptake of MMRV and MMR+V over Time at Look 1

Table 2. GS IPTW at look 1 showing primary results

Look	Days	MMR+V N	MMR+V Outcome (%)	MMRV N	MMRV Outcome (%)	MMR+V Adjusted % outcome	MMRV Adjusted % outcome	Adjusted Risk Difference *	IPTW test	Boundary	Error spent	Signal
1	364	12652	5(0.040)	2796	5(0.179)	0.043	0.074	0.031	0.794	1.297	0.025	No

*Adjusted stratified risk difference model applied using GS IPTW with sequential monitoring boundaries based on permutations.
Covariates included: Age, Sex and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting. Outcome(%)=Number(Risk%) of outcome within look and covariate category. Adj=Adjusted. RD=Risk Difference, Adj %Out=Adjusted Risk % from stratified IPTW model for a given exposure group, Adj RD=MMRV Adj %Out – MMR+V Adj %Out = stratified IPTW adjusted RD %. IPTW Test = Adj RD/Standard Error (Adj RD), and Boundary = Sequential Boundary to compare the IPTW Test Estimate.

b. Results by Analysis Time Point and Confounder Strata

The first semi-automated table, Table 3, provides a summary of outcome counts and risk percent by analysis time point and by confounder strata. For Analysis 1, this table contains only one column, but new columns will be added automatically as each new analysis is conducted. The table further provides information about the potential strength of confounders. As shown in Table 3, at Analysis 1 the percent with seizure is higher in older age groups compared to lower age groups, higher in females compared to males, and higher at site 15 compared to the other sites. Note there are only 10 outcomes in total and so the differences may not be statistically meaningful.

Table 3. GS IPTW at look 1 showing the first appendix table displaying outcome counts and risk % by look and covariate strata for MMRV and MMR+V

	Outcome (%) for Look 1
Total	10 (0.065)
Age in months	
11-12	3 (0.036)
13-14	3 (0.098)
15-16	2 (0.079)
17-19	1 (0.093)
20-23	1 (0.201)
Sex	
Male	4 (0.051)
Female	6 (0.080)
Site	
4	1 (0.035)
15	7 (0.109)
16	2 (0.032)

*Abbreviations: Outcome (%) = Number (Risk %) of outcome within look and covariate stratum.

The second semi-automated table, Table 4 summarizes outcome counts and risk percent by analysis time point and by confounder strata among the exposed group of interest (MMRV). For Analysis 1, this table only contains one column, but new columns will be added automatically as each new analysis is conducted. This table is designed to provide information about whether or not there is an interaction between a given confounder and exposure. Recall that if there is an interaction between exposure and site that the GS IPTW method implicitly accounts for this as it conducts a stratified analysis by site.

Table 4. GS IPTW at look 1 displaying outcome counts and risk % by look and covariate strata among MMRV only

	Outcome (%) for Look 1
Total	5 (0.179)
Age in months	
11-12	1 (0.089)
13-14	1 (0.144)
15-16	2 (0.312)
17-19	1 (0.417)
20-23	0 (0.000)
Sex	
Male	3 (0.212)
Female	2 (0.145)
Site	
4	0 (0.000)
15	5 (0.184)
16	0 (0.000)

*Abbreviations: Outcome (%) = Number (Risk %) of outcome within look and covariate stratum.

Table 5. GS IPTW at look 1 displaying demographics across analysis times

	N (%) for Look 1
Total, N (Row %)	15448 (100.0)
Age in months, N (Col %)	
11-12	8276 (53.6)
13-14	3053 (19.8)
15-16	2547 (16.5)
17-19	1075 (7.0)
20-23	497 (3.2)
Sex, N (Col %)	
Male	7909 (51.2)
Female	7539 (48.8)
Site, N (Col %)	
4	2829 (18.3)
15	6402 (41.4)
16	6217 (40.2)

The third semi-automated table, Table 5, provides demographics by analysis time. This displays any potential changes in the demographics of the entire cohort over time**.** Additionally, Table 6 provides demographics over time focusing exclusively on the exposure group of interest (MMRV). When a new medical product first comes onto the market, often only a subset of the population is initially exposed. Once the product is available for a longer period of time, it will often infiltrate a larger part of the market space. One of the attractive aspects of the IPTW method is that, under certain assumptions, the results are generalizable to the entire population who may eventually use the new product instead of just to those initially exposed. This is due to how the weighting in the IPTW method specifically, weights from IPTW are used to upweight those that were unlikely to receive their observed exposure and downweight those who were likely to receive their observed exposure based on their baseline confounder distribution. Those equally likely to receive either treatment are neutrally weighted. This process evens out the baseline covariate distribution to allow estimation of an unconfounded average effect in the entire population given no unmeasured confounders. However, to make this population estimate it is best to have some coverage of all confounders observed among both the exposed and comparator groups. Further this may explain differences compared to other statistical approaches such as exposure matching, which generalizes to a population that resembles the exposed group. These tables can be especially informative if the overall adjusted RD estimate changes substantially across analysis time points, potentially due to changes in the confounder distribution over time (e.g., New site is added after analysis 3).

Table 6. GS IPTW at look 1 displaying demographics across analysis times among MMRV (report Table A.4)

	N (%) for Look 1
Total, N (Row %)	2796 (100.0)
Age in months, N (Col %)	
11-12	1120 (40.1)
13-14	693 (24.8)
15-16	642 (23.0)
17-19	240 (8.6)
20-23	101 (3.6)
Sex, N (Col %)	
Male	1412 (50.5)
Female	1384 (49.5)
Site, N (Col %)	
4	23 (0.8)
15	2716 (97.1)
16	57 (2.0)

Figure 4 displays the uptake by site. For this example we included 4 sites, but at Analysis 1 there was no exposure uptake at one site so it was not included in the analysis. These figures show immediate, but slow, uptake at sites 4 and 16. At site 15, uptake was delayed until about week 20 and then it increased quickly.

Figure 4. GS IPTW at look 1 displaying uptake of MMR+V and MMRV for each site

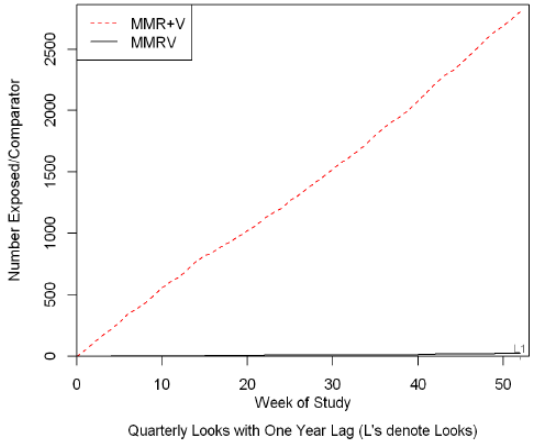


Figure A.1: Uptake of MMR+V and MMRV over Time for Site 4

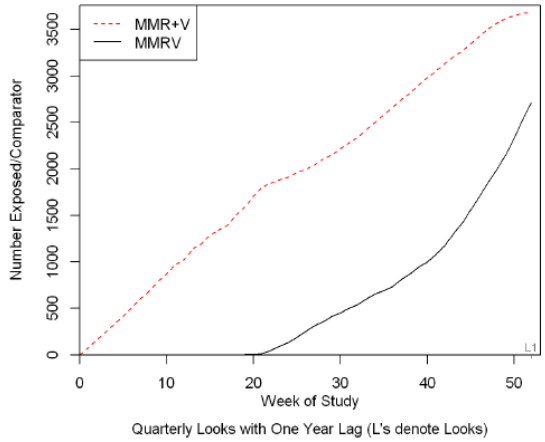


Figure A.2: Uptake of MMR+V and MMRV over Time for Site 15

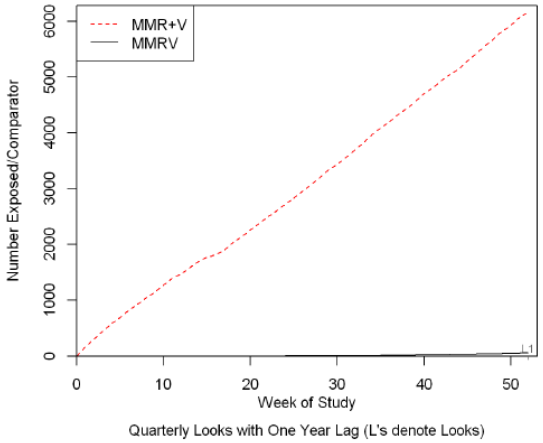


Figure A.3: Uptake of MMR+V and MMRV over Time for Site 16

Table 7, shows the site specific RD results at the current analysis. This table is important for assessing whether there is consistency in the estimated RD across sites and how the amount of information contributed by each site varies. For this example, the Analysis 1 results are strongly driven by site 15, as only 23 and 57 children at sites 4 and 16 were exposed to MMRV relative to 2,716 at site 15. Further, there are only one and two outcomes total at sites 4 and 16, respectively, compared to site 15 with 7 outcomes total.

Table 7. Current analysis (look 1) site specific results

Site	MMR+V N	MMR+V Outcome (%)	MMRV N	MMRV Outcome (%)	MMR+V Adjusted % outcome	MMRV Adjusted % outcome	Adjusted Risk Difference*
4	2806	1(0.036)	23	0(0.000)	0.036	0.000	-0.036
15	3686	2(0.054)	2716	5(0.184)	0.056	0.178	0.121
16	6160	2(0.032)	57	0(0.000)	0.032	0.000	-0.032

*Adjusted risk difference model applied using IPTW for each site (no Sequential).
Covariates Included: Age, Sex and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting. Outcome(%)=Number(Risk%) of outcome within look and covariate category. Adj=Adjusted. RD=Risk Difference, Adj %Out=Adjusted Risk % from site-specific IPTW model for a given exposure group. Adj RD=MMRV Adj %Out – MMR+V Adj %Out=site-specific IPTW adjusted RD %.

2. Analysis 4 When a Signal Was First Detected

We will now review an example for an analysis in which the sequential boundary is crossed and there is a signal indicating an elevated risk in the exposure of interest (MMRV). For the MMRV versus MMR+V example this occurred at Analysis 4 on day 637 (1.75 years after the start of surveillance). We will first look at the Table 8 information showing demographics by exposure group. At Analysis 4, we have a cohort of 34,823 children in which 17,502 received MMR+V and 17,321 received MMRV. Similar distributions of demographic characteristics were observed in both exposure groups except for site 15 in which 81.9% of MMRV recipients compared to only 23.9% of MMR+V recipients reside. We then visually display uptake over time by exposure group from GS IPTW Analysis 4 in Figure 5. Figure 5 shows that uptake stayed relatively steady for the MMR+V, but increased dramatically over time for MMRV.

The main results for Analysis 4 are displayed in GS IPTW Table 9. The last row shows the primary surveillance results for Analysis 4 which was conducted at day 637. Among 17,502 new users of MMR+V, 7 (0.040%) had a febrile seizure. Among 17,321 new users of MMRV, 18 (0.104%) had a febrile seizure. After adjusting at each site for age and sex using IPTW and combining site-specific RD estimates across sites into a site stratified estimate, the adjusted seizure rate among MMR+V was 0.028% compared to MMRV of 0.080% yielding an adjusted risk difference of 0.052%. The IPTW standardized RD test statistic was 2.235, which exceeded the critical boundary of 2.069 and so there was a statistically significant signal. For an analysis in which we observe a signal, we provide a sequential adjusted analysis p-value at the bottom of the footer. In this example, the sequential p-value was 0.0319 and we were able to detect an elevated risk of febrile seizure of 0.052% RD (1 extra febrile seizure per 1,923 exposed to MMRV relative to being given MMR+V) after 637 days of surveillance (or after 17,321 exposed children to MMRV). We further note that this elevated adjusted RD was consistent across analyses.

Table 8. GS IPTW for look 4 showing demographics of population by exposure group for MMRV and MMR+V example

	Total	MMR+V	MMRV
Total, N (Row%)	34823 (100.0)	17502 (50.3)	17321 (49.7)
Age, N (Col%)			
11m-12m	17728 (50.9)	10089 (57.6)	7639 (44.1)
13m-14m	7038 (20.2)	3267 (18.7)	3771 (21.8)
15m-16m	6171 (17.7)	2434 (13.9)	3737 (21.6)
17m-19m	2681 (7.7)	1143 (6.5)	1538 (8.9)
20m-23m	1205 (3.5)	569 (3.3)	636 (3.7)
Sex, N (Col%)			
Male	17798 (51.1)	9040 (51.7)	8758 (50.6)
Female	17025 (48.9)	8462 (48.3)	8563 (49.4)
Site, N (Col%)			
4	5090 (14.6)	4981 (28.5)	109 (0.6)
15	18353 (52.7)	4175 (23.9)	14178 (81.9)
16	11380 (32.7)	8346 (47.7)	3034 (17.5)

Figure 5. GS IPTW for look 4 showing uptake over time by exposure group for MMRV and MMR+V example

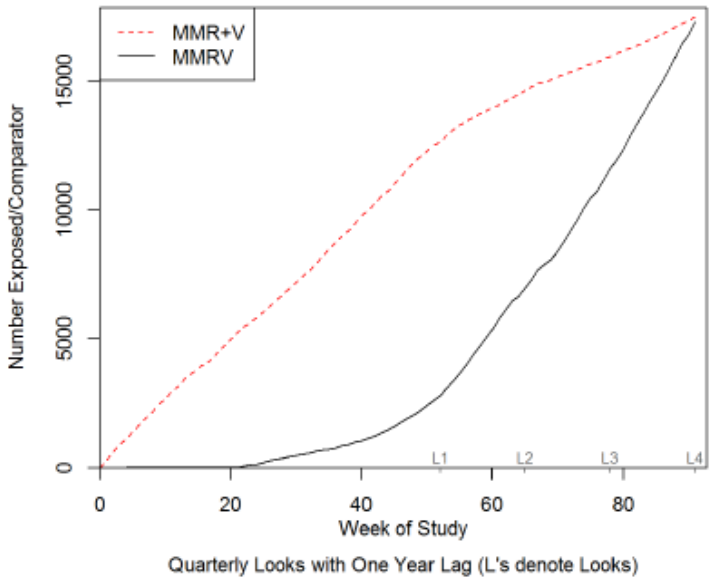


Figure 1: Total uptake of MMRV and MMR+V over Time at Look 4

Table 9. GS IPTW at look 4 showing primary results for MMRV and MMR+V example

Look:	Days	MMR+V N	MMR+V Outcome (%)	MMRV N	MMRV Outcome(%)	MMR+V Adj %Out	MMRV Adj %Out	Adj RD*	IPTW Test	Boundary	Error Spent	Signal
1:	364	12652	5(0.040)	2796	5(0.179)	0.043	0.074	0.031	0.794	1.297	0.028	No
2:	455	14633	7(0.048)	6970	10(0.143)	0.045	0.088	0.043	1.261	2.043	0.032	No
3:	546	15968	7(0.044)	11577	14(0.121)	0.036	0.086	0.051	1.790	2.074	0.032	No
4:	637	17502	7(0.040)	17321	18(0.104)	0.028	0.080	0.052	2.235	2.069	0.032	Yes

*Adjusted stratified risk difference model applied using GS IPTW with sequential monitoring boundaries based on permutations.
Covariates Included: Age, Sex, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Outcome(%)=Number(Risk %) of outcome within look and covariate category, Adj=Adjusted, RD=Risk Difference, Adj %Out=Adjusted Risk % from stratified IPTW model for a given exposure group, Adj RD= MMRV Adj%Out – MMR+V Adj %Out = stratified IPTW adjusted RD %, IPTW Test =Adj RD/Standard Error(Adj RD), and Boundary = Sequential Boundary to compare the IPTW Test Estimate.
Sequential P-Value at Signal: 0.0319

a. Appendix Results for GS IPTW at Signal (Look 4)

We will now look more into the data to assess if there is any obvious issues that might make us question our results. Table 10 shows how the outcome incidence rate changed over analysis time by confounder strata. Over time there are relatively similar incidence rates, and nothing seems to be changing drastically by site which is a good indication that coding practices did not change substantially over the surveillance period.

Table 10. GS IPTW at look 4 showing outcome and incidence rates by look and covariate strata

	Look 1	Look 2	Look 3	Look 4
Total, Outcome(%)	10 (0.065)	17 (0.079)	21 (0.076)	25 (0.072)
Age, Outcome(%)				
11m-12m	3 (0.036)	6 (0.053)	7 (0.050)	9 (0.051)
13m-14m	3 (0.098)	4 (0.094)	5 (0.090)	5 (0.071)
15m-16m	2 (0.079)	2 (0.054)	2 (0.041)	3 (0.049)
17m-19m	1 (0.093)	3 (0.192)	4 (0.195)	4 (0.149)
20m-23m	1 (0.201)	2 (0.275)	3 (0.329)	4 (0.332)
Sex, Outcome(%)				
Male	4 (0.051)	8 (0.072)	10 (0.071)	12 (0.067)
Female	6 (0.080)	9 (0.085)	11 (0.082)	13 (0.076)
Site, Outcome(%)				
4	1 (0.035)	1 (0.028)	1 (0.023)	1 (0.020)
15	7 (0.109)	10 (0.100)	13 (0.095)	17 (0.093)
16	2 (0.032)	6 (0.075)	7 (0.073)	7 (0.062)

Table 11 displays the site-specific RD analysis results. Here we see that the elevated risk is only occurring at Site 15, with an adjusted RD of 0.100%. The estimated RD at site 16 is close to null (0.006%) and Site 4 has little exposure uptake yielding little influence on the results. Therefore, if we were truly conducting this surveillance activity, to confirm our results it would be advantageous to attempt to add additional sites since the elevated rate is only observed in one site.

Table 11. GS IPTW showing Look 4 analysis site specific results

	Look 1	Look 2	Look 3	Look 4
Total, Outcome(%)	5 (0.179)	10 (0.143)	14 (0.121)	18 (0.104)
Age, Outcome(%)				
11m-12m	1(0.089)	4 (0.135)	5 (0.100)	7 (0.092)
13m-14m	1 (0.144)	2 (0.129)	3 (0.118)	3 (0.080)
15m-16m	2 (0.312)	2 (0.124)	2 (0.075)	3 (0.080)
17m-19m	1 (0.417)	2 (0.338)	3 (0.295)	3 (0.195)
20m-23m	0 (0.000)	0 (0.000)	1 (0.255)	2 (0.314)
Sex, Outcome(%)				
Male	3 (0.212)	5 (0.141)	7 (0.120)	9 (0.103)
Female	2 (0.145)	5 (0.146)	7 (0.122)	9 (0.105)
Site, Outcome(%)				
4	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
15	5 (0.184)	8 (0.131)	11 (0.133)	15 (0.106)
16	0 (0.000)	2 (0.240)	3 (0.166)	3 (0.099)

***Abbreviations: Outcome (%)=Number(Risk %) of outcome within look and covariate stratum.**

III. MS PROMPT: GS GEE

A. SUMMARY OF THE METHOD

This method performs regression estimation and group sequential testing using categorical baseline confounder adjustment through GEE for new user cohorts.

1. Design

The design assumes an active-comparator new user cohort design in which there is an exposure of interest and a concurrent control exposure. For short term exposures (i.e. one-time exposure (injection) or a short period of time (antibiotic)) it assumes a binary indicator of being exposed or unexposed and a binary outcome window within a pre-specified risk window following product initiation. We assume that a person is not included in the analysis until the completion of the outcome risk window has been fully observed so that all short term exposures have the same follow-up time.

For longer term exposure (i.e. drug taken over several months or years) it only uses the first indication of taking either the exposure of interest or comparator and counts the length of being exposed up to time stopping taking the either exposure (with a lag if desired), if the outcome of interest occurs, or if the person stops being enrolled. The outcome is assumed to be binary, occurring or not occurring, while the person is still being exposed (with a lag if desired).

2. Statistical Analysis

Group Sequential regression using GEE (GS GEE) is a flexible approach that uses regression to control for baseline categorical confounders. It uses a general GEE framework that can handle different exposure and outcome types. Specifically, for short term exposures it assumes a binomial outcome with a logit

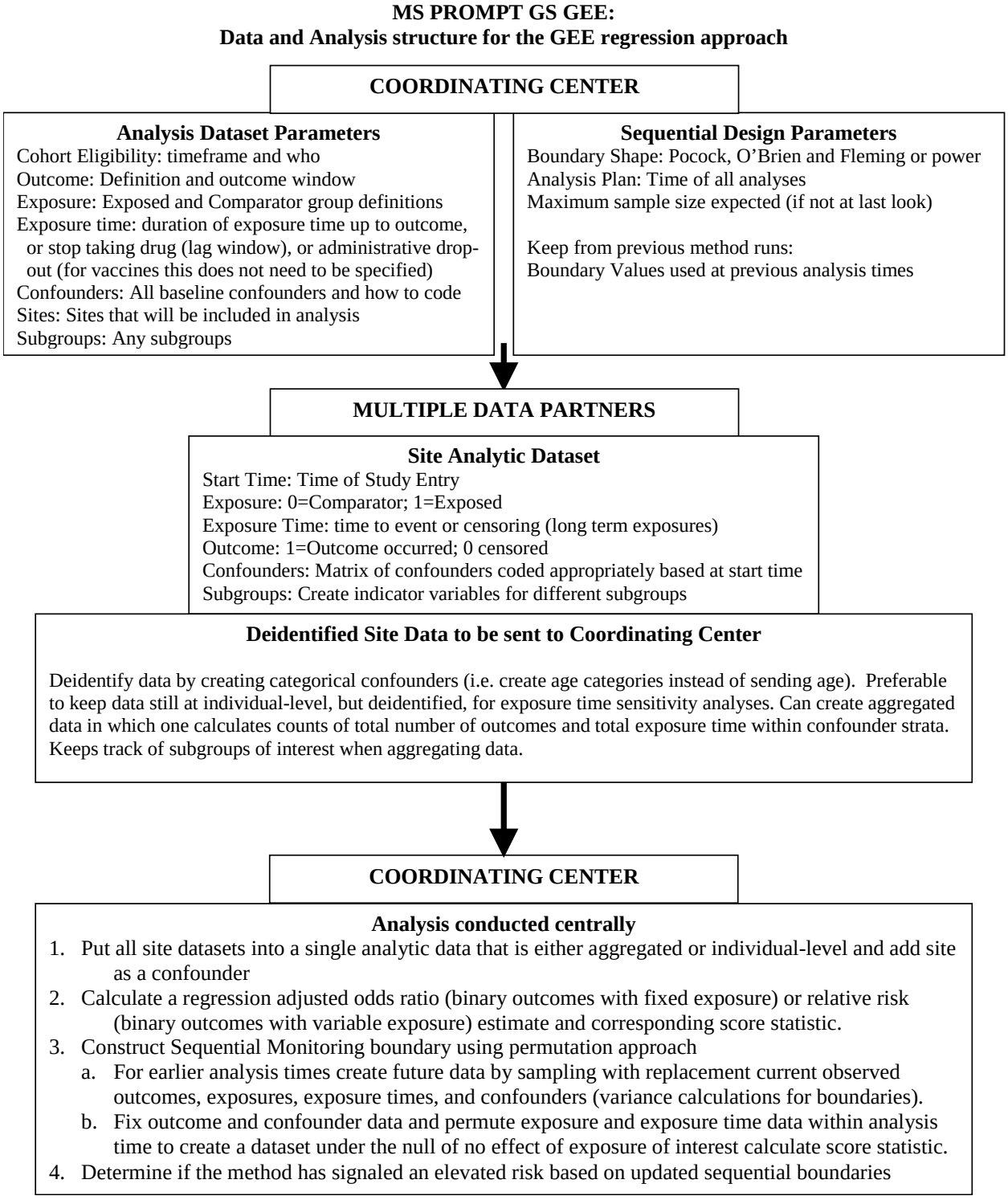
link function to estimate an adjusted odds ratio (OR) as the measure of elevated risk in an exposure group relative to a comparator group. For chronic exposures a Poisson regression approach is used that takes into account length of exposure and estimates an adjusted relative rate (RR). For both outcome types the GS GEE method calculates a score test statistic and uses this standardized test statistic to inform signaling for elevated risk or continuation of monitoring.

To incorporate group sequential monitoring, GS GEE uses a non-parametric permutation approach, that is particularly suited for rare outcomes, by flexibly simulating data under the null hypothesis of no elevated risk in the exposure group (i.e. OR=RR=1). It uses the unifying boundary approach[5] to define the boundary based on the permuted data, thus incorporating the concepts of both stopping at earlier analysis times and repeated testing (See Appendix for Statistical Details). It requires a specified number of analysis times, timing of analyses (based on observed, or expected sample size, at each analysis time), and a total expected sample size by end of study. Boundary shape is dependent on desired signaling thresholds earlier versus later in a study. Flatter boundaries will signal earlier for lower elevated risk, but will have less power to signal later in the study relative to boundaries that have a higher threshold to signal early in the study yielding more power to signal later. Boundary shape is viewed based on a standardized test statistic chosen based on what decisions rules are desired (i.e. low boundaries early are only desirable if an action is to be taken for a given signal) in combination with statistical criteria. Given the boundary shape function and permuted score test statistic under the null, we use this boundary to yield a score test statistic boundary that will be the stopping boundary for indicating if there is an elevated risk or if the study is to continue.

3. Specific Parameters That Need to Be Specified to Conduct Analyses

We assume that the user has already specified a standard set of dataset parameters (see Figure 6 Analysis Dataset Parameters) and used them to create a prospective new user cohort. Specifically, the exposure of interest and comparator have been defined and the outcome of interest has been defined as being either binary (i.e. happened within a pre-specified outcome window after being exposed to the exposure of interest or comparator) or Poisson type outcome that incorporates variable exposure time (i.e. outcome is 1 if that outcome occurs while being exposed to either the exposure of interest or comparator (with a potential outcome lag if that exposure stopped and allow outcomes a fixed window after exposure stopping) and 0 otherwise; follow-up time is from initiation of exposure of medical product up to either having the outcome, censoring due to disenrollment from health plan, death, or stopped taking exposure (with a potential outcome lag)). Further, a set of relevant confounders have been defined and categorized (e.g., age has been categorized as 20-29yrs, 30-39yrs, and so on).

Figure 6. Flow Diagram of the MS PROMPT GS GEE method including parameter specification and in general how the method works in the distributed data setting



Once the analytic dataset is created based on the dataset parameters, a second set of parameters are required to specify the sequential monitoring and application of the GS GEE method (see Figure 6 Sequential Design Parameters). Figure 6 further shows the flow of how the data and method moves from the central coordinating center and is distributed to sites for those interested in further details of the process. These additional parameters specify the details for how we will conduct prospective surveillance analyses with testing at multiple time points for evidence of an increased risk of a specific outcome in the exposed group of interest (typically a new medical product) compared to a comparator group. First, we must specify a 'look plan' that designates when each analysis will occur. Then the shape of the boundary must also be decided a priori. The existing code allows for flexibility in this choice, including specification of a flat boundary on the scale of the standardized test statistic, which has been used in previous safety surveillance evaluations. The user must also determine what sample size will be sufficient at the end of surveillance if no signal occurs (i.e., the maximum sample size). The maximum sample size is typically chosen to yield a certain level of power by the end of the surveillance to rule out excess risk that would concern the monitoring agency. This maximum sample size needed to achieve a specific level of power will vary depending on the look plan (more frequent monitoring requires a larger maximum sample size), shape of boundary (flatter boundary shape requires a larger maximum sample size), baseline rate of outcome (larger maximum sample for lower baseline outcome rates), proportion exposed (further away from 50% exposed the larger the maximum sample size), and confounder strength (stronger confounding larger maximum sample size). The complete list of sequential design parameters needed to run an analysis are summarized below:

Method and Sequential Parameters

- Outcome Type: Binary
- Adjust for Analysis Time: NO or YES (Adjust for indicator of analysis time)
- Maximum Sample Size: maximum sample size when surveillance will be complete given no signal
- Look Plan: plan of when the analysis will occur (e.g. first analysis after 30,000 observations and then 10 evenly spaced looks up Maximum Sample Size)
- Boundary Shape: Pocock[6], O'Brien Fleming[7], or power function

B. APPLICATION TO MMRV AND MMR+V

We will now walk through the surveillance results for the vaccine example comparing risk of febrile seizure 7-10 days after receipt of MMRV (exposure of interest) versus MMR+V (comparator) among children 12-23 months of age. We will first review Analysis 1 in detail. Then we will go over a subset of the results on Analysis 8, when a signal was detected. This review is intended to serve as a guide both to summarize the test case findings from the specific vaccine example and to showcase more generally what information can currently be provided when running the PROMPT: GS GS GEE module. The following analysis and sequential parameters have been set for this example:

- Outcome Type: Binary
- Adjust for Analysis Time: YES
- Maximum Sample Size: 118,328
- Look Plan: Look 1 at 1 year, then quarterly looks for an additional 2.5 years (i.e., at 364(12mths), 455(15mths), 546(18mths), 637(21mths), 728(24mths), 819(27mths), 910(30mths), 1001(33mths), 1092(36mths), 1183(39mths), and 1274(42mths) days since start of surveillance)

- Boundary Shape: Pocock

1. Analysis 1

We will first go over each page of the report generated for Analysis 1. We will start with a summary of the main features and inputs for Analysis 2 and will then discuss tables that provide further detail.

a. Primary Summary

Figure 7 summarizes the main features and inputs for the current analysis, including the comparison groups, list of confounders, indicator of adjustment for look, look plan, maximum sample size, and boundary shape. Table 12 summarizes the demographics of the current analysis dataset by exposure group, and Figure 8 displays the uptake of the exposure of interest and comparator.

Table 13 provides main surveillance results, including data from all analyses up to the current look. Correspondingly, Analysis 1 contains a single row with results from the first look. Analysis 1 occurred on day 364 from initial start date (9/6/2005). Among 12,652 new users of MMR+V and 2,726 new users of MMRV, 0.040% (5 events) and 0.179% (5 events) had febrile seizures, respectively. After adjusting for age, sex and site using GS GEE, the adjusted seizure rate among those exposed to MMR+V was 0.043% compared to 0.131% for those exposed to MMRV, yielding an adjusted OR of 3.05. The GS GEE standardized Score Test Statistic was 1.842, which did not cross the critical boundary of 3.407 and so there was no signal (See Statistical Appendix, Section VI, for how boundary was calculated). The first look spent <0.001 of the available total cumulative alpha of 0.05.

In addition to these first three parts of the primary results, it may also be important to assess differences by site, analysis time, and demographics to see if one site or confounder subgroup is especially influential, or if there is an indication that the results vary by analysis time. The next section will describe these important issues.

Figure 7. GS GEE parameters for MMRV and MMR+V example (taken directly from the Title Page of the output)

- Method: MS PROMPT GSGEE Logistic Regression
- Brief Description: Adjusted logistic regression model applied using GEE framework with sequential monitoring boundaries based on permutations.
- Estimate: Adjusted Odds Ratio
- Exposure of Interest: MMRV
- Control Comparator: MMR+V
- Outcome: Seizure
- Confounders: Age, Sex and Site
- Adjusted for Look Time: Yes
- Maximum Sample Size: 15448
- Look Times(Days): 364, 455, 546, 637, 728, 819, 910, 1001, 1092, 1183, 1274
- Boundary: Pocock using Unifying Boundary approach with 10000 permutations to derive boundary

Table 12. GS GEE for look 1 showing demographics of population by exposure group

	Total	MMR+V	MMRV
Total, N(Row%)	15448 (100.0)	12652 (81.9)	2796 (18.1)
Age, N(Col%)			
11m-12m	8276 (53.6)	7156 (56.6)	1120 (40.1)
13m-14m	3053 (19.8)	2360 (18.7)	693 (24.8)
15m-16m	2547 (16.5)	1905 (15.1)	642 (23.0)
17m-19m	1075 (7.0)	835 (6.6)	240 (8.6)
20m-23m	497 (3.2)	396 (3.1)	101 (3.6)
Sex, N(Col%)			
Male	7909 (51.2)	6497 (51.4)	1412 (50.5)
Female	7539 (48.8)	6155 (48.6)	1384 (49.5)
Site, N(Col%)			
2	0 (0.0)	0 (0.0)	0 (0.0)
4	2829 (18.3)	2806 (22.2)	23 (0.8)
15	6402 (41.4)	3686 (29.1)	2716 (97.1)
16	6217 (40.2)	6160 (48.7)	57 (2.0)

Figure 8. GS GEE for look 1 showing uptake over time by exposure group

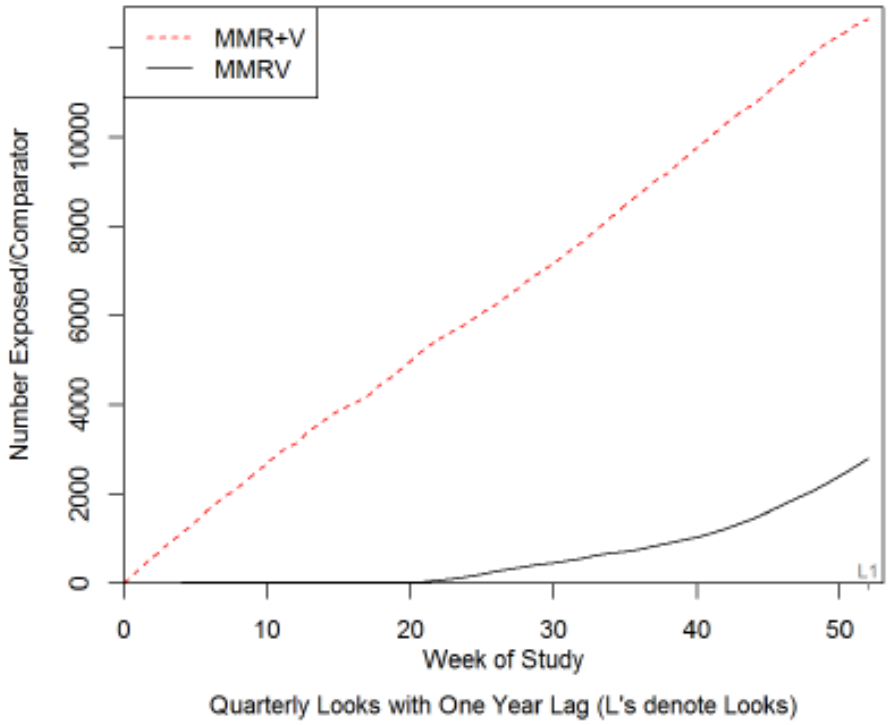


Figure 1: Total uptake of MMRV and MMR+V over Time at Look 1

Table 11. GS GEE report table 2 at look 1 showing primary results

Look:	Days	MMR+V N	MMR+V Outcome(%)	MMRV N	MMRV Outcome(%)	MMR+V Adj %Out	MMRV Adj %Out	Adj OR*	Score Test	Boundary	Error Spent	Signal
1:	364	12652	5(0.040)	2796	5(0.179)	0.043	0.131	3.05	1.842	3.407	0.000	No

***Adjusted logistic regression model applied using GEE framework with sequential monitoring boundaries based on permutations.**
Covariates Included: Age, Sex, Site, and indicator for each look.
Abbreviations: Outcome(%)=Number(Risk %) of outcome within look and covariate category, Adj= Adjusted, Adj %Out= Adjusted risk % from adjusted logistic regression model assuming entire population was either exposed or unexposed, Adj OR= Adjusted Odds Ratio comparing MMRV to MMR+V from logistic regression model, Score Test = Score Test statistic from GEE logistic regression model, and Boundary = Sequential Boundary to compare the Score Test Estimate.

b. Results by Analysis Time Point and Confounder Strata

Table 14 provides a summary of outcome counts and incidence rates by analysis time point and by confounder strata. For Analysis 1, this table contains only one column, but new columns will be added automatically as each new analysis is conducted. The table further provides information about the potential strength of confounders. As shown in Table 14, at Analysis 1 the percent with seizure is higher in older age groups compared to lower age groups, higher in females compared to males, and higher at site 15 compared to the other sites. Note there are only 10 outcomes in total and so the differences may not be statistically meaningful.

Table 12. GS GEE at look 1 displaying outcome counts and risk % by look and covariate strata for MMRV and MMR+V example

	Look 1
Total, Outcome (%)	10 (0.065)
Age, Outcome (%)	
11m-12m	3 (0.036)
13m-14m	3 (0.098)
15m-16m	2 (0.079)
17m-19m	1 (0.093)
20m-23m	1 (0.201)
Sex, Outcome (%)	
Male	4 (0.051)
Female	6 (0.080)
Site, Outcome (%)	
2	0 (NaN)
4	1 (0.035)
15	7 (0.109)
16	2 (0.032)

*Abbreviations: Outcome(%)=Number(Risk %) of outcome within look and covariate stratum.

The second semi-automated appendix table, Table 15 (or Table A2 in the report) summarizes outcome counts and incidence rates by analysis time point and by confounder strata among the exposed group of interest (MMRV). For Analysis 1, this table only contains one column, but new columns will be added automatically as each new analysis is conducted. This table is designed to provide information about whether or not there is an interaction between a given confounder and exposure.

Table 13. GS GEE at look 1 displaying outcome counts and risk % by look and covariate strata among MMRV only

	Look 1
Total, Outcome (%)	5 (0.179)
Age, Outcome (%)	
11m-12m	1 (0.089)
13m-14m	1 (0.144)
15m-16m	2 (0.312)
17m-19m	1 (0.417)
20m-23m	0 (0.000)
Sex, Outcome (%)	
Male	3 (0.212)
Female	2 (0.145)
Site, Outcome (%)	
2	0 (NaN)
4	0 (0.000)
15	5 (0.184)
16	0 (0.000)

*Abbreviations: Outcome(%)=Number(Risk %) of outcome within look and covariate stratum.

Table 14. GS GEE at look 1 displaying demographics across analysis times

	Look 1
Total, N (Row%)	15448 (100.0)
Age, N (Col%)	
11m-12m	8276 (53.6)
13m-14m	3053 (19.8)
15m-16m	2547 (16.5)
17m-19m	1075 (7.0)
20m-23m	497 (3.2)
Sex, N (Col%)	
Male	7909 (51.2)
Female	7539 (48.8)
Site, N (Col%)	
2	0 (0.0)
4	2829 (18.3)
15	6402 (41.4)
16	6217 (40.2)

Table 16 provides demographics by analysis time. This displays any potential changes in the demographics of the entire cohort over time**.** Additionally, Table 17 provides demographics over time focusing exclusively on the exposure group of interest (MMRV). When a new medical product first

comes onto the market, often only a subset of the population is initially exposed. Once the product is available for a longer period of time, it will often infiltrate a larger part of the market space. These tables can be especially informative if the adjusted OR changes substantially across analysis time points, potentially due to changes in the confounder distribution over time.

Table 15. GS GEE at look 1 displaying demographics across analysis times among MMRV

	Look 1
Total, N (Row%)	2796 (100.0)
Age, N (Col%)	
11m-12m	1120 (40.1)
13m-14m	693 (24.8)
15m-16m	642 (23.0)
17m-19m	240 (8.6)
20m-23m	101 (3.6)
Sex, N (Col%)	
Male	1412 (50.5)
Female	1384 (49.5)
Site, N (Col%)	
2	0 (0.0)
4	23 (0.8)
15	2716 (97.1)
16	57 (2.0)

Figure 9 displays the uptake by site. For this example we included 4 sites, but at Analysis 1 there was no uptake at one site so it was not included in the analysis. These figures show immediate, but slow, uptake at sites 4 and 16. At site 15, uptake was delayed until about week 20 and then it increased quickly.

Figure 9. GS GEE at look 1 displaying uptake of MMR+V and MMRV for each site (Excludes site 2 due to no uptake)

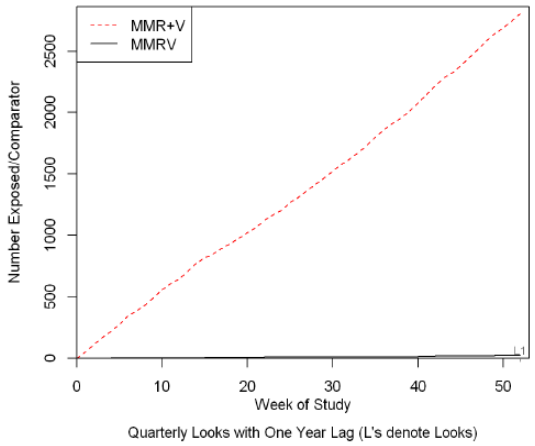


Figure A.1: Uptake of MMR+V and MMRV over Time for Site 4

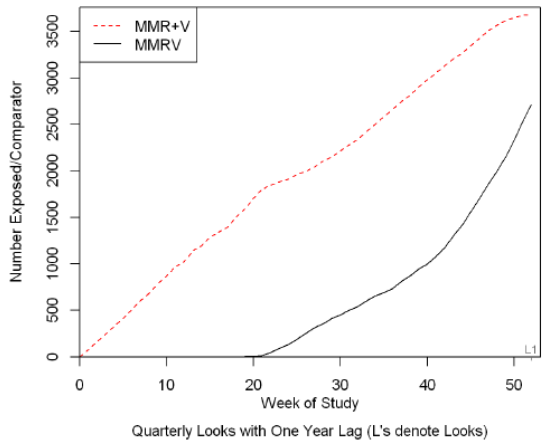


Figure A.2: Uptake of MMR+V and MMRV over Time for Site 15

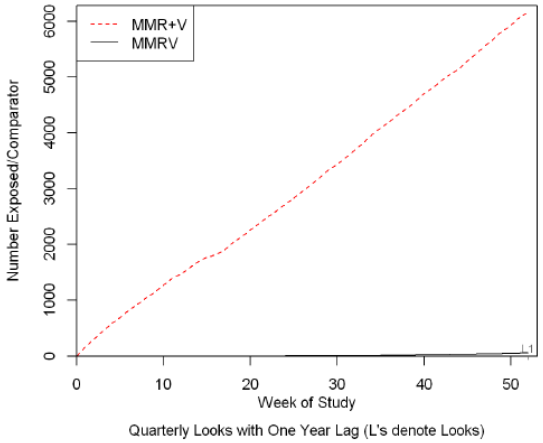


Figure A.3: Uptake of MMR+V and MMRV over Time for Site 16

Table 18 shows the site specific adjusted results at the current analysis. This table is important for assessing whether there is consistency in the estimated adjusted OR across sites and how the amount of information contributed by each site varies. For this example, the Analysis 1 results are mainly driven by site 15, as only 23 and 57 children at sites 4 and 16 were exposed to MMRV relative to 2,716 at site 15. Further, there are only one and two outcomes total at sites 4 and 16, respectively, compared to site 15 with 7 outcomes total.

Table 16. Current analysis site specific results (report Table A.5)

Site	MMR+V N	MMR+V Outcome(%)	MMRV N	MMRV Outcome(%)	MMR+V Adj %Out	MMRV Adj %Out	Adj OR(SE)*
2	0	0( )	0	0( )			( )
4	2806	1(0.036)	23	0(0.000)	0.036	0.000	0.00(Inf)
15	3686	2(0.054)	2716	5(0.184)	0.050	0.158	3.16(2.31)
16	6160	2(0.032)	57	0(0.000)	0.026	0.000	0.00(Inf)

*Site-specific adjusted logistic regression model (no Sequential).
Covariates Included: Age, Sex and indicator for each look.
Abbreviations: Outcome(%)=Number(Risk %) of outcome within look and covariate category, Adj=Adjusted, Adj %Out=Adjusted Risk % from site-specific adjusted logistic regression model assuming entire site population was either exposed or unexposed, and Adj OR(SE)= Adjusted Odds Ratio (Standard Error) comparing MMRV to MMR+V from site-specific logistic regression model.

2. Analysis 8 When a Signal Was First Detected

We will now review an analysis in which the sequential boundary is crossed and there is a signal indicating an elevated risk in the exposure of interest (MMRV). For the MMRV versus MMR+V example this occurred at Analysis 8 on day 1001 (2.75 years after the start of surveillance). We will first look at the Table 19 information showing demographics by exposure group. At Analysis 8, we have a cohort of 83,370 children in which 35,137 received MMR+V and 48,233 received MMRV. Similar distributions of demographic characteristics were observed in both exposure groups except for site 15 in which had 81.5% of MMRV recipients compared to only 20.8% of MMR+V recipients. We then visually display uptake over time by exposure group from GS GEE Analysis 8 in Figure 9. Figure 1 shows that uptake stayed very low for MMRV up to 50 weeks into the study and then dramatically increased over time surpassing MMR+V around study week 95 with both vaccines steadily increasing thereafter.

The main results for Analysis 8 are displayed in GS GEE Table 20. The last row shows the primary surveillance results for Analysis 8 which was conducted at day 1001. Among 35,137 new users of MMR+V, 13 (0.037%) had a febrile seizure. Among 48,233 new users of MMRV, 45 (0.093%) had a febrile seizure. After adjusting for age, sex, and site using logistic GEE the adjusted seizure rate among MMR+V was 0.038% compared to MMRV of 0.091% yielding an adjusted OR of 2.37. The GS GEE Score test statistic was 5.783, which exceeded the critical boundary of 3.832 and so there was a statistically significant signal. For an analysis in which we observe a signal, we provide a sequential adjusted analysis p-value at the bottom of the footer. In this example, the sequential p-value was 0.030 and we were able to detect an elevated risk of febrile seizure of 2.37 OR after 1001 days of surveillance (or after 48,233 exposed children to MMRV). We further note that this elevated adjusted OR was consistent across analyses.

Table 19. Demographics of population by exposure group for MMRV and MMR+V at look 8

	Total	MMR+V	MMRV
Total, N(Row%)	83370 (100.0)	35137 (42.1)	48233 (57.9)
Age, N(Col%)			
11m-12m	45105 (54.1)	21207 (60.4)	23898 (49.5)
13m-14m	16752 (20.1)	6101 (17.4)	10651 (22.1)
15m-16m	13024 (15.6)	4798 (13.7)	8226 (17.1)
17m-19m	5717 (6.9)	2030 (5.8)	3687 (7.6)
20m-23m	2772 (3.3)	1001 (2.8)	1771 (3.7)
Sex, N (Col%)			
Male	42600 (51.1)	18052 (51.4)	24548 (50.9)
Female	40770 (48.9)	17085 (48.6)	23685 (49.1)
Site, N (Col%)			
2	10992 (13.2)	8730 (24.8)	2262 (4.7)
4	8088 (9.7)	7686 (21.9)	402 (0.8)
15	46655 (56.0)	7321 (20.8)	39334 (81.5)
16	17635 (21.2)	11400 (32.4)	6235 (12.9)

Figure 10. GS GEE for look 8 showing uptake over time by exposure group for MMRV and MMR+V example

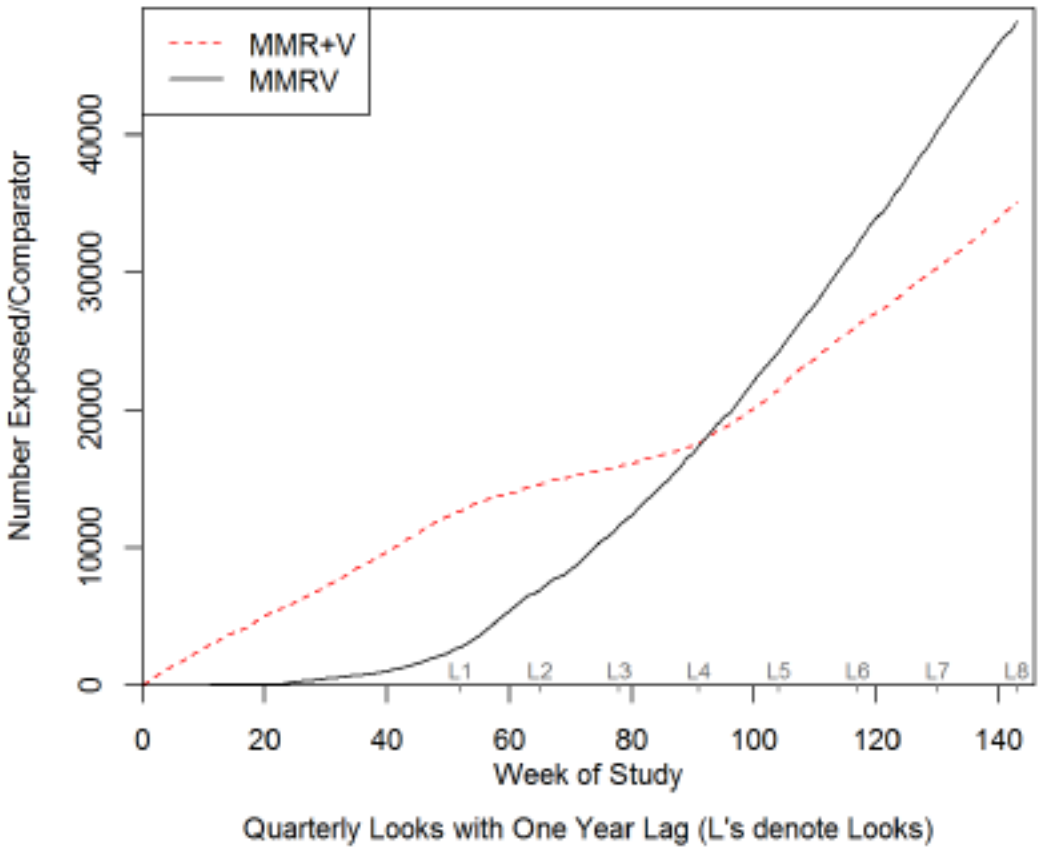


Figure 1: Total uptake of MMRV and MMR+V over Time at Look 8

Table 20. Primary results for MMRV and MMR+V example at look 8

Look:	Days	MMR+V N	MMR+V Outcome (%)	MMRV N	MMRV Outcome (%)	MMR+V Adj %Out	MMRV Adj %Out	Adj OR*	Score Test	Boundary	Error Spent	Signal
1:	364	12652	5(0.040)	2796	5(0.179)	0.043	0.131	3.05	1.842	3.407	0.000	No
2:	455	14633	7(0.048)	6970	10(0.143)	0.047	0.150	3.20	2.791	4.329	0.012	No
3:	546	15968	7(0.044)	11577	14(0.121)	0.044	0.119	2.69	2.787	4.493	0.012	No
4:	637	17560	7(0.040)	17376	18(0.104)	0.045	0.093	2.09	2.124	3.678	0.017	No
5:	728	21491	8(0.037)	24195	21(0.087)	0.042	0.079	1.86	2.038	3.873	0.019	No
6:	819	26169	10(0.038)	32123	29(0.090)	0.041	0.085	2.07	3.683	3.802	0.026	No
7:	910	30385	12(0.039)	40326	36(0.089)	0.045	0.081	1.79	2.513	3.601	0.029	No
8:	1001	35137	13(0.037)	48233	45(0.093)	0.038	0.091	2.37	5.783	3.832	0.038	Yes

***Adjusted logistic regression model applied using GEE framework with sequential monitoring boundaries based on permutations.**
Covariates Included: Age, Sex, and Site.
Abbreviations: Outcome(%)=Number (Risk %) of outcome within look and covariate category. Adj=Adjusted, Adj %Out= Adjusted Risk % from adjusted logistic regression model assuming entire population was either exposed or unexposed, Adj OR= Adjusted Odds Ratio comparing MMRV to MMR+V from logistic regression model, Score Test = Score Test statistic from GEE logistic regression model, and Boundary = Sequential Boundary to compare the Score Test Estimate.
Sequential P-Value at Signal: 0.03

a. Appendix Results for GS GEE at Signal (Look 8)

We will now look more into the data to assess if there is any obvious issues that might make us question our results. **Table 21** shows how the outcome incidence rate changed over analysis time by confounder strata. Over time there are relatively similar incidence rates, and nothing seems to be changing drastically by site which is a good indication that coding practices did not change substantially over the surveillance period.

Table 22 displays the site-specific adjusted OR analysis results. Here we see that the elevated risk is consistent across sites 2, 15, and 16 with adjusted OR ranging from 1.88 to 3.08. Site 4 has little exposure uptake yielding little influence on the results and is not able to calculate a site specific adjusted OR due to no outcomes in MMRV. Having a consistent elevated risk assures us that our overall results are likely indicative of a true elevated risk of febrile seizure due to MMRV.

Table 21. GS GEE outcome and incidence rates by look and covariate strata at look 8

	Look 1	Look 2	Look 3	Look 4	Look 5	Look 6	Look 7	Look 8
Total, Outcome(%)	10 (0.065)	17 (0.079)	21 (0.076)	25 (0.072)	29 (0.063)	39 (0.067)	48 (0.068)	58 (0.070)
Age, Outcome(%)								
11m-12m	3 (0.036)	6 (0.053)	7 (0.050)	9 (0.051)	11 (0.046)	16 (0.052)	18 (0.048)	23 (0.051)
13m-14m	3 (0.098)	4 (0.094)	5 (0.090)	5 (0.071)	6 (0.066)	9 (0.077)	11 (0.078)	13 (0.078)
15m-16m	2 (0.079)	2 (0.054)	2 (0.041)	3 (0.049)	3 (0.038)	4 (0.041)	6 (0.052)	8 (0.061)
17m-19m	1 (0.093)	3 (0.192)	4 (0.195)	4 (0.149)	5 (0.151)	6 (0.148)	9 (0.185)	9 (0.157)
20m-23m	1 (0.201)	2 (0.275)	3 (0.329)	4 (0.331)	4 (0.256)	4 (0.204)	4 (0.172)	5 (0.180)
Sex, Outcome(%)								
Male	4 (0.051)	8 (0.072)	10 (0.071)	12 (0.067)	14 (0.060)	17 (0.057)	20 (0.055)	25 (0.059)
Female	6 (0.080)	9 (0.085)	11 (0.082)	13 (0.076)	15 (0.067)	22 (0.077)	28 (0.081)	33 (0.081)
Site, Outcome(%)								
2	0 (Nan)	0 (NaN)	0 (NaN)	0 (0.000)	2 (0.069)	4 (0.072)	5 (0.060)	6 (0.055)
4	1 (0.035)	1 (0.028)	1 (0.023)	1 (0.020)	1 (0.017)	1 (0.015)	1 (0.014)	1 (0.012)
15	7 (0.109)	10 (0.100)	13 (0.095)	17 (0.093)	19 (0.079)	25 (0.080)	33 (0.085)	38 (0.081)
16	2 (0.032)	6 (0.075)	7 (0.073)	7 (0.062)	7 (0.054)	9 (0.061)	9 (0.056)	13 (0.074)

Table 22. GS GEE Look 8 analysis site specific results

Site	MMR+V N	MMR+V Outcome(%)	MMRV N	MMRV Outcome(%)	MMR+V Adj %Out	MMRV Adj %Out	Adj OR(SE)*
2	8730	4(0.046)	2262	2(0.088)	0.063	0.118	1.88(2.38)
4	7686	1(0.013)	402	0(0.000)	0.013	0.000	0.00(Inf)
15	7321	3(0.041)	39334	35(0.089)	0.041	0.088	2.15(1.83)
16	11400	5(0.044)	6235	8(0.128)	0.051	0.156	3.08(1.77)

Site-specific adjusted logistic regression model (no Sequential).
Covariates Included: Age, Sex.
Abbreviations: Outcome(%)=Number(Risk %) of outcome within look and covariate category, Adj=Adjusted, Adj %Out= Adjusted Risk % from site-specific adjusted logistical regression model assuming entire site population was either exposed or unexposed, and Adj OR(SE)= Adjusted Odds Ratio (Standard Error) comparing MMRV to MMR+V from site-specific logistic regression model.

IV. COMPARISON OF MMRV RESULTS ACROSS METHODS AND PREVIOUS PUBLISHED FINDINGS

As has been shown in this report both statistical methods, GS IPTW and GS GEE, found evidence for elevated risk of febrile seizure among children vaccinated with the combination vaccine MMRV relative to two separate injections of MMR and V (MMR+V) vaccines. Using the same analysis look plan of waiting to look until one year after licensure (15,558 children vaccinated with 2,796 vaccinated with MMRV) and then quarterly looks there after (between 6,000-12,000 vaccinated children with 4,000-8000 vaccinated with MMRV) the GS IPTW method signaled at the 4th analysis at 1.75 years (17,376 MMRV doses) since licensure compared to the GS GEE method at the 8th analysis at 2.75 years (48,233 MMRV doses) since licensure. This resulted in 30,857 additional doses of MMRV within our study population before the GS GEE method would signal compared to the GS IPTW method. The weighted RD estimated at the time of signal from the GS IPTW was 5.2 per 10,000 doses with a relative risk of 2.86 (0.080/0.028). The adjusted OR estimated at the time of signal from the GS GEE model was 2.37 resulting in an estimated RD of 5.3 per 10,000 doses (9.1-3.8 from the adjusted risk estimates using logistic regression). Therefore, both approaches estimated similar magnitude of risk at their corresponding time of signal, but the GS IPTW was able to signal 1 year earlier (30,857 MMRV doses) than the GS GEE method.

This example shows the capability of the methods in the FDA Sentinel data system (using a subset of sites), but it is also important to compare to findings previously published examining this vaccine comparison in the CDC Vaccine Safety Datalink (VSD). The VSD has been actively monitoring new vaccines using sequential monitoring methods. In February 2006 they launched an active surveillance study to monitor evidence for elevated rates of Febrile Seizure amongst new MMRV recipients relative to a historical control design.[8,9] Specifically, they used the population of children that received MMR vaccine with or without varicella from 2000 to 2006 and calculated age, sex, and site adjusted expected risk estimates of febrile seizures for the prospective cohort of newly vaccinated children to MMRV. The study used a continuous sequential monitoring boundary which assumes that one monitors after every MMRV vaccination, but in reality the data available was weekly. They used the Poisson MaxSPRT[10]

sequential monitoring method which used a Poisson likelihood ratio test statistic assuming the expected rates of outcomes from the historical controls is known (not estimated) and monitors until a fixed number of exposed MMRV recipients (paper did not report the maximum sample size planned). The VSD study detected an elevated rate of febrile seizures after 43,353 MMRV doses were administered and estimated an adjusted RR of ~2 (exact RR not reported). They conducted a follow-up study using a prospective control (MMR+V recipients) and chart reviewing all outcomes and found an estimated adjusted OR of 2.3 and estimated RD of 5 per 10,000 vaccinated.

The results from the VSD study are extremely comparable to our findings using FDA Sentinel data in terms of magnitude of risk at the time of signal (OR or RR, RD: VSD 2.3 , 5 per 10,000; GS IPTW 2.86, 5.2 per 10,000; GS GEE 2.37, 5.3 per 10,000). The GS IPTW method found the elevated risk the quickest after 17,376 MMRV doses while the VSD Poisson MaxSPRT found an elevated risk after 43,353 MMRV doses and the GS GEE method after 48,233 MMRV doses. These results corroborate the findings of an elevated risk of febrile seizures for MMRV recipients compared to MMR+V recipients. Further study comparing these three statistical methods and whether the time to detection is shorter for GS IPTW in general should be evaluated.

V. REFERENCES

1. Mini-Sentinel Methods Development: Sequential Testing Working Group Report. 2011. (Accessed at http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Sequential-Testing-Report.pdf.)
2. Mini-Sentinel Methods Development: Statistical methods for estimating causal risk differences in the distributed data setting for postmarket safety outcomes. 2012. (Accessed at http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_PRISM_Statistical-Methods-for-Estimating-Causal-Risk-Differences.pdf.)
3. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 2004;23:2937-60.
4. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association 1984;79:516-24.
5. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics 1999;55:874-82.
6. Pocock SJ. Interim Analyses for randomized clinical-trials - The Group Sequential Approach. Biometrics 1982;38:153-62.
7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.
8. Centers for Disease Control and Prevention; Advisory Committee on Immunization Practices Update: recommendations from the Advisory Committee on Immunization Practices (ACIP) regarding administration of combination MMRV vaccine. MMWR Morb Mortal Wkly Rep 2008;57:258-60.
9. Klein NP, Fireman B, Yih WK, et al. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. Pediatrics 2010;126:e1-8.
10. Kulldorff M, Davis RL, Kolczakâr M, Lewis E, Lieu T, Platt R. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. Sequential Analysis: Design Methods and Applications 2011;30:58 - 78.
11. Cook AJ, Tiwari RC, Wellman RD, et al. Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the Mini-Sentinel pilot. Pharmacoepidemiology and Drug Safety 2012;21:72-81.
12. Nelson JC, Cook AJ, Yu O, et al. Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. Pharmacoepidemiology and Drug Safety 2012;21:62-71.
13. Cook AJ, Wellman RJ, Tiwari RC, Nelson JC. Group Sequential Methods for observational data incorporating confounding through estimating equations with application in Post-Marketing Vaccine/Drug Surveillance. Submitted 2013.
14. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf 2010;19:858-68.
15. Rotnitzky A, Jewell NP. Hypothesis-testing of regression parameters in semiparametric generalized linear-models for cluster correlated data. Biometrika 1990;77:485-97.
16. Zeger SL, Liang KY, Albert PS. Models for longitudinal data - a generalized estimating equation approach. Biometrics 1988;44:1049-60.

VI. STATISTICAL APPENDIX

A. GS IPTW DETAILS

Details of the GS IPTW method have been published in detail in a previous Mini-Sentinel methods report[2], but this section will briefly discuss the statistical assumptions underlying the method and provide detail on what specifically is being displayed in the GS IPTW report (Table 2). We will first introduce the methods assuming that the analysis is being performed at a single site (pooled (non-distributed) data setting) and without sequential monitoring of the outcome (Last Table of Appendix Site-Specific Results). Then in Section VI.A.2, we will introduce the stratified IPTW that properly incorporates the distributed data structure. Finally in Section VI.A.3 we will propose how to extend both standard and stratified IPTW methods for group sequential monitoring.

1. IPTW Risk Difference Estimation: Single Site Setting

Assume at a single site, *s*, we have outcome Y_{si} (i=1,..,N_s) equal to 1 if subject *i* at site *s* experiences the outcome of interest and 0 otherwise, with exposure X_{si}(t) equal to 1 if subject *i* at site *s* is exposed to the medical product of interest and equal to 0 otherwise, and let $\boldsymbol{Z}_{si}$ be a set of measured baseline confounders. Then define the propensity score, e_{si}, as the probability of receiving (ie. being exposed to) the medical product X_{si} given confounders $\boldsymbol{Z}_{si}$, so that $e_{si}=P(X_{si}|\boldsymbol{Z}_{si})$. We estimate e_{si} using a standard logistic regression model, $logit(E(Xsi|\boldsymbol{Z}si))= \beta_{x,z}\boldsymbol{Z}_{si}$, where $\beta_{x,z}$ is estimated using the maximum likelihood approach. This is typically done in practice and yields $\hat{e} = (1+\exp(-\hat{\beta}_{x,z}\mathbf{Z}_{si}))^{-1}$. These propensity scores will be used as the inverse probability weights to upweight individuals who were estimated to be unlikely to receive the treatment, but actually did receive the treatment, while downweighting individuals who were estimated to be likely to receive the treatment and did receive the treatment. Similarly among those that did not actually receive the treatment, the inverse probability weights will upweight those estimated to be likely to receive the treatment and downweight those estimated to be not likely to receive the treatment. This evens out the baseline covariate distribution, across exposed and unexposed populations, to allow one to estimate an unconfounded population average effect estimate.

There are numerous approaches available to estimate the risk difference using IPTW and propensity scores [3]. For this report we have chosen one weighting approach for which details are given below. We initially included a doubly robust estimate, but this was found to be infeasible for the rare event setting (even when the probability of outcome was as high as 5%) since doubly robust estimates require modeling the probability of outcome conditional on confounders within the exposed group and separately modeling the same quantity within the unexposed group. Specifically, because of the small number of events, at least one of the models often failed to be estimable. Therefore we used a standard approach originally proposed by Rosenbaum et al[4] which takes the following form,

$$\hat{\Delta}_s = \left(\sum_{i=1}^{N_s}\frac{X_{si}}{\hat{e}_{si}}\right)^{-1}\sum_{i=1}^{N_s}\frac{X_{si}Y_{si}}{\hat{e}_{se}} - \left(\sum_{i=1}^{N_s}\frac{1-X_{si}}{1-\hat{e}_{si}}\right)^{-1}\sum_{i=1}^{N_s}\frac{(1-X_{si})Y_{si}}{1-\hat{e}_{se}} = \hat{\mu}_{s1} - \hat{\mu}_{s0}\,.$$

The estimated variance of $\hat{\Delta}_s$ is derived using the empirical sandwich method[3] taking into account that the e_{si} are estimated. The formula for the variance is given by,

$$\hat{\mathrm{V}}\left(\hat{\Delta}_s\right)=\frac{1}{N_s^2}\sum_{i=1}^{N_s}\left[\frac{X_{si}(Y_{si}-\hat{\mu}_{s1})}{\hat{e}_{si}}-\frac{(1-X_{si})(Y_{si}-\hat{\mu}_{s0})}{1-\hat{e}_{si}}-(X_{si}-\hat{e}_{si})\hat{\mathbf{H}}^T\hat{\mathbf{E}}^{-1}\mathbf{Z}_{si}\right]^2,$$

where

$$\hat{\mathbf{E}}^{-1}=N_s^{-1}\sum_{i=1}^{N_s}\hat{e}_{si}(1-\hat{e}_{si})\mathbf{Z}_{si}\mathbf{Z}_{si}^T$$

and

$$\hat{\mathbf{H}}=\frac{1}{N_s}\sum_{i=1}^{N_s}\left[\frac{X_{si}(Y_{si}-\hat{\mu}_{s1})(1-\hat{e}_{si})}{\hat{e}_{si}}+\frac{(1-X_{si})(Y_{si}-\hat{\mu}_{s0})\hat{e}_{si}}{1-\hat{e}_{si}}\right]\mathbf{Z}_{si}.$$

If the e_{si} are not estimated, such as in the case of known sample weights, then the variance of $\hat{\Delta}_s$ can be estimated as,

$$\tilde{\mathrm{V}}(\hat{\Delta}_s)=\frac{1}{N_s^2}\sum_{i=1}^{N_s}\left[\frac{X_{si}(Y_{si}-\hat{\mu}_{s1})}{e_{si}}-\frac{(1-X_{si})(Y_{si}-\hat{\mu}_{s0})}{1-e_{si}}\right]^2,$$

which is larger than $\hat{\mathrm{V}}\left(\hat{\Delta}_s\right)$. This variance, $\tilde{\mathrm{V}}(\hat{\Delta}_s)$, will be used later when describing a permutation approach for the distribution of the standardized test statistic under the null.

It should be noted that bootstrapping is the most standard approach for obtaining IPTW variance estimators. However, we chose this empirical estimator approach because it is simpler and computationally faster to use making it more practical to implement, especially in the context of a distributed data setting.

The single site approach detailed here is used for the site-specific risk difference estimates presented in the last appendix table showing results by individual site. The next section will extend this result to a stratified IPTW method which allows for an overall, population-level estimate which combines each of the site specific estimates.

2. Stratified IPTW Risk Difference: Multi-Site Estimate

A variety of approaches exist for combining data across sites. The most straightforward approach, which is used in this report, is to use stratified modeling and treat each site's estimate as independent. Specifically, for the risk difference with site-specific estimate, $\hat{\Delta}_s$, a valid overall population estimate, $\hat{\Delta}$, is

$$\hat{\Delta}=\frac{\sum_{s=1}^{S}w_s\hat{\Delta}_s}{\sum_{s=1}^{S}w_s},$$

with estimated variance

$$\hat{\mathrm{V}}(\hat{\Delta})=\frac{\sum_{s=1}^{S}w_s^2\hat{\mathrm{V}}(\hat{\Delta}_s)}{\left[\sum_{s=1}^{S}w_s\right]^2},$$

where w_s can be the sample size of the site, N_s, or the inverse of the variance of the estimator from that site, $\hat{V}(\hat{\Delta}_s)$. However, due to potential instability of the site-specific variance estimates in the rare event setting, we found that weighting with the sample size performed much better, and we therefore use this approach in this report.

Another quantity of interest that is obtainable from this approach is an adjusted probability of outcome per exposure group (adjusted risk). Specifically, the site specific RD= $\hat{\Delta}_s = \hat{\mu}_{s1} - \hat{\mu}_{s0}$, which is the average estimated probability of observing the outcome given X=1 minus the average estimated probability of observing the outcome given X=0. Therefore we can then calculate a site stratified estimated adjusted risk as the following,

$$\hat{\mu}_X = \frac{\sum_{s=1}^{S} w_s \hat{\mu}_{sX}}{\sum_{s=1}^{S} w_s}.$$

We use this quantity in the report when presenting adjusted risk by exposure group. The next section will extend these approaches for use in group sequential monitoring.

3. Group Sequential IPTW (GS IPTW)

Now that we are in the context of group sequential monitoring we must introduce the concept of multiple analysis times. Specifically, we assume that accruing data will be analyzed at specific time points (t=1,…,T). We also assume that an individual i at site s is either exposed to the medical product of interest, X_{si}(t)=1, or not, X_{si}(t)=0 and either has the outcome of interest, $Y_{si}(t)$=1, or does not ,$Y_{si}(t)$=0, before analysis time t. Note that since we are assessing acute outcome events with short follow-up windows it is standard to only include participants in the study population after their short follow-up window (e.g. 45 days) has elapsed so that all participants have the same follow-up time. Additional common data lag time issues have been discussed elsewhere and will be not discussed in this report.[11,12] We assume all within site baseline confounders, $\mathbf{Z}_{si}$, are measured and are not dependent on time. Further, we assume that the cumulative number of participants observed at site s up to analysis time t is $N_s(t)$ yielding the cumulative total number of observed people across sites at analysis time t is $N(t)$= $\sum_{s=1}^{S} N_s(t)$.

The same null hypothesis is tested at each analysis time t, H_O: $\Delta(t)$=0, and if the test statistic at analysis t exceeds a pre-defined critical boundary, $c(t)$, it signals a significantly elevated rate of events in the exposed group at analysis t; otherwise, the study continues to the next analysis time until the pre-defined end of the evaluation, $N(T)$. At each analysis, new information accumulates, which includes new participants since the last analysis. Different approaches for incorporating updated data yield different assumptions that need to be accounted for in the calculation of the critical boundary. The critical boundary can be chosen in numerous ways, but it must maintain the overall type I error rate across all analyses, taking into account both multiple testing and the skewed distribution of the test statistic that results when one conditions on whether or not earlier test statistics exceeded the specified critical value. To form a boundary it is necessary to define a test statistic, the variability of the test statistic over time, the shape of the boundary, the number of analysis times and when they will occur, the α-level

(type I error) and either the maximum sample size at end of study or overall power. To begin, we first assume that the maximum sample size and number of observations per analysis time are known and allow power to vary. We used a general unifying boundary definition developed by Kittleson and Emerson.[5] This approach defines the boundary as a general function of time $c(t)=au(t)$ where $u(t)$ is a function dependent on the proportion of statistical information (e.g., sample size) up to time t and is of the form $u(t)=(N(T)/N(t))^{1-2\omega}$, where $\omega>0$ is a fixed parameter depending upon the desired design (e.g. $u(t)=1$ is Pocock-like[6] and $u(t)=(N(T)/N(t))^{0.5}$ is O'Brien and Fleming[7]). One solves for the constant a using an iterative simulation approach to hold the overall type I error at α.

For our application in the rare event setting we have chosen a non-parametric permutation approach to solve for *a,* which has the advantage of relaxing standard parametric assumptions. If the null hypothesis, H_O: $\Delta(t)=0$, is true, this implies that within each site X_{si} is independent of Y_{si} conditional on confounders $\mathbf{Z}_{si}$. Therefore, to derive a permutation test under the null one can simply permute all X's while fixing the outcome and confounder data as observed, $((Y_{s,1}, \mathbf{Z}_{s,1}),\ldots, (Y_{s,Ns(t)}, \mathbf{Z}_{s,Ns(t)}))$, resulting in a permutation, $(X^p_{s,(1)},\ldots, X^p_{s,(Ns(t))})$, where p indicates the p^{th} of the Nperm permutations. However, since we are randomly permuting X's, the propensity score for the permuted dataset is constant $P(X^p_{si}|\mathbf{Z}_{si})= P(X^p_{si})$, since X^p_s is independent of $\mathbf{Z}$, for all p. Once we fix the propensity scores to be constant, we must incorporate this into the estimate of the variance of the estimator for the permuted data. Specifically, for the permuted data the estimate is not $\hat{V}(\hat{\Delta}_s)$, but $\tilde{V}(\hat{\Delta}_s)$. Keeping propensity scores constant allows for computational efficiency since the propensity score does not need to be estimated for all permutations. Further, in reality we observe variability in the proportion exposed over time, which directly affects the variability of the test statistic and therefore we have implemented the permutation test by permuting within a given analysis time. To do this, we assume that the data are ordered by time of entry into study such that, for analysis at time *t,* the new data observed at analysis time t since t-1 for site s is indexed by $\{N_s(t-1)+ 1\}$ to $\{N_s(t)\}$, and for the first analysis time has index {1} to $\{N_s(t)\}$. Given this ordering of the data the permutation approach proceeds as follows:

Step 1: Within each analysis t for each site s, simulate data by fixing observed outcomes $(Y_{s,Ns(t-1)+1},\ldots,Y_{s,Ns(t)})$ and permuting $X_{s,Ns(t-1)+1},\ldots,X_{s,Ns(t)}$ to create $X^p_{s,Ns(t-1)+1},\ldots,X^p_{s,Ns(t)}$ to obtain Nperm permuted datasets (p=1,…,Nperm).

Step 2: For each permuted dataset p at each analysis time t and each site s calculate the site specific adjusted RD, $\hat{\Delta}_s(t)$, and variance, $\tilde{V}(\hat{\Delta}_s(t))$ from permuted data up to time t ($Y_{s,1}(t)$, $X^p_{s,1}$),…,($Y_{Ns(t)}(t)$, $X^p_{Ns(t)}$) and fixing the propensity scores $e_{si} = \sum_{i=N_s(t-1)+1}^{N_s(t)} X_{si}/(N_s(t)-N_s(t-1))$.

Step 3: For each permutation dataset p at each analysis time t calculate the stratified IPTW estimate, $\hat{\Delta}(t)$, and variance, $\tilde{V}(\hat{\Delta}(t))$ yielding the standardized IPTW test statistic, $Z^p(t)$.

Step 4: For each permuted dataset calculate $C^p_{\max} = \sup_t \frac{Z^p(t)}{u(t)}$ which is the maximum value of the standardized test statistic across time for that permutation taking into account the desired shape of the boundary.

Step 5: Estimate, a, as $\hat{a} = C^{(1-\alpha)}_{\max}$ which is the (1-α) percentile of $C^p_{\max}$.

Step 6: Boundary at time t is $c(t)=\hat{a}\,u(t)$.

This simulation framework requires that we have a complete dataset ($Y_{xi}(t)$, $\boldsymbol{Z}_{si}$, $X_{si}(t)$) for all sites and all observation times. However, this is not practical at earlier analysis times $t < T$. To solve this, at times $t < T$ we can instead make assumptions about how the data will look at future analysis times. Specifically, to derive the permutation approach under the null we only need to know the prevalence of X_{si} and Y_{si} at future looks since $P(X_s|\boldsymbol{Z}_s)=P(X_s)$ under the null and the total sample size is $N(T)$. Therefore, to approximate the future prevalence of X_s and Y_s, we can sample the future observations, $N(T)$-N(t), by sampling with replacement from the observed (X_{si}, Y_{si}). At each site we sample $(N(T)- N(t))xN_s(t)/N(t)$ future observations within site s assuming that the proportion of future observations that site s contributes is similar to the proportion it contributes currently (e.g. $N_s(t) / N(t)= N_s(T) / N(T)$). This will create a complete dataset necessary to perform the permutation approach previously described for all analyses.

In practice, at each new analysis time we keep the prior boundary values, $c(1),\ldots,c(t\text{-}1)$ since these were the signaling thresholds used at previous analysis times and each analysis time is defined to be conditional on the prior analyses. The simulation plan is slightly altered to take into account the amount of error spent at previous analysis times and incorporating not having all observed data across all analysis times. Specifically we follow the following simulation outline:

Step 1: If at analysis time $t<T$: Create one complete dataset by sampling with replacement $(N(T)\text{-}N(t))xN_s(t)/N(t)$ observations from $(Y_{si}(t),X_{si}(t))$ (i = 1,…, $N_s(t)$) at each site s.

Given a complete observed dataset:

Step 2: Within each analysis t for each site s, simulate data by fixing observed outcomes $(Y_{s,Ns(t-1)+1},\ldots,Y_{s,Ns(t)})$ and permuting $X_{s,Ns(t-1)+1},\ldots,X_{s,Ns(t)}$ to create $X^p_{s,Ns(t-1)+1},\ldots,X^p_{s,Ns(t)}$ to obtain Nperm permuted datasets (p=1,…,Nperm).

Step 3: For each permuted dataset p at each analysis time t at each site s calculate the site specific adjusted RD, $\hat{\Delta}_s(t)$, and variance, $\tilde{\mathrm{V}}(\hat{\Delta}_s(t))$ from permuted data up to time t ($Y_{s,1}(t)$, $X^p_{s,1}$),…,($Y_{Ns(t)}(t)$, $X^p_{Ns(t)}$) and fixing the propensity scores $e_{si} = \sum_{i=N_s(t-1)+1}^{N_s(t)} X_{si}/(N_s(t)-N_s(t-1))$.

Step 4: For each permutation dataset p at each analysis time t calculate the stratified IPTW estimate, $\hat{\Delta}(t)$, and variance, $\tilde{\mathrm{V}}(\hat{\Delta}(t))$ yielding the standardized IPTW test statistic, $Z^p(t)$.

Step 5: For analysis times ($j<t$) already observed and have previous boundaries c(1),…,c(j):

Calculate the cumulative error spent at analysis time j as:

$$\hat{\alpha}(j) = \frac{\sum_{p=1}^{Nperm} I\left(Z^p(1) \geq c(1) \bigcup \cdots \bigcup Z^p(j) \geq c(j)\right)}{Nperm}$$

and for permutation datasets which cross the previous boundaries set the current analysis time standardized IPTW test statistic, $Z^p(t)$, to be something large such as 10,000. Do this to make sure that permutation will be treated as signaling in the next step.

Step 6: For each permuted dataset calculate $C^p_{\max} = \sup_t \frac{Z^p(t)}{u(t)}$ which is the maximum value of the standardized IPTW test statistic across time for that permutation taking into account the desired shape of the boundary.

Step 7: Estimate the current analysis time, *a*, as $\hat{a}_t = C_{\max}^{(1-\alpha)}$ which is the (1-α) percentile of $C_{\max}^{p}$.

Step 8: Boundary at time *t* is $c(t)=\hat{a}_t\, u(t)$, which takes into account previous boundaries and error spent.

This is how the boundaries are calculated for the current Sentinel application.

a. Sequential p-values

Given the sequential monitoring boundaries it is important to further quantify the level of statistical significance either at the time of signal or at the end of study surveillance. To calculate such a p-value one must make certain decisions about how to order a series of test statistics over time. In a one time analysis it is straightforward to order a given test statistic, *S,* by whether the data realization k is greater than data realization m if the test statistic $S^k > S^m$. However, there are numerous approaches to choose the ordering of data in sequential monitoring. In our context we only need to order the permuted data realizations compared to observed data realizations at the time of signal (observed *Z*(t)>c(t)) or end of analysis time T. We defined a permuted data realization to be more extreme than the observed data realization if one of the following conditions occurred:

Permuted data realization signaled at previous analysis times: $Z^p(1) \geq c(1) \bigcup \cdots \bigcup Z^p(t) \geq c(t)$ **or**

Permuted data realization did not signal at previous analysis times, but the permuted current analysis time realization is greater than the observed standardized test statistic,

$$Z^p(t) \geq Z(t) \,|\, \left\{ Z^p(1) < c(1) \bigcap \cdots \bigcap Z^p(t-1) < c(t-1) \right\}$$

Then the empirical sequential p-value is

$$P = \frac{\sum_{p=1}^{Nperm} \mathrm{I}\left(Z^p(1) \geq c(1) \bigcup \cdots \bigcup Z^p(t-1) \geq c(t-1) \right) + \mathrm{I}\left(Z^p(t) \geq Z(t) \,|\, Z^p(1) < c(1) \bigcap \cdots \bigcap Z^p(t-1) < c(t-1) \right)}{Nperm}$$

$$= \hat{\alpha}(t-1) + \frac{\sum_{p=1}^{Nperm} \mathrm{I}\left(Z^p(t) \geq Z(t) \,|\, Z^p(1) < c(1) \bigcap \cdots \bigcap Z^p(t-1) < c(t-1) \right)}{Nperm}$$

which is the cumulative error spent up to analysis time *t* combined with the probability of the non-signaling permuted datasets observing a more extreme value than the current observed test statistic. Note that the empirical sequential p-value at the time of signal will always be less than the cumulative error spent up to that analysis time, $\hat{\alpha}(t)$.

b. Details of GS IPTW Report Quantities

We will now go over each column of the results table, Table 2 (next page), of the GS IPTW report for at a signal for analysis 4 comparing MMR+V(X=0) to MMRV(X=1). The first two columns specify the look number and date of look. The 3rd column specifies cumulative number of those exposed to MMR+V at each look. The 4th column provides the cumulative number of outcomes and percent of the MMR+V with outcome (Number of Outcomes/Number of Exposures *100). The 5th and 6th column reports columns 3 and 4 for MMRV. The 7th (MMR+V Adj %Out) and 8th (MMRV Adj %Out) columns provide the site stratified adjusted risk estimates within exposure group using

$\hat{\mu}_0$ and $\hat{\mu}_1$ as described in Section VI.A.2. The 9th column (Adj RD) is $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$. The 10th column (IPTW Test) is the standardized IPTW test statistic, Z(t), as specified in Section VI.A.2. The 11th column (Boundary) is the sequential boundary derived using the methods detailed in this appendix, c(t). The 12th column (Error Spent) is the cumulative error spent up to a given analysis time $\hat{\alpha}(t)$. Note that additional error may not necessarily be spent at each analysis time since the test is based on a standardized test statistic boundary and not an error spending function. This may be advantageous since at earlier analysis times to follow an error spending rule may make the boundary on the test statistic smaller than desired. For example, we are attempting to have a flat boundary on the standardized IPTW test statistic scale so even though no error under the NULL was spent at analysis one we still could have signaled under the alternative hypothesis. If we had forced an error spending function the first boundary value would have been even smaller. However, later in the analysis with more data it would have made it more difficult to signal which is counterintuitive to gaining more statistical information. Even when attempting to have a flat boundary we still had fluctuations due to the nature of not being able to know the outcome, exposure, and confounder distributions at future looks. If they had stayed similar to look 1 then the boundary would have stayed close to the original 1.297, but since they did change drastically we properly took into account the actual observed distributions and appropriately updated the boundary to hold the overall type I error of 0.05. The final column (Signal) is just a Yes/No indicator if the IPTW Test Statistic crossed the boundary at a given look.

Since this analysis signaled we also report at the bottom of the table the Sequential P-Value at Signal of 0.0319. Note that it is smaller than the error spent at look 4 of 0.031. This is the sequential p-value as outlined in the previous section.

B. GS GEE METHOD DETAILS

Details of the GS GEE method have been published in detail in a previous methods task order report[1,13], but this section will briefly discuss the statistical assumptions underlying the method and provide detail on what specifically is being displayed in the GS GEE report (Table 2). We will start with data notation, detail the method and how the boundary is calculated, and finish with summarizing specifically each column of Table 2.

1. Data Specification and Notation

We assume that accruing data on new users of medical product of interest (MPI) or comparator is analyzed at specified times (t=1,…,T). We also assume that each individual i (i = 1,…,$N(t)$) is either exposed to the MPI, X_i=1, or not exposed, X_i=0, and either has the outcome of interest occurring before the end of analysis t, $Y_i(t)$=1, or does not $Y_i(t)$=0. The exposure time, $E_i(t)$, denotes the cumulative exposure time prior to analysis t. It could be a single time exposure window (e.g., vaccine: $E_i(t)$=1 for all individuals) or a chronic exposure (time on either MPI or comparator), for which assumptions of the exposure time and outcome relationship must be made (constant risk or change in risk due to exposure duration). What is currently applied in Sentinel is that we censor participant's exposure time at the date of disenrollment, occurrence of outcome, or discontinuation of use of the initial assigned treatment (allowing for a lag in follow-up specified by user). Further, participants are censored if they switch exposure groups and begin taking the other medical product (i.e. an exposed individual starts taking the

comparator medical product). These assumptions are consistent with incident user cohort studies which are currently being used in post-marketing surveillance[14].

Further, we assume that there is a set of baseline confounders, $\mathbf{Z}_i$, associated with individual *i*, which can be comprised of variables such as site, age, sex, and health conditions. When using aggregate data these confounders are often categorized to form a set of categorical confounders, Z_i^c. For example a continuous confounder, such as age, can be categorized into 5 or 10 year age groups. When one further would like to adjust for time we have allowed **Z** to also include indicators for look time and therefore across analyses times **Z** will differ in number of confounders. (Note that compared to GS IPTW **Z** includes site as one of the baseline confounders instead of conducting a site-stratified estimate)

2. Group Sequential Generalized Estimating Equations (GS GEE)

In this section describe the GS GEE method. We first define the relevant generalized estimating equations and score statistic used by our approach. We then introduce how we have implemented the sequential boundary calculations and corresponding sequential p-value. Finally we end with a detailed description of all values reported in the main GS GEE summary report Table 2.

a. Generalized Estimating Equations

Assume the marginal expectation of $Y_i(t)$ given covariates, $\boldsymbol{Z}_i$, exposure of interest, X_i, and exposure time $E_i(t)$, is $E(Y_i(t)|X_i, \boldsymbol{Z}_i, E_i(t))=\mu_i(t)$, where $\mu_i(t)$ is linked to X_i, $\boldsymbol{Z}_i$, and $E_i(t)$ through a link function g(.), such that

$$g(\mu_i(t)) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_Z \mathbf{Z}_i + f_\theta(E_i(t)),$$

where β_X is the effect parameter for exposed versus comparator, and $\boldsymbol{\beta}_z$ is a px1 vector of unknown regression parameters. Typically g(.) is logit for a binary outcome or logarithm for a Poisson outcome. The corresponding marginal variance, dependent on $\mu_i(t)$, is $Var(Y_i(t)|X_i,\mathbf{Z}_i,E_i(t))$, which is typically $\mu_i(t)(1-\mu_i(t))$ for binary outcomes and $\mu_i(t)$ for Poisson outcomes. The exposure link function, $f_\theta(.)$, would typically be ignored for a single time exposure or specified as the logarithmic function if using a Poisson model. However, to allow for flexibility this has been kept general.

Define the following score equations for analysis time *t* for β_X and $\boldsymbol{\beta}_z$ based on the first moment of $Y_i(t)$ as,

$$\mathrm{U}_{\beta(t)}(\boldsymbol{\beta}) = \begin{pmatrix} \mathrm{U}_{\beta_X(t)}(\beta_X) \\ \mathrm{U}_{\beta_Z(t)}(\boldsymbol{\beta}_Z) \end{pmatrix} = \begin{pmatrix} \sum_{i-1}^{n_t} \mathrm{U}_{i,\beta_X(t)}(\beta_X) \\ \sum_{i-1}^{n_t} \mathrm{U}_{i,\beta_Z(t)}(\boldsymbol{\beta}_Z) \end{pmatrix} = \begin{pmatrix} \sum_{i-1}^{n_t} \partial \mathrm{g}(\mu_i(t))/\partial\beta_X \\ \sum_{i-1}^{n_t} \partial \mathrm{g}(\mu_i(t))/\partial\boldsymbol{\beta}_Z \end{pmatrix}.$$

Using $\mathrm{U}_{\beta(t)}(\boldsymbol{\beta})$ we denote $S(t)$ as the standard generalized score statistic using a robust sandwich variance estimator[15] for testing if Ho: $\beta_X=0$. By using a generalized score statistic we are only assuming that the mean model is correctly specified.[16] In the next section we will use this score test statistic $S(t)$ at analysis time *t*, to develop sequential monitoring boundaries.

b. Observational Group Sequential Monitoring Boundary

The general purpose of group sequential boundaries is to be able to conduct multiple testing of a null hypothesis of interest while incorporating a stopping boundary. Both sufficient power and short time-to-detection of a potential signal, if it exists, is desired while still holding the overall type I error. For the method currently applied in Sentinel, the interest is in a one-sided alternative hypothesis of $\beta_X > 0$; however, it is easy to incorporate two-sided hypotheses or futility boundaries commonly used in clinical trials. For our boundary formulation we modified a well established simulation approach using the unifying family of boundaries[5]. This approach allows for the application of a wide range of commonly used boundary shapes including the Pocock-like boundary, which is at on the standardized test statistic scale[6] and the O'Brien and Fleming boundary, which is proportional to $\sqrt{n}$ on the standardized test statistic scale[7]. Specifically, the boundary is defined as $c(t) = au(t)$ where $u(t)$ is a function dependent on the proportion of statistical information (e.g. sample size) up to time t and is of the form $u(t)= \left(N(t)/N(T)\right)^{1-2\omega}$ where $\omega > 0$ is a fixed parameter depending upon design (specifically for Pocock-like $u(t) = 1$ and O'Brien and Fleming $u(t)= \sqrt{N(t)/N(T)}$) and a is solved iteratively by permuting the data under H_0 to hold the type 1 error at α.

To form a boundary it is necessary to define a test statistic, the variability of the test statistic over time, the shape of the boundary, the number of analysis times and when they occur, the α-level (type I error) and either maximum sample size at end of study or overall power. We first assume that the maximum sample size and number of observations per analysis time is known and allow power to vary. For observational studies to determine the variability of the test statistic over time one must also assume the distribution at each analysis time of all variables in the model including outcome, exposure and all confounder distributions. We then discuss how to alter this boundary selection process to incorporate earlier non-pre-specified analysis times, variable number of observations *N(t)* per analysis time and unknown future data distributions (exposed, outcome, and confounder distributions).

To accommodate rare events, we have chosen a non-parametric permutation approach to solve for *a,* which has the advantage of relaxing standard parametric assumptions. Under the null hypothesis $\beta_X(t) = 0$ for all t since $Y_i(t)|\boldsymbol{Z}_i,E_i(t)$ is independent of X_i. Therefore, we can permute observed exposures, X, while fixing the observed set $(Y(t),\boldsymbol{Z},E(t))$. Since we are analyzing data at times $t = 1,\ldots,T$, and in practice the variability in the proportion exposed may directly affect the variability of the test statistic, it is important to permute X within analysis time t. To do this, we assume that the data are ordered by time of entry into study such that, for analysis at time *t,* the new data observed at analysis time t since t-1 is indexed by $\{N(t\text{-}1)+ 1\}$ to $\{N(t)\}$ and for the first analysis time has index $\{1\}$ to $\{N(1)\}$. Given this ordering of the data the permutation approach proceeds as follows:

Step 1: Within each analysis t, simulate data by fixing $(Y_{N(t-1)+1}, \boldsymbol{Z}_{N(t-1)+1}, E_{N(t-1)+1}),\ldots,(Y_{N(t)}, \boldsymbol{Z}_{N(t)}, E_{N(t)})$ and permuting $X_{N(t-1)+1},\ldots,X_{N(t)}$ to create $X^p_{N(t-1)+1},\ldots,X^p_{N(t)}$ to obtain Nperm permuted datasets (p=1,…,Nperm).

Step 2: For each permuted dataset p at each analysis time t calculate the score test statistic $S^p(t)$ from permuted data up to time t $(Y_1(t), \boldsymbol{Z}_1, E_1(t), X^p_1),\ldots,(Y_{N(t)}(t), \boldsymbol{Z}_{N(t)}, E_{N(t)}(t), X^p_{N(t)})$.

Step 3: For each permuted dataset calculate $C^{p}_{\max} = \sup_{t} \frac{S^{p}(t)}{u(t)}$, which is the maximum value of the score statistic across time for that permutation taking into account the desired shape of the boundary.

Step 4: Estimate, a, as $\hat{a} = C^{(1-\alpha)}_{\max}$ which is the (1-α) percentile of $C^{p}_{\max}$.

Step 5: Critical boundary at time t is $c(t) = \hat{a}\, u(t)$.

This simulation framework requires that we have a complete dataset ($Y_i(t)$, $\boldsymbol{Z}_i$, $E_i(t)$,X_i) for all observation times i=1,...,*N(T)*. However, this is not practical at earlier analysis times t < T. To solve this, at times t < T we can instead make assumptions about how the data will look at future analysis times. Specifically, we will assume that future data will look like the current outcome, exposure, and confounder data. To approximate the future distributions of *Y*, *X*, *E*, and ***Z***, we can sample the future observations, N(t)+1 to N(T), by sampling with replacement from the observed ($Y_i(t)$,X_i,$\boldsymbol{Z}_i$,$E_i(t)$) (i = 1,…,N(t)). This will create a complete dataset necessary to perform the permutation approach described previously for all analysis times.

In practice, at each new analysis time we keep the prior boundary values, $c(1),\ldots,c(t\text{-}1)$ since these were the signaling thresholds used at previous analysis times and each analysis time is defined to be conditional on the prior analyses. The simulation plan is slightly altered to take into account the amount of error spent at previous analysis times and incorporating not having all observed data across all analysis times. Specifically we follow the following simulation outline:

Step 1: If at analysis time $t<T$: Create one complete dataset by sampling with replacement $N(T) - N(t)$ observations from ($Y_i(t)$,X_i,$\boldsymbol{Z}_i$,$E_i(t)$) (i = 1,…,$N(t)$).

Given a complete observed dataset:

Step 2: Within each analysis t, simulate data by fixing ($Y_{N(t-1)+1}$, $\boldsymbol{Z}_{N(t-1)+1}$, $E_{N(t-1)+1}$),…,($Y_{N(t)}$, $\boldsymbol{Z}_{N(t)}$, $E_{N(t)}$) and permuting $X_{N(t-1)+1},\ldots,X_{N(t)}$ to create $X^{p}_{N(t-1)+1},\ldots,X^{p}_{N(t)}$ to obtain Nperm permuted datasets (p=1,…,Nperm).

Step 3: For each permuted dataset p at each analysis time t calculate the score test statistic $S^{p}(t)$ from permuted data up to time t ($Y_1(t)$, $\boldsymbol{Z}_1$, $E_1(t)$, X^{p}_1),…,($Y_{N(t)}(t)$, $\boldsymbol{Z}_{N(t)}$, $E_{N(t)}(t)$, $X^{p}_{N(t)}$).

Step 4: For analysis times ($j<t$) already observed and have previous boundaries $c(1),\ldots,c(j)$:
Calculate the cumulative error spent at analysis time j as:

$$\hat{\alpha}(j) = \frac{\sum_{p=1}^{Nperm} I\left(S^{p}(1) \geq c(1) \cup \cdots \cup S^{p}(j) \geq c(j)\right)}{Nperm}$$

and for permutations datasets in which cross the previous boundaries set the current analysis time score statistic, $S^{p}(t)$, to be something large such as 10,000. Do this to make sure that permutation will be treated as signaling in the next step.

Step 6: For each permuted dataset calculate $C^{p}_{\max} = \sup_{t} \frac{S^{p}(t)}{u(t)}$ which is the maximum value of the score statistic across time for that permutation taking into account the desired shape of the boundary.

Step 4: Estimate the current analysis time, a, as $\hat{a}_t = C^{(1-\alpha)}_{\max}$ which is the (1-α) percentile of $C^{p}_{\max}$.

Step 5: Critical boundary at time *t* is $c(t)=\hat{a}_t\, u(t)$, which takes into account previous boundaries and error spent.

These are how the boundaries are calculated for the current Sentinel application.

c. Sequential p-values

Given the sequential monitoring boundaries it is important to further quantify the level of statistical significance either at the time of signal or at the end of study surveillance. To calculate such a p-value one must make certain decisions about how to order a series of test statistics over time. At a one-time analysis it is straightforward to order a given test statistic, *S,* by the data realization k is greater than data realization m if the test statistic $S^k > S^m$. However, there are numerous approaches to choose the ordering of data in sequential monitoring. In our context we only need to order the permuted data realizations compared to observed data realizations at the time of signal (observed S(t)>b(t)) or end of analysis time T. We defined a permuted data realization to be more extreme than the observed data realization if one of the following conditions occurred:

Permuted data realization signaled at previous analysis times: $S^p(1) \geq c(1) \bigcup \cdots \bigcup S^p(t) \geq c(t)$ or

Permuted data realization did not signal at previous analysis times, but the permuted current analysis time score statistic, $S^p(t)$, is greater than the observed score statistic,

$$S^p(t) \geq S(t) \mid \left\{S^p(1) < c(1) \bigcap \cdots \bigcap S^p(t-1) < c(t-1)\right\}$$

Then the empirical sequential p-value is

$$P = \frac{\sum_{p=1}^{Nperm} \mathrm{I}\left(S^p(1) \geq c(1) \bigcup \cdots \bigcup S^p(t-1) \geq c(t-1)\right) + \mathrm{I}\left(S^p(t) \geq S(t) \mid S^p(1) < c(1) \bigcap \cdots \bigcap S^p(t-1) < c(t-1)\right)}{Nperm}$$

$$= \hat{\alpha}(t-1) + \frac{\sum_{p=1}^{Nperm} \mathrm{I}\left(S^p(t) \geq S(t) \mid S^p(1) < c(1) \bigcap \cdots \bigcap S^p(t-1) < c(t-1)\right)}{Nperm}$$

which is the cumulative error spent up to analysis time *t* combined with the probability of the non-signaling permuted datasets observing a more extreme value than the current observed test statistic. Denote the empirical sequential p-value at the time of signal will always be less than the cumulative error spent up to that analysis time, $\hat{\alpha}(t)$.

d. Details of GS GEE Report Quantities

We will now go over each column of the results table, Table 2 (next table), of the GS GEE report for at a signal for analysis 8 comparing MMR+V(X=0) to MMRV(X=1). The first two columns specify the look number and date of look. The 3rd column specifies cumulative number of those exposed to MMR+V at each look. The 4th column provides the cumulative number of outcomes and percent of the MMR+V with outcome (Number of Outcomes/Number of Exposures *100). The 5th and 6th column reports columns 3 and 4 for MMRV. The 7th (MMR+V Adj %Out) and 8th (MMRV Adj %Out) columns provide an adjusted for confounding measure of percent of each exposure group

with outcome. Specifically, first run the following standard logistic regression model at analysis time *t*,

$$\text{logit}(E(Y_i)) = \beta_0 + \beta_X D_i + \boldsymbol{\beta}_Z \mathbf{Z}_i \quad \text{for } i=1,\ldots,N(t)$$

and estimate all of the regression parameters. Then for each observation calculate the estimated probability of outcome if they were on MMR+V,

$$\hat{\mu}_i(X=0) = \frac{\exp(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_Z \mathbf{Z}_i)}{1+\exp(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_Z \mathbf{Z}_i)},$$

and average these estimated probabilities to obtain MMR+V adjusted percent outcome,

$100\sum_{i=1}^{n_t} \frac{\hat{\mu}_i(X=0)}{n_t}$. Similar calculation is done for MMRV adjusted percent outcome, but replace $\hat{\mu}_i(X=0)$ with,

$$\hat{\mu}_i(X=1) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_X + \hat{\boldsymbol{\beta}}_Z \mathbf{Z}_i)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_X + \hat{\boldsymbol{\beta}}_Z \mathbf{Z}_i)}.$$

The 9th column (Adj OR) is from the same model as the adjusted percent outcome and is the $\exp(\hat{\beta}_X)$. The 10th column (Score Test) is the generalized score test statistic, S(t), as specified in this appendix. The 11th column (Boundary) is the sequential boundary derived using the methods detailed in this appendix, Section VI.B.2, *c*(t). The 12th column (Error Spent) is the cumulative error spent up to a given analysis time $\hat{\alpha}(t)$. Note that additional error may not necessarily be spent at each analysis time since the test is based on a standardized test statistic boundary and not an error spending function. This may be advantageous since at earlier analysis times to follow an error spending rule may make the boundary on the test statistic smaller then desired. For example, we are attempting to have a flat boundary on the score test statistic scale so even though no error under the NULL was spent at analysis one we still could have signaled under the alternative hypothesis. If we had forced an error spending function the first boundary value would have been even smaller. However, later in the analysis with more data it would have made it more difficult to signal which is counterintuitive to gaining more statistical information. Even when attempting to have a flat boundary we still had fluctuations due to the nature of not being able to know the outcome, exposure, and confounder distributions at future looks. If they had stayed being similar to look 1 then the boundary would have stayed close to the original 3.407, but since they did change drastically we properly took into account the actual observed distributions and appropriately updated the boundary to hold the overall type I error of 0.05. The final column (Signal) is just a Yes/No indicator if the Score Test Statistic crossed the boundary at a given look.

Since this analysis signaled we also report at the bottom of the table the Sequential P-Value at Signal of 0.03. Note that it is smaller than the error spent at look 8 of 0.038. This is the sequential p-value as outlined in the previous section.