MINI-SENTINEL METHODS DEVELOPMENT

STATISTICAL METHODS FOR ESTIMATING CAUSAL RISK DIFFERENCES IN THE DISTRIBUTED DATA SETTING FOR POSTMARKET SAFETY OUTCOMES

Prepared by: Andrea J. Cook, PhD[1,2], Robert D. Wellman, MS[1], Tracey L. Marsh, MS[2,3], Ram C. Tiwari, PhD[4], Michael D. Nguyen, MD[5], Estelle Russek-Cohen, PhD[5], Zhen Jiang, PhD[5], 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 Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 5. Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD

a l-NJŎK 1, 2012

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

MINI-SENTINEL METHODS DEVELOPMENT

STATISTICAL METHODS FOR ESTIMATING CAUSAL RISK DIFFERENCES IN BOTH THE DISTRIBUTED DATA SETTING AND SEQUENTIAL MONITORING

I. INTRODUCTION 4

II. SUMMARY OF FEATURES OF THE NEW IPTW DISTRIBUTED DATA APPROACH DEVELOPED IN PRISM – ACTIVITY 12 5

III. COMPARISON OF EXISTING BASELINE CONFOUNDING ADJUSTMENT APPROACHES USING PROPENSITY SCORES 7

A. WHAT IS A PROPENSITY SCORE? 7
B. PROPENSITY SCORE MATCHING 8
C. PROPENSITY SCORE STRATIFICATION 9
D. PROPENSITY SCORE WITH INVERSE PROBABILITY OF TREATMENT WEIGHTING 10

IV. STATISTICAL METHODS FOR ESTIMATING THE RISK DIFFERENCE USING INVERSE PROBABILITY OF TREATMENT WEIGHTING IN THE DISTRIBUTED DATA SETTING WITH EXTENSION TO GROUP SEQUENTIAL MONITORING 13

A. INVERSE PROBABILITY OF TREATMENT WEIGHTING (IPTW) APPROACHES WITH NO DISTRIBUTED DATA STRUCTURE OR SEQUENTIAL MONITORING 13
1. *Estimating propensity scores in a single site setting* *13*
2. *Generalized Weighted Least Squares Regression (GWLS) for Risk Difference Estimates* *14*
3. *Risk difference estimates derived directly to incorporate propensity score estimation* *14*

B. INCORPORATING THE DISTRIBUTED DATA SETTING 15
1. *Stratified IPTW Method* *15*
2. *Permutation Test for the Rare Event Data Setting* *16*

C. INCORPORATING GROUP SEQUENTIAL MONITORING 18
1. *Group Sequential Data Framework* *18*
2. *Group Sequential Lan-Demets Method (GS LD)* *19*
3. *Group Sequential IPTW Permutation Approach (GS WPerm)* *19*

V. SIMULATION STUDY TO EVALUATE PERFORMANCE OF IPTW METHODS 20

A. ONE TIME STUDY ANALYSIS 21
1. *Data Structure* *21*
2. *Equivalently sized sites* *22*
a. Prevalence of Exposure is 50% 23
b. Prevalence of Exposure is 25% 23
3. *Mini-Sentinel like site distribution* *23*
a. Prevalence of Exposure is 0.50 23
b. Prevalence of Exposure is 0.25 24
4. *Very Rare Event Setting* *24*
5. *Interaction with exposure by site* *24*

B. GROUP SEQUENTIAL STUDY ANALYSIS 25
1. *Data Structure* *25*
2. *Evenly sized sites* *25*
a. Prevalence of Exposure is 50% 25
b. Prevalence of Exposure is 0.25 26

3. Mini-Sentinel like site distribution 26

a. Prevalence of Exposure is 0.50 26

b. Prevalence of Exposure is 0.25 27

VI. DISCUSSION AND FUTURE WORK 27

A. TASK ORDER DELIVERABLES 27

VII. TABLES AND FIGURES FOR SIMULATION STUDY 29

VIII. REFERENCES 37

I. INTRODUCTION

There is a pressing public health need to monitor the safety of marketed medical products. Therapeutic and prevention products, such as vaccines, drugs, and devices, go through rigorous clinical trials evaluating efficacy and safety before being approved, but these trials are generally not of sufficient size to systematically detect rare adverse events and do not always include participants similar to the population that receives them after their marketing. Therefore, the Food and Drug Administration (FDA) has begun to utilize large multi-site healthcare databases to conduct postmarket surveillance studies for medical product safety. The FDA's Sentinel Initiative is an example of a program designed to improve the evaluation of safety across a large array of FDA-regulated medical products. This task order focuses on the use of Sentinel data for the evaluation of the safety of vaccines for pre-specified acute, short-term, outcomes.

This report presents new statistical methods developed during the task order for assessing safety of a vaccine with a single time exposure using a prospective cohort observational study design with existing electronic healthcare data for pre-specified safety outcomes. It has been assumed that we are in the multi-site distributed data setting in which individual-level covariate data cannot be readily combined across sites due to privacy concerns or propriety information policies. This task order has specifically focused on developing methods that estimate a risk difference since it is generally the key quantity of interest for informing important policy decisions. The methods developed assume that outcomes are acute with everyone having the same outcome window (e.g. 45 days) and therefore outcomes are binary. The new methods can be used to estimate and test for risk differences in both one-time and sequential studies. The research aim is to determine whether, for a prespecified set of safety outcomes, there is an excess rate of observed events in recipients of the vaccine of interest compared with a single comparison group. In this task order, we consider a concurrent control group defined to be comparable to those taking the vaccine of interest after controlling for confounders. For example, when evaluating a new vaccine for safety, an appropriate comparison group could be those coming into the office for a well visit or those who received injections of a comparable vaccine. However, we would still need to control for confounders such as sex, age, comorbidities, and site.

The methods proposed in the task order use site-specific propensity scores with inverse probability of treatment weighting (IPTW) to control for confounding. The approach works well in settings in which the outcome is rare, but where there are numerous confounders that still need to be accounted for. Standard regression methods that directly adjust for these confounders can have estimation problems as the number of confounders increase. In the distributed data setting individual data is not available to fit such regression models. The new method keeps the individual-level data at the site by fitting site-specific propensity models to estimate the probability of exposure given the confounders and then creates site-specific adjusted risk difference estimates using IPTW. These site-specific risk differences are combined across sites using proper statistical techniques to create an adjusted overall risk difference estimate and variance incorporating the variability of the site-specific propensity score models. Further extensions are proposed to the group sequential monitoring setting in which multiple analyses are planned to proactively assess safety issues when a new vaccine is released on the market. The report begins with a brief summary of the new method followed by an overview of different propensity score approaches. Then in Section IV the new statistical approach is presented in detail. We finish with an evaluation of these new methods with a simulation study in Section V.

II. SUMMARY OF FEATURES OF THE NEW IPTW DISTRIBUTED DATA APPROACH DEVELOPED IN PRISM – ACTIVITY 12

Purpose: To develop a new method for the distributed data setting to control for multiple confounders in the concurrent control design with a single time exposure, e.g. vaccine, assessing elevated rates of rare acute outcomes when the quantity of interest is a risk difference. The method proposed is applicable to both a single-time analysis and a group sequential analysis design.

Exposure Type: Acute Exposure (e.g. vaccine)

Outcome Type: Binary Outcome (e.g. acute events that occur in a fixed follow-up period, e.g. 45days)

Quantity of Interest: Risk Difference

Confounding Control: Inverse Probability of Treatment Weighting

- Works well in situations where there are enough exposed and unexposed participants to estimate propensity scores, but outcome is rare causing regression and other adjustment approaches to run into difficulties. The required proportion exposed or unexposed depends on total sample size and number of confounders. If you have a large dataset (>100,000 participants) even with a small proportion of exposed, or unexposed, (<10%) this method can still be viable as the propensity score can often be estimated even with a large number of confounders. However in small datasets a larger proportion of exposed subjects are required as the number of confounders increases. If the proportion exposed (or unexposed) is smaller or similar to the probability of outcome, then this method does not have obvious advantages over standard regression approaches.

Distributed Data Setting: The method involves combining stratified covariate-adjusted risk difference estimates across sites and allows for propensity score models to be constructed separately at individual sites. Specifically,

1. Construct Propensity Score (PS) models, P(exposed|confounders), at each site controlling for pre-specified confounders
2. At each site calculate an adjusted risk difference estimate with corresponding variance using each site's PS model as defined in the report Section IV.A.3
3. To account for differential variability (due to varying sample sizes) across sites, and to not assume normality (which may not hold due to rare outcome prevalence), use an exact method that creates a large number of permuted datasets where outcomes are fixed, but observed exposures are permuted. Calculate adjusted risk difference estimates and corresponding variance using estimates specified in report Section IV.B.2.
4. De-identified data combined across sites: Number of observations, the observed estimated risk difference, the observed estimated variance, and a dataset of permuted estimated risk differences and variances are provided by the sites
5. An overall estimate of adjusted risk difference is calculated as specified in Section III.B.1 and the distribution of this test statistic is calculated using the permuted datasets as outlined in Section III.B.2.

- This method appropriately accounts for site as a confounder and correctly incorporates site-related variability in the risk difference and in the variance estimates. Other methods (e.g., PS matching) typically do not.

Group Sequential Monitoring (appropriate for non-sequential one-time analysis setting as well): Uses an exact permutation approach (versus making asymptotic assumptions) to estimate appropriate distribution of test statistic, including variability of site specific propensity scores, which is important in a rare event setting. Assuming that your test statistic is normally distributed when events are rare may not be appropriate, yielding inflation of type I error. Further, a permutation approach allowed us to easily incorporate differential variability in the PS across sites since we are estimating site-specific PS models while remaining computationally feasible.

Advantages:

1. Estimates the risk difference, a quantity commonly of interest to decision-makers (previous approaches use relative risk)
2. By using a propensity score, can control for a large number of confounders even when events are rare, given enough exposed and unexposed participants
3. Identifiable data remains at sites; only summarized data need to be combined for central analyses
4. Permits a causal inference interpretation given correct specification of the propensity score model, thus mimicking a clinical trial
5. Is highly efficient – generally has more power compared with other propensity score confounding control approaches such as matching (although this is not the case if one includes in the population participants that are either extremely likely, or not likely, to receive the exposure of interest; requires identification of a good control group and restriction or trimming to reduce this problem)

Disadvantages:

1. Less well known compared to other confounding approaches
2. Need to be careful to include a representative unexposed population otherwise very large weights can inflate the variance. This occurs because of how the inverse probability of treatment weighting works. The weight used for the unexposed population is the inverse of their estimated probability of being exposed. If you include in the unexposed population people who were very unlikely to be exposed (e.g. include in the unexposed population those who are 65yrs or older and the recommended schedule for the vaccine, or exposure of interest, is for only those 45-65yrs old) then the estimated probability of being exposed is very small for this subset of the unexposed population. If the estimated probability of being exposed is very small then the inverse probability of treatment weight will be very large, which inflates the variance. Methods such as trimming and restriction can help this problem, but the best practice is to choose a good comparison group first to mitigate this problem in advance.

III. COMPARISON OF EXISTING BASELINE CONFOUNDING ADJUSTMENT APPROACHES USING PROPENSITY SCORES

Numerous approaches have been proposed to deal with confounding that arises in observational cohort studies with concurrent comparators. An increasingly common approach is use of a propensity score. A propensity score is the estimated probability of treatment selection conditional on observed baseline covariates. Propensity scores can be used to control for confounding in a variety of ways, such as matching based on the propensity score, stratifying on the propensity score, or adjusting for the propensity score in a regression setting.[1-3] A less common, but potentially advantageous approach, is to address confounding by using inverse probability of treatment weighting (IPTW).[3-5] IPTW has been shown to allow for a causal interpretation and to closely mimic randomized clinical trials if certain assumptions are met. Use of propensity score approaches in distributed data settings (i.e. where propensity models are constructed separately at sites and then combined across sites) has not been sufficiently studied and may be difficult depending upon which propensity-based confounding control approach is applied. These difficulties are discussed in detail in each specific propensity score method section. Further, incorporating sequential monitoring over time may yield additional complications that could lead to preferences for certain propensity score approaches over others. In this section, we will provide a short summary of the different available propensity score approaches and potential limitations/advantageous that may occur, highlighting issues relevant in distributed data settings and for sequential monitoring. In subsequent sections, we will focus on one specific propensity score approach, IPTW, and how it can be applied in the Mini-Sentinel setting.

A. WHAT IS A PROPENSITY SCORE?

A propensity score is defined as the probability of treatment selection conditional on observed baseline covariates; that is, P($X|Z$) where X is 1 if a subject receives the exposure of interest and 0 otherwise, and Z is a vector of confounders such as age and sex. In large-scale postmarket surveillance, an important additional confounder is site, S, since the probability of receiving the exposure of interest (EOI) and/or having the outcome of interest recorded in the available electronic records may be different across sites. The probability of receiving the EOI may differ by site due to drug protocols, physician preference, or other site-specific characteristics. The probability that an outcome is recorded may differ by site due to differing patient-specific characteristics that are not measured by observed baseline confounders, e.g. race or frailty are often unmeasured in electronic health records. Site variation in outcomes may also exist due to differential use of ICD-9 codes when relying on claims records for capturing outcomes. Due to these specific differences, site is typically also included as a potential confounder in an overall propensity score model for P($X|Z$, S). Alternatively, propensity scores can be calculated within each site, i.e. using a model for $P_s(X|Z)$, where the subscript s refers to a particular site. Site-specific propensity score models intrinsically account for interactions between site and other important variables that are included in the models; however site-specific propensity score models result in additional variability among estimated propensity scores because the sample size for each site-specific model is smaller than if the data were pooled across sites in a single propensity score model. Any approach that uses site specific propensity models should take into account this differential variability when conducting an analysis so that small sites will not "inappropriately" contribute to estimation "like a large site", even though their estimates are much less reliable. In a distributed data setting, often only site-specific propensity score estimates may be possible to compute since individual-level covariate data cannot be readily combined across sites due to privacy concerns or proprietary information policies.

A further complication in using a propensity score approach arises when the study design requires sequential monitoring over time, as if often the case for postmarket vaccine and drug safety evaluations. At the first analyses time, only a small portion of data is available to estimate propensity scores. As time passes, more data become available to better estimate the propensity scores. Certain (design-based) approaches, such as matching or stratification, have difficulty in taking advantage of such new information, since, once a subject is matched, or placed into a propensity stratum, that subject must remain in the same matched set, or specified propensity stratum, for all subsequent analyses. The result of this is reduced control of confounding at early times points stemming from that fact that propensity scores estimated early in the study are less accurate than those estimated later. IPTW and adjusted approaches are not hindered by this limitation. We will discuss this issue further in future sections.

In the rest of this section we will briefly explore and describe the different aforementioned approaches for confounding control in more detail and summarize their potential strengths/weaknesses.

B. PROPENSITY SCORE MATCHING

One popular method of confounding control in observational studies is to match an exposed participant to an unexposed participant based on the similarity of their respective estimated propensity scores. This 1:1 matching can be extended to exposure matching ratios such as 1:M, where M is fixed and is the number of unexposed participants matched to a single exposed participant. This approach is an extension of an exposure matched design that uses individual baseline categorical covariates, such as site, age categories, and sex, to match an exposed participant to an unexposed participant within the same confounder stratum.

Often propensity score matching is applied in combination with standard exposure matching by first stratifying by important confounders such as site, sex, and broad age categories, and then using propensity score matching within strata to implement the best matching scenario. This ensures that, for example, in a study where the outcome is AMI, for which site, age, and sex are strong confounders, a male at site A aged 40-65yrs will not be matched to a female at site B age 70-85yrs simply because they have comparable propensities of exposure.

Given a matched dataset where the matching is done well, the data analysis is straightforward using conditional regression approaches, and confounding is generally well controlled. In practice, matching is often ignored at the analysis phase, but this has been shown to be inefficient (i.e. larger confidence intervals) and may introduce bias especially in the setting of time-to-event data[1]. Matching methods can also be applied in distributed data settings by treating site as a matching covariate. However, to our knowledge, there are no available methods for handling the differential variability of the propensity score across sites. A method that does not address this differential variability, essentially ignores the fact that confounding may not be well-controlled at some sites due to smaller sample sizes and imprecise estimates of propensity scores. This issue may be even more pronounced in the group sequential setting since at earlier analyses imprecision in propensity score estimates resulting from smaller amounts of data can occur, making matching potentially less effective. Further, when one analyzes the data frequently over time, it may be difficult to find a good match for all exposed individuals at each analysis time point. This can result in insufficient confounding control and increased bias. Below we detail the general advantages and limitations of the propensity score matching approach.

Advantages:

- Simple (fixed sampling ratio)
- Intuitively reflects a clinical trial setting, but the population to which the results are generalizable may be altered depending on how matching is conducted
- Reduces sample size making chart review and other data collection more feasible.
- If unexposed population is very different from exposed population then this approach will, when matching is possible, appropriately restricts to those that have the potential to be exposed. Choosing a good unexposed population and/or restricting are alternative ways to handle this issue.

Disadvantages:

- Does not use entire available cohort, resulting in a loss of efficiency
- Reduction of bias relies on how well the matching is done, something that is difficult to assess and may be more easily compromised in a sequential monitoring setting where matching occurs within each new (and possibly small) increment of data
- If the population receiving the exposure of interest is highly specialized then the ability to generalize the results to a broader population may be compromised, potentially hiding adverse events that would occur in the broader population

Limitations:

- Matching methods that take into account differential uncertainty of estimated propensity scores across sites and analysis times in the distributed data and group sequential monitoring settings have not been developed. Potential for future work using bootstrapping, but may be difficult to implement in practice.

C. PROPENSITY SCORE STRATIFICATION

Propensity score stratification is very similar to propensity score matching except that it uses the entire cohort instead of a matched subset. In practice stratification is done by calculating propensity scores and then forming strata based on percentiles of the propensity score distribution. For example, one can create 10 propensity score strata by using the following propensity score percentile categories 0-10%, 11-20%, … , 91-100%. Choice of the number of strata depends on balancing the tradeoff between bias and inefficiency (variance). As the number of strata increases, bias decreases, but so does efficiency. This trade-off is particularly important in the rare event setting since only strata with at least one event are informative. Therefore, the number of subjects informing analyses essentially reduces to only those participants that reside within a stratum with at least one event. Hypothetically, if there were only 5 events in 5 strata, a population of 10,000 in 10 strata might be reduced to, say, 5,000 in 5 informative strata. This occurs because approaches using stratification typically condition on strata in the analysis phase to control for differences across strata.

In the distributed data and group sequential monitoring settings, similar issues arise for propensity score stratification as were described for propensity score matching. A promising method for group sequential monitoring that involves stratification has been developed by Li et al 2011[6], but has been shown to have limitations when very frequent monitoring occurs or if a large number of strata are needed to control for confounding [7]. Further work still needs to be conducted to incorporate the variability of estimated propensity score models over sites and time. The following list gives a general description of advantages and limitations of propensity score stratification.

Advantages:

- As long as an outcome exists in every stratum, the entire available cohort is used and the results are generalizable to the entire cohort population
- Has the flexibility to handle effect modification by site by requiring strata to be formed within site

Disadvantages:

- Selecting the number of strata to establish the appropriate bias-variance tradeoff can be difficult.
- Results are generalizable only to the population from which the subcohort with informative strata (at least 1 outcome in stratum) was drawn.

Limitations:

- Stratification methods that take into account differential uncertainty of estimated propensity scores across sites in the distributed data setting and across analysis times in the group sequential monitoring setting have not been developed. There is potential for future work using bootstrapping, but it may be difficult to implement in practice.

D. PROPENSITY SCORE WITH INVERSE PROBABILITY OF TREATMENT WEIGHTING

Inverse probability of treatment weighting (IPTW) has been used in numerous contexts including sample survey designs and confounding control in observational studies.[4, 8-10] The basic idea is to re-weight the observed sample (a subset of a larger population) using baseline covariate information so that an effect, such as a risk difference, that generalizes to the larger population of interest can be estimated. For example, a survey might be undertaken to assess a specific population, such as all adults aged 20-65 that reside within the city limits of Seattle, and a random sample of this population is identified for study. However, if fewer young people (e.g., 20-35yr olds) complete the survey due to response bias, our random sample will be older than, and no longer representative of, the desired population. If the sampled 20-35yr olds are representative of other 20-35 yr olds, we can account for this by upweighting the observed and undersampled 20-35yr olds to appropriately represent the proportion of all 20-35yr olds living in Seattle, and downweight the oversampled 50-65yr olds to appropriately represent the actual proportion of 50-65yr olds. Downweighting and upweighting allows the researcher to estimate, for example, the approval rating of a mayor in the entire population of Seattle instead of just among those that picked up the phone.

Table 1: Description of how IPTW handles up and downweighting of observations

	Unlikely to receive exposure of interest	Equally likely to receive either treatment	Likely to receive exposure of interest
Exposed	Upweighted	Neutral-weight	Downweighted
Unexposed	Downweighted	Neutral-weight	Upweighted

A similar construct can be applied to control for confounding using IPTW. Propensity scores are used as inverse probability weights to upweight those that were unlikely to receive the treatment that they actually did receive, while downweighting those that were more likely to receive the treatment and did receive the treatment (Table 1). Similarly, among those that did not receive the treatment, the inverse probability weights will upweight those that were likely to receive the treatment, but did not actually receive the treatment, and downweight those that were unlikely not to receive the treatment and did not receive it. Those that are 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. Under the assumption that there is no unmeasured confounding, IPTW mimics a randomized clinical trial, and the estimate is termed the causal effect estimate since it is estimating the effect as if the *entire* population received the exposure of interest relative to the effect if the *entire* population was unexposed. This is different then estimating the effect among those that received the exposure of interest relative to the effect among those that did not receive the exposure of interest conditional on covariates (conditional regression methods).

A variety of approaches have been used in practice to estimate causal effects using IPTW weights.[3] These approaches take into account the variability of the weights in estimating the variance, usually through the use of bootstrapping.[11] In Sections IV and V of this report we will propose and evaluate two extensions of these approaches: 1) to the distributed data setting for a single time analysis and 2) to the distributed data setting with a group sequential analysis. In addition, these methodological extensions are designed to estimate a causal risk difference between those exposed to a vaccine of interest and a comparison population and to perform well statistically even in the rare event setting, both of which are useful when evaluating postmarket safety endpoints. In the group sequential analysis setting we are able to directly account for the change in variability of the propensity score over time, but still use all of the data available at the new analysis to calculate updated, and more stable, propensity scores to estimate an adjusted risk difference. The proposed method does not require that those who were exposed (and therefore entered the study) earlier keep their original propensity score weights, but it instead allows their propensity scores to be updated using the most recent information while still holding the statistical properties that we desire (e.g., unbiased estimation for both the risk difference and variance and type I error). Below are general descriptions of advantages and limitations of this methodology.

Advantages:

- Uses the entire cohort and results are generalizable to population of interest
- Estimates effects with a causal interpretation under appropriate assumptions
- Generalizes to the entire cohort instead of to a restricted cohort
- Can flexibly handle effect modification by site by separately modeling propensity scores at each site and then appropriately combining data across sites taking account differential variability
- Has been shown to reduce the most bias and have the highest power compared to other propensity score approaches.[12, 13]

Disadvantages:

- Less familiar and less well understood in the general research community
- Bias/variance tradeoff needs to be taken into account. Observations with very large weights can inflate variance so methods such as trimming and restriction need to be assessed. [8] However, trimming and restriction may increase bias so a careful understanding of tradeoffs should be obtained before applying these approaches by conducting sensitivity analyses

An overview outlining the general differences between propensity score approaches can be found in Table 2.

Table 2: Comparison of Propensity Score Approaches to Adjust for Baseline Confounding

	Gold Standard: Regression Model	**PS Matching**	**PS Stratification**	**PS IPTW (new)**
Basic Description	Controls for confounding by adjusting for all individual confounders directly in model	Fixed ratio matching of exposed and unexposed by similar PS. Reduces population to a smaller subset of original dataset	Matches entire exposed and unexposed cohort by PS strata. Most methods limit to those with informative strata (at least one outcome in strata)	Re-weight observed data by inverting the PS to generate a more generalizable population (upweight low PS and downweight high PS)
Challenges for Application in Distributed Settings	▪ If data are combined across sites then confidentially issues ▪ Estimating site specific regression models usually not feasible (rare event and many confounders)	▪ Site-specific PS models imply more variable PS estimates ▪ No available methods incorporate this variability across sites. ▪ Can be difficult to determine amount of confounding control. With different models across sites: What is a good match?? How well have we matched at different sites (especially at small sites)? How do we interpret extent/quality of overall confounder adjustment across sites?	▪ Site-specific PS models imply more variable PS estimate ▪ No available methods incorporate this variability across sites. ▪ Number of strata can be large (number of sites times number of PS strata) which reduces number of informative strata ▪ Larger can have more strata then smaller sites and therefore better confounder control. How do we deal with this in the analysis and interpret results accordingly?	▪ Site-specific PS models imply more variable PS estimate ▪ Can handle variability across sites by correctly calculating the stratified variance. ▪ Sensitive to choice of exposure population and making certain observations too informative (restriction and trimming necessary)
Challenges to Application Sequentially	▪ Assumptions such as normality may not be met in rare event setting	▪ Earlier matching (and thus confounding control) may be poor ▪ Must keep matching fixed (no re-matching as more data accrues) to properly calculate test statistics over time and to not allow change in population (those included in the unexposed initially should not be dropped from future analysis) ▪ How well are we controlling for confounding?	▪ More heterogeneity (and thus more residual confounding) may exist within strata at earlier analyses due to more variable PS estimation ▪ Must keep people in original confounding strata to make sequential inference ▪ Number of strata can be large since it grows by number of analyses x number of sites x number of PS strata.	▪ Earlier confounder control may be less well done, but this is incorporated in the variability of the PS estimated risk difference. ▪ Assumptions need to be made about future proportion of exposed to estimate sequential boundaries

IV. STATISTICAL METHODS FOR ESTIMATING THE RISK DIFFERENCE USING INVERSE PROBABILITY OF TREATMENT WEIGHTING IN THE DISTRIBUTED DATA SETTING WITH EXTENSION TO GROUP SEQUENTIAL MONITORING

As discussed in previous sections there are several approaches for using propensity scores to deal with confounding arising from observational cohort studies. However, there is sparse literature evaluating the performance of IPTW methods in a distributed data setting like Mini-Sentinel (where typically individual level data remains at sites and only de-identified summary data is available for analysis) or detailing how to use them sequentially over time. In this section we extend IPTW methods to two specific postmarket safety surveillance settings: 1) Data are distributed across multiple sites (i.e. propensity models are formed independently at sites and then combined across sites), and 2) Data (that are either distributed or not distributed across sites) are group sequentially monitored over time. In Section V we will report the details and results of a simulation study evaluating the proposed methods and comparing them to existing approaches.

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. We will review IPTW statistical methods that are available for estimating the risk difference in this simple observational setting. Then in Section IV.B we will propose a new method that properly incorporates the distributed data structure. Finally in Section IV.C we will propose how to extend both distributed and non-distributed IPTW methods for group sequential monitoring.

A. INVERSE PROBABILITY OF TREATMENT WEIGHTING (IPTW) APPROACHES WITH NO DISTRIBUTED DATA STRUCTURE OR SEQUENTIAL MONITORING

The IPTW approach can be applied using estimating equations to estimate quantities of interest including odds ratios (OR), relative risks (RR), and risk differences (RD). In this task order, we focused on the risk difference (RD) since it is a highly relevant quantity of interest for medical decision making and policy changes, particularly for vaccines. However, it is also possible to similarly extend approaches involving RRs and potentially to ORs, but this is less straightforward. Numerous studies have already assessed the performance of applying the IPTW approach for OR and RR effect estimation,[13-15] but few have evaluated these approaches for the RD.[3, 12] In this section we will first detail general methods for the pooled (non-distributed) data setting, including introducing notation and presenting standard approaches for IPTW risk difference estimates.

1. Estimating propensity scores in a single site setting

Assume at a single site, *s*, we have outcome Y_{si} ($i=1, \dots , N_s$), with treatment X_{si} , equal to 1 if subject *i* at site *s* has the exposure of interest, and equal to 0 otherwise, and let $\boldsymbol{Z}_{si}$ be a set of measured confounders. Define, X_{si} to be 1 or 0 if the subject *i* at site s is exposed to the treatment or not. Then define the propensity score, $e_{si,}$ as the probability of receiving, i.e. being exposed to, the treatment X_{si} given confounders $\boldsymbol{Z}_{si}$, so that e_{si}=P(X_{si}=1|$\boldsymbol{Z}_{si}$). We can estimate e_{si} using logistic regression assuming a logistic model, logit(E(Y_{si}))=$\boldsymbol{Z}_{si}\boldsymbol{\beta}_{\boldsymbol{Z}}$, where $\boldsymbol{\beta}_{\boldsymbol{Z}}$ is estimated using the maximum likelihood approach. This is typically done in practice and yields $\hat{e}_{si}$ =(1+exp(-$\boldsymbol{Z}_{si}\boldsymbol{\beta}_{\boldsymbol{Z}}$))$^{-1}$. 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 to 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 a population unconfounded average effect estimate.

2. Generalized Weighted Least Squares Regression (GWLS) for Risk Difference Estimates

This approach uses standard generalized weighted least squares regression to estimate a risk difference. Specifically it assumes a linear regression model, but incorporates confounding adjustment using inverse probability weighting. To calculate a risk difference we assume the following outcome distribution,

$$Y_{si} \sim \mathrm{N}(\beta_0+\beta_X X_{si}, \sigma^2 w_{si}),$$

where

$$w_{si} = \begin{cases} \dfrac{1}{\hat{e}_{si}} & \text{if } X_{si} = 1 \\ \dfrac{1}{(1-\hat{e}_{si})} & \text{if } X_{si} = 0 \end{cases} .$$

In this setup, since our outcome Y_{si} is binary, β_0 estimates the probability of outcome in the unexposed group and β_X is the risk difference estimate of interest. σ^2 is a nuisance parameter estimating the variability of Y_{si}, but the IPTW are incorporated in the variance component of the model to give more weight (lower variance) to those less likely to receive the treatment that they actually received and less weight (larger variance) to those more likely to receive the treatment that they actually received. The model parameters can be estimated using standard weighted least squares regression. However, since the outcome is binary and not normally distributed it is better to estimate the variance using robust standard errors.[16] This approach does not take into account the variability of the propensity scores. The following methods will further this approach to account for the fact that the propensity scores are estimated and may be differential across sites.

3. Risk difference estimates derived directly to incorporate propensity score estimation

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. 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[5] 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 taking into account that e_{si} is estimated. The formula for the variance is given by,

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

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

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

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.

In the following section, we will extend this approach to the distributed data setting in which propensity scores and corresponding test statistics are estimated at each site and then combined across sites to test scientific hypotheses. We will use the scenario where data are pooled across sites to evaluate the operating characteristics and potential information loss of the combined site-specific estimates as part of our simulation study.

B. INCORPORATING THE DISTRIBUTED DATA SETTING

1. Stratified IPTW Method

A variety of approaches exist for combining data across sites. The most straightforward approach is to use a stratified modeling approach 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

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

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, $\hat{\mathrm{V}}(\hat{\Delta}_s)$. 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 present this approach in our simulation study. Note that other effect estimates, such as a relative risk, could be weighted using a similar approach.

Another standard approach would be to pool the individual-level data across sites and treat site as a covariate. This model will be treated as the gold standard since efficiency gains should be realized by using pooled data, but if there is an interaction with confounders by site it may lead to a biased estimate. A third approach for combining data across sites, which we will call the naive approach, is to combine estimates but ignore the fact that the propensity scores were modeled differently across sites. Specifically, using the site-specific propensity weights as if they were derived in a single model pooled across sites and just calculating a weighted risk difference estimate directly.

In Section V we will evaluate approaches using different assumptions to address which method is preferred. Since the naïve method fails to account for the uncertainty and potential differential sites effects that arise when propensity scores estimated from different site-specific models are combined together, we do not assess it here. Our simulation study evaluates a standard linear regression approach adjusting for covariates directly in the model with robust standard errors for estimation of the risk difference with data pooled across sites (referred to as GLM), the IPTW approach with data pooled across sites using unstratified estimator from Section IV.A.3, and the new stratified IPTW estimator (referred to as IPTW_s) which estimates site-specific IPTW estimates and then combines them across sites in the manner detailed in Section IV.B.1. This framework allows us to assess the differences between the commonly used GLM regression adjustment approach and the IPTW approach in the ideal situation where individual-level data are shared (pooled) across sites, as well as the performance of the stratified estimator relative to these other two.

2. Permutation Test for the Rare Event Data Setting

The previous sections have developed a framework for mean and variance estimation of an IPTW risk difference estimate and addressed issues specific to the distributed data setting. These estimates have been shown to be asymptotically normal using standard central limit theory reasoning.[3] However, in the rare event setting it is often preferable to use non-parametric derivations when performing tests of statistical significance in order to better hold important statistical properties such as type I error. In this section we will describe one such method, a permutation approach, for this one-time analysis setting and in the Section IV.C.3 will extend this approach to the group sequential monitoring setting.

The null hypothesis of interest in the postmarket safety setting is that there is no treatment effect, i.e. H_0: Δ_s=0, implies that X_{si} is independent of Y_{si} conditional on confounders $\boldsymbol{Z}_{si}$. Therefore, to derive a permutation test under the null, which is used to estimate p-values, one can simply permute all X's while fixing the outcome and confounder data as observed, $((Y_{s(1)}, \boldsymbol{Z}_{s(1)}), \ldots , (Y_{s(Ns)}, \boldsymbol{Z}_{s(Ns)}))$, resulting in a permutation, $\boldsymbol{X}_s^p$ =($X_{s(1)}^p$, … , $X_{s(Ns)}^p$), where p indicates the pth of N$_{\text{perm}}$ permutations. However, since we are randomly permuting X's we observe that the propensity score for the permutated dataset is constant, i.e. P($\boldsymbol{X}_s^p$ |$\boldsymbol{Z}$)=P($\boldsymbol{X}_s^p$)= $\sum_{i=1}^{N_S} X_{si} / N_s$, since X_s^p is independent of $\boldsymbol{Z}$**,** for all p. Once we fix the propensity scores to be constant, we must incorporate this into the estimate of variance of the estimator for the permuted data. Specifically, for the permuted data the estimated is not $\mathrm{V}(\hat{\Delta}_s)$ but $\tilde{\mathrm{V}}(\hat{\Delta}_s)$, where the correction due to estimated propensities has been removed. Keeping the propensity scores constant allows for computational efficiency when calculating the following stepped permutation algorithm:

For p=1,…,N_p where N_p is large,

Step 1: Within each site permute observed X's to form ($X_{s(1)}^p$, … , $X_{s(ns)}^p$)
Step 2: Set all propensity scores within each site s to be, e_{si}= $\sum_{i=1}^{N_S} X_{si} / N_s$.
Step 3: Calculate $\hat{\Delta}_s$ and $\tilde{\mathrm{V}}(\hat{\Delta}_s)$ on the permuted data to create $\hat{\Delta}_s^p$ and $\tilde{\mathrm{V}}^p(\hat{\Delta}_s)$, respectively.
Step 4: If not in the distributed data setting then return $Z_s^p = \hat{\Delta}_s^p \Big/ \sqrt{\tilde{\mathrm{V}}^p(\hat{\Delta}_s)}$, otherwise calculate $\hat{\Delta}^p$ and $\tilde{\mathrm{V}}^p(\hat{\Delta})$ from the site specific estimates and return $Z^p = \hat{\Delta}^p \Big/ \sqrt{\tilde{\mathrm{V}}^p(\hat{\Delta})}$.

Given the N_p permuted test statistics under the null one can calculate an empirical, one-sided p-value as the following,

$$P = \frac{\sum_{p=1}^{N_p} I(Z^p \geq \hat{Z})}{N_p}$$

where $\hat{Z}$ is the standardized test statistic computed from the observed, non-permuted dataset using risk difference, $\hat{\Delta}_s$,and variance of the risk difference, $\hat{\mathrm{V}}(\hat{\Delta}_s)$, in the non-distributed setting, or risk difference, $\hat{\Delta}$, and variance, $\hat{\mathrm{V}}(\hat{\Delta})$, in the distributed data setting. We use a one-sided p-value instead of two-sided since the hypothesis often evaluated is if the exposure of interest has a higher risk of outcome compared to the unexposed population. If a two side p-value is of interest then simply taking the absolute value of both the permuted test statistics and observed test statistic will result in a two-sided p-value.

Note that the permutation test will provide tests of statistical significance, but not provide 95% confidence intervals for the risk difference estimates. One can do this using the Wald-type estimates with the variance, $\hat{\mathrm{V}}(\hat{\Delta})$, or can conduct non-parametric bootstrap approach which will mimic the permutation approach for hypothesis testing. The bootstrap approach requires sampling with replacement from the observed data, ((Y_{s1}, $\boldsymbol{Z}_{s1}$, X_{s1}), … , (Y_{sNs}, $\boldsymbol{Z}_{sNs}$,X_{sNs}), a full set of bootstrapped data (($Y_{s(1)}$, $\boldsymbol{Z}_{s(1)}$, $X_{s(1)}$), … , ($Y_{s(Ns)}$, $\boldsymbol{Z}_{s(Ns)}$,$X_{s(Ns)}$). On each of the bootstrapped datasets the propensity model must be refit at each site and the adjusted risk difference estimates must be recalculated. Requiring the propensity model to be refit each time adds significant amounts of computational time and potential for issues with model convergence, which makes the permutation approach much more feasible especially in the next section when discussing extension to group sequential monitoring.

C. INCORPORATING GROUP SEQUENTIAL MONITORING

Another nuance of postmarket vaccine safety surveillance is the need to detect elevated rates of adverse events as quickly as possible, which motivates the use of routine monitoring. When multiple tests are performed over time and stopping rules for signaling a safety problem are defined, formal group sequential monitoring approaches are necessary to hold the overall false positive error rate (type I error). We will propose two general group sequential monitoring methods that incorporate IPTW. The first method uses a standard group sequential approach derived in the context of randomized clinical trials for more common outcomes.[17] The second approach is derived in the context of flexible regression methods for the rare adverse event setting.[18] In the simulation study we will evaluate performance of these methods in the rare event setting and with a relatively common testing frequency, which may be important for active surveillance studies.

1. Group Sequential Data Framework

Now that we are in the context of group sequential monitoring we must introduce the concept of multiple analysis times. Specifically, 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 vaccine, X_{si}(t)=1, or not exposed, 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 everyone has the same follow-up time. There are other data lag time issues that are also standard that have been discussed elsewhere and will be not discussed in this report.[19, 20] Further, assume that the cumulative number of participants observed at site *s* up to analysis time *t* is $N_s(t)$. If in the distributed data setting, further assume that 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 who were either exposed or unexposed to the vaccine. 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. A general review of sequential monitoring boundaries has been presented by Emerson et al[21] and is beyond the scope of this report, but we will present an approach specific to the observational surveillance setting and one general existing method used in randomized clinical trials that can be applied in an observational setting.

2. Group Sequential Lan-Demets Method (GS LD)

This is a standard group sequential method developed by Lan and Demets[17] that assumes a normal distribution to derive a sequential monitoring boundary using a specified error-spending function. The derived boundary can then be used to compare any normally distributed test statistic. Specifically, an error spending approach uses the concept of cumulative alpha or type I error, $\alpha(t)$, defined as the cumulative amount of type I error spent at analysis t and all previous analyses, 1, … , t-1. We assume that $0<\alpha(1)< \alpha(2)<\ldots< \alpha(T)= \alpha$, where α is the overall type I error to be spent across the evaluation period. The function $\alpha(t)$ can be any increasing monotonic function that preserves family-wise error[21], but there are several common approaches including the Pocock boundary function $\alpha(t)$=log(1+(exp(1)-1)$N(t)/N(T))\alpha$, O'Brien-Fleming boundary function $\alpha(t)= 2\left(1-\Phi\left(Z_{1-\alpha/2}/\sqrt{N(t)/N(T)}\right)\right)$, and the general power boundary function $\alpha(t)=(N(t)/N(T))^h\alpha$ for $h>0$. The most commonly used boundary function for safety evaluations has been a flat, Pocock-like, boundary on a standardized test statistic scale. Compared to an O'Brien Fleming boundary, which is commonly used in efficacy studies, this boundary spends more α at earlier versus later analyses given the amount of statistical information, or sample size, observed up to time t. This flat boundary has been discussed as Pocock-like, but a Pocock boundary when testing more frequently (quarterly or more often) is not completely flat.[7] For further discussion of boundary shapes and statistical trade-offs between them in practice for postmarket surveillance see Nelson et al[19].

Given the error spending boundary function, Lan and Demets developed an asymptotic conditional sequential monitoring boundary for any asymptotically normal test statistic based on independent increments of data. This boundary can be computed and used to compare to almost any standardized test statistic, including one that controls for confounding. For our setting of IPTW risk difference the standardized test statistic is, $Z\_s(t)= \hat{\Delta}_s(t)\Big/\sqrt{\hat{V}(\hat{\Delta}_s(t))}$ for non-distributed data, and $Z(t)=\hat{\Delta}(t)\Big/\sqrt{\hat{V}(\hat{\Delta}(t))}$ for distributed data. The value of $Z(t)$ (resp $Z\_s(t)$) is then compared to the asymptotic conditional monitoring boundary developed by Lan and Demets resulting either in a decision to stop if $Z(t)$ (resp $Z\_s(t)$) exceeds the monitoring boundary or in a decision to continue collecting additional data (if at end of study then stop for no evidence of elevated risk). This is an appealing approach because the boundary is very simple to calculate and relies on a well-defined asymptotic distribution. However, in practice with rare events and frequent testing (which implies a small amount of new information between analyses) the asymptotic properties of the boundary fail to hold. This problem is analogous to the scenario where an exact test may be preferred to an asymptotically normal test when the sample size is small. The new method that we introduce in the next section has sought to address the shortcomings of this approach to allow for better statistical performance in a wider variety of settings.

3. Group Sequential IPTW Permutation Approach (GS WPerm)

This method will extend the permutation approach initially presented in Section IV.B.2 for use in the group sequential setting. For the group sequential boundary formation we will use a general unifying boundary definition developed by Kittleson and Emerson.[22] 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 design (e.g. $u(t)$=1 is Pocock-like and $u(t)=(N(T)/N(t))^{0.5}$ is O'Brien and Fleming-like). One solves for the constant *a* using an iterative simulation approach to hold the overall type I error at α. Specifically, *a* can be obtained via the following stepwise process:

Step 1: At analysis time *t* permute data following the permutation approach in Section IV.B.2 to obtain a permuted standardized test statistic $Z^p(t)$.

Step 2: Choose a value of *a*; if $Z^p(t) \geq a$, then Sig_p=1 (signal) and stop, otherwise continue to next *t*+1.

Step 3: If $t=T$, then Sig_p=0 (no signal).

This process is repeated a large number of times, N_{perm}, and the empirical α-level for the boundary is calculated as $\hat{\alpha} = \sum_{p=1}^{N_p} \mathrm{Sig_p} / N_{perm}$. One solves for *a* by repeating the simulation and changing *a* until $\hat{\alpha} = \alpha$, the desired type I error.

This simulation framework requires that we have a complete dataset, (X_{si}, $Y_{si}(t)$, $\mathbf{Z}_{si}$), for all analysis time points t=1, … ,T . However, this is not practical at earlier analysis times $t<T$. To get around this, at times $t<T$ we can instead make assumptions about how the data look at future time points. Specifically, to derive the permutation approach under the null we only need to know the prevalence of *X* and *Y* at future looks since P($X|Z$)=P(X) under the null. Therefore, to approximate the future prevalence of *X* and *Y,* we can sample the future observations, $N(T)$-$N(t)$, by sampling with replacement from the observed (X_{si},Y_{si}). This will create a complete dataset necessary to perform the permutation approach described previously for all analyses.

In practice, at each new analysis time we will keep the prior critical values $c(1)$, … ,$c(t$-1) since these were the signaling thresholds used at previous analysis times. Using these values, we will then solve for the current analysis time critical value $c(t)$ using the newly updated observed information which may have different prevalence of outcomes compared to what we had assumed during previous analysis times and potentially different sample sizes than initially planned. Allowing both different expected outcome prevalence and sample size at a given look affects the variability of the estimator and therefore the corresponding signaling threshold *c*. Therefore, at each analysis time the boundary will be updated in order to mimic the original boundary family (i.e. stay constant if Pocock-like boundary), but it will move slightly compared to the initially planned boundary in order to keep the overall type I error constant as the variability of the data changes over time.

V. SIMULATION STUDY TO EVALUATE PERFORMANCE OF IPTW METHODS

The following simulation study was conducted to evaluate the operating characteristics of a stratified IPTW estimator of the risk difference due to a specific exposure of interest in the context of postmarket safety surveillance. As detailed above, IPTW is used to control for confounding due to variables other than site, and stratification is used to control for confounding due to site. This is because in the postmarket surveillance setting obtaining sufficient amounts of data to detect relatively rare adverse events often requires the involvement of multiple Data Partners at different sites. Barriers to effective data sharing, such as privacy concerns and proprietary information policies, makes pooling of individual-level data across sites rarely used unless deemed critical to the question of interest, and so stratified methods that allow for efficient use of site-specific, individual-level data are desirable. Furthermore, it is necessary to develop methods capable to conduct single time analyses, as well as group sequential analyses on accruing data. In both instances, it is important to understand the operating characteristics of proposed estimators in order to effectively evaluate quantities such as the false positive rate (type I error) and the average time to detection of a true signal.

Section V.A below details the design and results of simulations evaluating the proposed stratified IPTW estimator for a two-year, one-time study analysis compared to the IPTW estimator and the gold standard adjusted regression estimator applied to data pooled across sites. Section V.B provides the same information but in the context of a two-year study where multiple analyses are conducted at particular intervals over the study period, i.e., a group sequential analysis. In both the one-time and group sequential analyses the risk difference due to the exposure of interest is varied, the two-year incidence of outcome in the unexposed portion of the population varies from 1% to 5%, the proportion of exposed individuals in the study population is varied from 25% to 50%, the proportionate distribution of the sample among the sites changes from an equal sample size at each of three sites to a Mini-Sentinel-like site distribution of one small site (10%) and two larger sites (45% each), and the strength of confounding by site takes on three different configurations: 1) smaller odds of exposure in two sites compared to a third , 2) no difference in exposure odds between the three sites and 3) greater odds of exposure in two of the sites relative to the third. The strength of confounding by other variables such as age and sex is held fixed as well as the total sample size of 10,000 across simulations. In the case of a one-time study analysis, we further assess a very rare event setting where the two-year incidence of outcome in the unexposed is 0.02% and the sample size is increased to 100,000, as well as a setting where the effect of certain confounders on exposure differs by site, i.e., site modifies the effect of another variable with respect to exposure.

A. ONE TIME STUDY ANALYSIS

1. Data Structure

For this type of scenario, the outcome of interest, $Y_i(t)$, is a binary outcome that is 1 if the outcome occurred during the fixed follow-up period (e.g. 45 days after taking vaccine), or 0 otherwise. Below is the specific step-wise simulation design for creating a dataset of N study participants for ($i=1,\ldots,N$);

1) Start date, D_i, is the time in which individual i is enrolled in the study and this is uniformly distributed throughout the two-year (720 day) study, D_i ~ Discrete Uniform(1,719);
2) Site distribution, S_i~Multinomial($p=(p_1,p_2,p_3)$), represents the proportionate distribution of study participants among three sites. In this study we have explored two potential site distributions. The first is an equal distribution of participants across the three sites with p=(1/3,1/3,1/3). The second, sometimes referred to as a Mini Sentinel-like distribution, yields a distribution where two sites are quite large relative to the third with p=(0.10, 0.45, 0.45). The variable S_i is generated from the multinomial distribution, and then the two corresponding binary (dummy) variables, S_{i1} and S_{i2}, are generated for use in regressions and calculations using design matrices.
3) Confounder distributions : The simulations performed for this study include a simple binary confounder (sex), Z_{1i}, which is distributed as Bernoulli(0.50), and a continuous confounder (age), Z_{2i}, which is distributed as Uniform(35,65) and then centered at 50 and scaled so that a one-unit change is equivalent 10 years. Additionally, the site variable comes from a multinomial distribution as detailed in 2) above.
4) Exposure distribution conditional on confounders, two different scenarios:
    a. Site is a confounder but does not interact with other confounders:

$$X_i \mid \mathbf{S}_i, \mathbf{Z}_i \sim Bernoulli\left(\frac{\exp(\beta_{x,0} + \beta_{x,s}\mathbf{S}_i + \beta_{x,z}\mathbf{Z}_i)}{1+\exp(\beta_{x,0} + \beta_{x,s}\mathbf{S}_i + \beta_{x,z}\mathbf{Z}_i)}\right)$$

b. Site is a confounder and interacts with other confounders:

$$X_i \mid \mathbf{S}_i, \mathbf{Z}_i \sim Bernoulli\left(\frac{\exp(\beta_{x,0} + \beta_{x,s}\mathbf{S}_i + \beta_{x,z}\mathbf{Z}_i + \beta_{x,sz}\mathbf{S}_i\mathbf{Z}_i)}{1+\exp(\beta_{x,0} + \beta_{x,s}\mathbf{S}_i + \beta_{x,z}\mathbf{Z}_i + \beta_{x,sz}\mathbf{S}_i\mathbf{Z}_i)}\right)$$

In both scenarios a and b, exp($\beta_{x,0}$) = P(X_i|$\mathbf{S}_i$=0, $\mathbf{Z}_i$=0), which in the confounding scenario considered in this study represents the probability of exposure for men (or women depending on which sex is coded as 1 or 0) in site 1 who are 50 years old. Furthermore, for each coefficient other than $\beta_{x,0}$, exp(β) is the odds ratio of exposure associated with a one unit change in that particular variable holding the other variables fixed. For each simulation we solved for $\beta_{x,0}$ so that the overall probability *of X* was fixed at either 50% or 25% across all simulation configurations. Details of the methods used for finding this solution are provided in the appendix.

5) Outcome distribution conditional on exposure and confounders:

The distribution of the outcome Y_i is Bernoulli with mean defined by the following logistic model:

$$Y_i \mid X_i, \mathbf{S}_i, \mathbf{Z}_i \sim Bernoulli\left(\frac{\exp(\beta_{y,0} + \beta_{y,x}X_i + \beta_{y,s}\mathbf{S}_i + \beta_{y,z}\mathbf{Z}_i)}{1+\exp(\beta_{y,0} + \beta_{y,x}X_i + \beta_{y,s}\mathbf{S}_i + \beta_{y,z}\mathbf{Z}_i)}\right).$$

Markov Chain Monte Carlo integration is used to set the marginal, population-level probability of an event, P(*Y*|*X*=0) to the desired level, 1%, 5% or 0.02% in the very rare case, and then setting the marginal probability of an event in the exposed to P(*Y*|*X*=0)+*RD* where *RD* defines the desired risk difference, which, in this study, is varied between 0 (no effect) and P(*Y*|*X*=0)*1.5. In essence this involves solving for $\beta_{y,0}$ above such that P(*Y*|*X*=0) equals the desired probability of outcome in the unexposed, and then using that solution to solve for $\beta_{y,x}$ such that P(*Y*|*X*=1)=P(*Y*|*X*=0)*1.5. The details of this procedure are provided in the appendix.

This simulation set-up allows us to vary strength of confounding by manipulating exp($\beta_{y,z}$) or exp($\beta_{x,z}$), and to set the marginal probabilities of exposure and outcome.

2. Equivalently sized sites

Situation:

- One-time analysis comparing methods across probability of the outcome of 1% and 5% for a two-year study with 10,000 participants
- Sample is equally distributed across three sites, i.e. *p*=(1/3, 1/3, 1/3)
- Three different effect sizes are examined: No effect (Power = Type I error), as well as 1.25 and 1.5 times the probability of outcome in the unexposed group. The resulting true risk differences are given in Tables 1a & 1b
- The odds ratio of exposure due to a one unit change in the binomial and continuous confounders is fixed at 2 and the site effect varies in magnitude from 0.5 to 1 (no effect) to 2 and is fixed for two of the sites relative to the third
- The odds ratio of outcome relative to a one unit change in the binomial, continuous and site confounders is fixed at 2

a. Prevalence of Exposure is 50%

Results for this section are provided in Table 1a, Section V11.
Main Conclusions:

- All three methods appear to approximately hold type I error
- As probability of outcome increases, power increases
- Stratified IPTW estimator performs comparably to both adjusted regression (GLM) and IPTW estimators using data pooled across sites
- Change in site-specific confounding strength and/or direction only modestly effects power

b. Prevalence of Exposure is 25%

Results for this section are provided in Table 1b, Section V11.
Main Conclusions:

- All three methods approximately hold type I error for 1% prevalence by consistent inflation of type I error occurs for the GLM method for 5% prevalence especially as two sites are more likely to be exposed compared to one site
- As probability of outcome increases power increases, though power tends to be lower than in the case of 50% exposure, especially for the IPTW estimators
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites

3. Mini-Sentinel like site distribution

Situation:

- One-time analysis comparing methods across probability of the outcome of 1% and 5% for a two-year study with 10,000 participants
- 10% of study sample comes from one site and the remaining 90% are split evenly across the other two sites, i.e. *p*=(0.10, 0.45, 0.45)
- Three different effect sizes are examined: No effect (Power = Type I error), as well as 1.25 and 1.5 times the probability of outcome in the unexposed group. The resulting true risk differences are given in Tables 2a & 2b
- The odds ratio of exposure due to a one unit change in the binomial and continuous confounders is fixed at 2, and the site effect varies in magnitude from 0.5 to 1 (no effect) to 2 and is the same for two of the sites relative to the third
- The odds ratio of outcome relative to a one unit change in the binomial, continuous and site confounders is fixed at 2

a. Prevalence of Exposure is 0.50

Results for this section are provided in Table 2a, Section V11.
Main Conclusions:

- All three methods appear to approximately hold type I error, though type I error is slightly inflated when site confounding set at 0.5 and slightly deflated when site confounding is set at 2 for all estimators
- As probability of outcome increases power increases
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites
- At a true relative risk of 1.25, when confounding by site is present, power is decreased for all estimators at both 1% and 5% incidence (as confounding effect increases).

b. Prevalence of Exposure is 0.25

Results for this section are provided in Table 2b, Section V11.
Main Conclusions:

- All three methods appear to approximately hold type I error, though type I error is slightly deflated for all estimators when site confounding is set at 0.5 in the case of 1% incidence
- Type I error for the adjusted regression estimator (GLM) is elevated for the case of 5% incidence for all site confounding settings similar to what was found for equally sized site distribution
- Type I error is slightly deflated for the IPTW and stratified IPTW estimators when no site confounding is present in the case of 5% incidence
- As probability of outcome increases power increases, though power tends to be lower than in the case of 50% exposure (when accounting for elevated type I error)
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites

4. Very Rare Event Setting

Results for this section are provided in Table 3a and 3b, Section V11.
Situation:

- One-time analysis with ***probability of outcome of 0.02%*** comparing methods across prevalence of exposure of 50% and 25% for a two-year study with 100,000 participants
- Site distribution is varied from uniform (table 3a) to Mini-Sentinel like (table 3b)
- Odds ratios for binomial, continuous and site confounders set to 2 for both exposure and outcome

Main Conclusions:

- All three methods appear to approximately hold type I error under the uniform site distribution
- Type I error is inflated under the MS-like site distribution where prevalence of exposure is 50%, though this does not occur when prevalence of exposure is 25%
- Stratified IPTW estimator performs similarly to pooled data estimators
- Power increases slightly as prevalence of exposure increases
- Detection limit is large for achieving acceptable power, RR between 2 and 3 for 80-90% power even when sample size is at 100,000.

5. Interaction with exposure by site

Results for this section are provided in Table 4, Section V11.
Situation:

- One-time analysis with comparing methods across probability of outcome of 1% and 5%, prevalence of exposure of 50% and 25% and both uniform and MS-like site distributions for a two-year study with 10,000 participants.
- In the model for exposure, the effect of the continuous confounder (age) varies by site. The odds ratios of exposure for sites 1, 2 and 3 are 0.5, 1.0 and 2.0, respectively, for a 10 year increase in age
- Odds ratios for binomial confounder in the exposure model is set to 2

- Odds ratios for the binomial, continuous and site variables in the outcome model are set to 2
- Three models were tested: 1) The IPTW estimator for data pooled across sites using the correct weighting model which includes an interaction between age and site; 2) the IPTW estimator for the pooled data where the weighting model is misspecified by ignoring the interaction between age and site; and 3) the stratified IPTW estimator which using site-specific weighting models and thus should mimic the correctly specified model for the pooled data
- Effect sizes between a relative risk of 1 and 2 are explored, with the corresponding true risk differences given in table 4

Main Conclusions:

- Type I error is inflated for the misspecified model in all cases. Especially as prevalence increases to 0.05 with type error as high as 15%
- Stratified IPTW estimator performs similarly to the pooled estimator with the correct weighting model

B. GROUP SEQUENTIAL STUDY ANALYSIS

1. Data Structure

In the case of sequential analyses, the data structure is exactly as described in Section V.A.1 with the addition of multiple analyses over time. In this case we consider performing an analysis at each look, *l* for *l*=1, … , *L*. The looks occur at time t_l using a subset of the data such that $D_i < t_l$, i.e., each participant included in the study at look *l* was exposed to the exposure of interest prior to time t_l.

2. Evenly sized sites

Situation:

- A two-year group sequential study with 10,000 total participants where the first look occurs at 180 days with quarterly looks thereafter for a total of 7 looks
- Sample is equally distributed across three sites, i.e. *p*=(1/3, 1/3, 1/3)
- Compares methods across probability of the outcome of 1% and 5%
- Odds ratios of exposure for binary and continuous confounders set at 2 for all simulations, while the odds ratio of exposure for two sites relative to the third varies over the values 0.5, 1.0 and 2.0
- Odds ratios for outcome relative to binary, continuous and site confounders set at 2 for all simulations
- Evaluation metrics are power, average time to detection of a signal and average time to study end

a. Prevalence of Exposure is 50%

Results for this section are provided in Table 5a, Section V11.
Main Conclusions:

- Type I error is slightly inflated for all estimators in the case of 1% incidence of outcome, but particularly so for the IPTW estimators when site confounding is present

- Power may still be slightly inflated for the IPTW estimators as a relative risk of 1.25 in the case of 1% incidence, but power appears to converge for all three estimators at a relative risk of 1.5. In this case, the IPTW estimators have slightly lower average time to detection, 7-14 days shorter, and study end than the GLM estimator
- In the case of 5% incidence of outcome all estimators approximately hold the type I error, the IPTW estimators tend to have slightly lower power, 1.7%-2.9% difference, at moderate effect sizes (RR=1.25), but times to study end are similar across the three estimators
- As probability of outcome increases, power increases
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites

b. Prevalence of Exposure is 0.25

Results for this section are provided in Table 5b, Section V11.
Main Conclusions:

- Type I error is slightly deflated for all estimators in the case of 1% incidence of outcome
- Power is comparable across the three estimators, but time to study end tends to be shorter for the IPTW methods in the 1% prevalence case
- In the case of 5% incidence of outcome all estimators approximately hold the type I error for the case of 0.5 and no site confounding but when site confounding is 2, the GLM estimator shows signs of inflation and the IPTW estimators show signs of deflation. This trend continues for the power at a RR of 1.25 which makes interpretation of differences time to detection and study end difficult
- As probability of outcome increases, power increases and power is slightly lower compared to the case of 50% probability of exposure

3. Mini-Sentinel like site distribution

Situation:

- A two-year group sequential study with 10,000 total participants where the first look occurs at 180 days with quarterly looks thereafter for a total of 7 looks
- 10% of study sample comes from one site and the remaining 90% are split evenly across the other two sites, i.e. *p*=(0.10, 0.45, 0.45)
- Compares methods across probability of the outcome of 1% and 5%
- Odds ratios of exposure for binary and continuous confounders set at 2 for all simulations, while the odds ratio of exposure for two sites relative to the third varies over the values 0.5, 1.0 and 2.0
- Odds ratios for outcome relative to binary, continuous and site confounders set at 2 for all simulations
- Evaluation metrics are power, average time to detection of a signal and average time to study end

a. Prevalence of Exposure is 0.50

Results for this section are provided in Table 6a, Section V11.
Main Conclusions:

- Type I error is moderately inflated for all estimators in the case of 1% incidence of outcome, particularly when there is no site confounding or the OR of exposure for site is set at 2
- Type I error for all estimators tends to be less inflated in the case of 5% incidence, which is again more prominent when the OR for exposure due to site is set at 2 versus 0.5
- Time to detection and study end is comparable in all instances though variability in power makes minor differences between estimators difficult to interpret
- As probability of outcome increases, power increases
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites

b. Prevalence of Exposure is 0.25

Results for this section are provided in Table 6b, Section V11.
Main Conclusions:

- Type I error is moderately deflated for all estimators in the case of 1% incidence of outcome
- Type I error for GLM estimator is slightly inflated in the case of 5% incidence of outcome
- Time to detection and study end is comparable in all instances though variability in power makes minor differences between estimators difficult to interpret
- As probability of outcome increases, power increases, though power tends to be lower than in the case of 50% exposure, especially for the IPTW estimators
- Stratified IPTW estimator performs similarly to estimators using data pooled across sites

VI. DISCUSSION AND FUTURE WORK

This task order has developed a new statistical method, stratified IPTW estimation using propensity scores, for conducting postmarket safety studies in a distributed data setting. This method can be used to estimate and test for differences in risk in both one-time and sequential studies. We have specifically focused on developing methods that estimate a risk difference since it is generally the key quantity of interest for informing important policy decisions. Using simulation studies we have assessed the performance of these methods relative to standard regression techniques applied to pooled (non-distributed) data. The simulation results show that stratified IPTW methods perform comparably to pooled estimators in the rare event setting when confounding is present. Moreover, the proposed extension to group sequential monitoring also proved to be comparable to standard methods with non-distributed data. Therefore, our overall conclusion is that the stratified IPTW methods are a viable approach to estimating risk differences in the postmarket surveillance safety setting when data is distributed among many sites. Future work comparing this IPTW stratified approach to propensity score matching and propensity score stratification still need to be explored. Extensions to relative risks and odd ratio stratified regression approaches need to be developed and evaluated when the quantity of interest is not a relative risk.

A. TASK ORDER DELIVERABLES

- Mini-Sentinel Report including non-statistical write-up of methods, statistical write-up of methods, and results of the simulation evaluation
- Code: We have created R code to run all of the IPTW methods including non-distributed one-time analysis, distributed one-time analysis, non-distributed group sequential, and distributed

group sequential. The code has been given to the FDA, but is also available at http://faculty.washington.edu/acook/software.html

VII. TABLES AND FIGURES FOR SIMULATION STUDY

Table 1a. Power across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Single*; Probability of Exposure*: 50%; *Outcome confounding strength:* OR of 2; *Site distribution*: Uniform

Effect Size			Confounders				Power					
	RD		OR (X|Z,S)				P(Y)=1%			P(Y)=5%		
RR	P(Y)=1%	P(Y)=5%	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1.00	0.00	0.00	2	2	0.5	0.5	0.059	0.058	0.058	0.049	0.039	0.039
			2	2	1	1	0.052	0.054	0.048	0.049	0.049	0.046
			2	2	2	2	0.042	0.050	0.048	0.051	0.049	0.052
1.25	0.0025	0.0125	2	2	0.5	0.5	0.298	0.303	0.305	0.848	0.817	0.816
			2	2	1	1	0.309	0.310	0.300	0.798	0.797	0.796
			2	2	2	2	0.283	0.304	0.293	0.799	0.791	0.795
1.50	0.005	0.025	2	2	0.5	0.5	0.687	0.678	0.677	0.999	0.998	0.999
			2	2	1	1	0.683	0.677	0.680	1.000	1.000	1.000
			2	2	2	2	0.688	0.696	0.686	1.000	1.000	0.999

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 1b. Power across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Single*; Probability of Exposure*: 25%; *Outcome confounding strength:* OR of 2; *Site distribution*: Uniform

Effect Size			Confounders				Power					
	RD		OR (X|Z,S)				P(Y)=1%			P(Y)=5%		
RR	P(Y)=1%	P(Y)=5%	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1.00	0.00	0.00	2	2	0.5	0.5	0.045	0.049	0.046	0.060	0.053	0.057
			2	2	1	1	0.052	0.049	0.044	0.061	0.043	0.040
			2	2	2	2	0.050	0.042	0.048	**0.090**	0.041	0.048
1.25	0.0025	0.0125	2	2	0.5	0.5	0.286	0.269	0.276	0.796	0.734	0.725
			2	2	1	1	0.319	0.293	0.287	0.855	0.788	0.786
			2	2	2	2	0.334	0.275	0.277	0.901	0.775	0.782
1.50	0.005	0.025	2	2	0.5	0.5	0.659	0.627	0.646	0.997	0.992	0.994
			2	2	1	1	0.686	0.649	0.647	0.998	0.997	0.998
			2	2	2	2	0.734	0.648	0.648	0.999	0.995	0.997

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 2a. Power across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Single*; Probability of Exposure*: 50%; *Outcome confounding strength:* OR of 2; *Site distribution*: Mini-Sentinel

Effect Size			**Confounders**				**Power**					
	RD		OR (X|Z,S)				P(Y)=1%			P(Y)=5%		
RR	P(Y)=1%	P(Y)=5%	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1.00	0.00	0.00	2	2	0.5	0.5	0.063	**0.067**	0.064	0.059	0.054	0.059
			2	2	1	1	0.049	0.051	0.053	0.056	0.056	0.056
			2	2	2	2	0.043	0.043	0.042	0.039	0.038	0.040
1.25	0.0025	0.0125	2	2	0.5	0.5	0.314	0.308	0.304	0.825	0.809	0.816
			2	2	1	1	0.322	0.326	0.323	0.834	0.829	0.829
			2	2	2	2	0.301	0.302	0.308	0.803	0.790	0.791
1.50	0.005	0.025	2	2	0.5	0.5	0.700	0.707	0.702	1.000	1.000	1.000
			2	2	1	1	0.680	0.678	0.675	1.000	1.000	0.999
			2	2	2	2	0.705	0.718	0.703	1.000	1.000	1.000

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 2b. Power across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Single*; Probability of Exposure*: 25%; *Outcome confounding strength:* OR of 2; *Site distribution*: Mini-Sentinel

Effect Size			**Confounders**				**Power**					
	RD		OR (X|Z,S)				P(Y)=1%			P(Y)=5%		
RR	P(Y)=1%	P(Y)=5%	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1.00	0.00	0.00	2	2	0.5	0.5	0.037	**0.032**	**0.034**	**0.067**	0.046	0.051
			2	2	1	1	0.047	0.046	0.046	**0.068**	**0.033**	**0.032**
			2	2	2	2	0.058	0.055	0.049	**0.072**	0.047	0.048
1.25	0.0025	0.0125	2	2	0.5	0.5	0.296	0.274	0.275	0.815	0.754	0.747
			2	2	1	1	0.297	0.271	0.265	0.853	0.757	0.760
			2	2	2	2	0.298	0.270	0.271	0.859	0.748	0.753
1.50	0.005	0.025	2	2	0.5	0.5	0.652	0.622	0.620	0.999	0.999	0.999
			2	2	1	1	0.701	0.657	0.654	0.999	0.996	0.996
			2	2	2	2	0.686	0.639	0.635	0.999	0.996	0.996

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 3a. Power for very rare event setting (P(Y)=0.02%) comparing estimators across varying prevalence of exposure case of a binary (sex), continuous (age) and categorical confounder (site) with a binary outcome.
Analysis type: Single*; Outcome confounding strength:* OR of 2; *Site distribution*: Uniform

Effect Size			Confounders				Power					
			OR (X|Z,S)				P(X)=50%			P(X)=25%		
P(Y)	RD	RR	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
0.0002	0.0000	1.0	2	2	2	2	0.054	0.063	0.064	0.045	0.057	0.058
0.0002	0.0002	2.0	2	2	2	2	0.571	0.574	0.564	0.492	0.508	0.496
0.0002	0.0004	3.0	2	2	2	2	0.931	0.926	0.924	0.924	0.905	0.901
0.0002	0.0006	4.0	2	2	2	2	0.993	0.993	0.990	0.992	0.992	0.992
0.0002	0.0008	5.0	2	2	2	2	1.000	1.000	1.000	1.000	1.000	1.000

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 3b. Power for very rare event setting (P(Y)=0.02%) comparing estimators across varying prevalence of exposure case of a binary (sex), continuous (age) and categorical confounder (site) with a binary outcome.
Analysis type: Single*; Outcome confounding strength:* OR of 2; *Site distribution*: Mini-Sentinel

Effect Size			Confounders				Power					
			OR (X|Z,S)				P(X)=50%			P(X)=25%		
P(Y)	RD	RR	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
0.0002	0.0000	1.0	2	2	2	2	**0.081**	**0.089**	**0.088**	**0.044**	0.052	0.054
0.0002	0.0002	2.0	2	2	2	2	0.547	0.557	0.562	0.520	0.545	0.532
0.0002	0.0004	3.0	2	2	2	2	0.935	0.933	0.932	0.912	0.911	0.918
0.0002	0.0006	4.0	2	2	2	2	0.998	0.993	0.994	0.992	0.991	0.988
0.0002	0.0008	5.0	2	2	2	2	1.000	1.000	1.000	0.999	0.999	0.999

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05
Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 4. Power comparing IPTW estimators with correct specification of weighting model using pooled data, misspecified weighting model using pooled data and stratified weighting (IPTW_s) for a setting where a continuous confounder (age) interacts with site relative to exposure and probability of outcome, probability of exposure and site distribution are varied.
Analysis type: Single*; Outcome confounding strength:* OR of 2 for binary (sex), continuous (age) and categorical (site) variables;
Exposure confounding strength: OR of 2 for binary (sex) variable, ORs of 0.5, 1.0 and 2.0 for continuous variable (age) at sites 1, 2 and 3

P(Y)	**P(X)**	**Effect Size**		**Power**			**Power**		
				Uniform Site Distribution			MS Like Site Distribution		
		RD	RR	Correct Model	Misspecified Model	IPTW_S	Correct Model	Misspecified Model	IPTW_S
0.01	0.5	0.000	1.0	0.054	**0.096**	0.056	0.069	**0.081**	0.067
		0.001	1.1	0.112	0.175	0.113	0.133	0.168	0.136
		0.003	1.3	0.406	0.539	0.403	0.410	0.464	0.412
		0.005	1.5	0.715	0.827	0.716	0.731	0.779	0.723
		0.010	2.0	0.990	0.997	0.990	0.993	0.994	0.992
	0.25	0.000	1.0	0.053	**0.094**	0.051	0.064	**0.082**	0.067
		0.001	1.1	0.104	0.173	0.109	0.114	0.131	0.116
		0.003	1.3	0.316	0.449	0.310	0.353	0.401	0.347
		0.005	1.5	0.613	0.777	0.614	0.638	0.686	0.638
		0.010	2.0	0.964	0.994	0.968	0.979	0.985	0.981
0.05	0.5	0.000	1.0	0.054	**0.157**	0.054	0.052	**0.073**	0.048
		0.005	1.1	0.268	0.543	0.267	0.296	0.389	0.294
		0.015	1.3	0.929	0.988	0.925	0.939	0.971	0.941
		0.025	1.5	0.999	1.000	0.999	1.000	1.000	1.000
	0.25	0.000	1.0	0.043	**0.153**	0.046	0.045	0.064	0.046
		0.005	1.1	0.235	0.481	0.238	0.234	0.297	0.230
		0.015	1.3	0.860	0.963	0.854	0.852	0.911	0.863
		0.025	1.5	0.993	0.999	0.994	0.999	1.000	0.999

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05

Abbreviations: X=exposure; Y=outcome event; RR=relative risk; RD=risk difference; OR=odds ratio; Correct model=IPTW estimator with interaction included in weighting model using pooled data; Misspecified model=IPTW estimator where weighting model ignores interaction using pooled data; IPTW_s=stratified IPTW estimator

Table 5a. Power, days to rejection, and days to study end across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.

Analysis type: Sequential; *Frequency:* First look at 180 days, quarterly thereafter; *Duration:* Two years; *Probability of Exposure*: 50%; *Outcome confounding strength:* OR of 2 for all variables; *Site distribution*: Uniform

P(Y)	Effect Size		Confounders				Power			Time to Detection			Time to Study End		
			OR (X|Z,S)												
	RR	RD	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1%	1.00	0.00	2	2	0.5	0.5	0.061	**0.065**	**0.069**	357.0	336.5	341.7	697.9	695.1	693.9
			2	2	1	1	0.058	0.064	0.063	339.8	312.2	311.4	698.0	693.9	694.3
			2	2	2	2	0.059	**0.075**	**0.071**	334.1	343.2	339.7	697.2	691.7	693.0
	1.25	0.0025	2	2	0.5	0.5	0.253	0.267	0.277	410.2	392.0	394.8	641.6	632.4	629.9
			2	2	1	1	0.255	0.270	0.266	398.1	378.7	375.2	637.9	627.8	628.3
			2	2	2	2	0.255	0.277	0.272	396.7	385.3	396.1	637.6	627.3	631.9
	1.50	0.005	2	2	0.5	0.5	0.598	0.594	0.603	399.0	392.0	392.8	528.0	525.2	522.7
			2	2	1	1	0.595	0.598	0.604	390.6	377.6	381.0	524.0	515.3	515.3
			2	2	2	2	0.612	0.608	0.596	399.1	384.7	386.3	523.6	516.2	521.1
5%	1.00	0.00	2	2	0.5	0.5	0.051	0.045	0.050	347.6	326.0	345.6	701.0	702.3	701.3
			2	2	1	1	0.055	0.056	0.060	337.1	343.9	331.5	698.9	698.9	696.7
			2	2	2	2	0.056	0.054	0.054	329.5	328.3	330.0	698.1	698.9	698.9
	1.25	0.0125	2	2	0.5	0.5	0.770	0.741	0.743	389.0	385.9	383.1	465.1	472.4	469.7
			2	2	1	1	0.728	0.718	0.711	377.7	375.0	372.5	470.8	472.3	473.0
			2	2	2	2	0.743	0.719	0.717	396.5	388.0	388.6	479.6	481.3	482.4
	1.50	0.025	2	2	0.5	0.5	0.998	0.998	0.998	235.6	242.2	242.5	236.5	243.2	243.5
			2	2	1	1	0.999	0.999	1.000	240.0	244.1	243.9	240.5	244.6	243.9
			2	2	2	2	0.998	0.996	0.995	239.6	243.5	242.9	240.6	245.4	245.3

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05

Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 5b. Power, days to rejection, and days to study end across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Sequential; *Frequency:* First look at 180 days, quarterly thereafter; *Duration:* Two years; *Probability of Exposure*: 25%; *Outcome confounding strength:* OR of 2 for all variables; *Site distribution*: Uniform

P(Y)	Effect Size		Confounders				Power			Time to Detection			Time to Study End		
			OR (X|Z,S)				GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
	RR	RD	Z1	Z2	S1	S2									
1%	1.00	0.00	2	2	0.5	0.5	**0.032**	0.038	0.040	388.1	374.2	389.3	709.4	706.9	706.8
			2	2	1	1	0.044	0.045	0.046	405.0	394.0	383.5	706.1	705.3	704.5
			2	2	2	2	0.039	**0.032**	**0.034**	396.9	351.6	367.9	707.4	708.2	708.0
	1.25	0.0025	2	2	0.5	0.5	0.190	0.212	0.227	436.3	423.3	427.0	666.1	657.1	653.5
			2	2	1	1	0.225	0.238	0.240	462.4	446.6	446.3	662.0	654.9	654.3
			2	2	2	2	0.240	0.218	0.223	437.3	422.3	414.9	652.1	655.1	652.0
	1.50	0.005	2	2	0.5	0.5	0.537	0.564	0.580	425.5	421.4	418.0	561.9	551.6	544.9
			2	2	1	1	0.583	0.571	0.571	429.5	417.8	420.1	550.6	547.5	548.7
			2	2	2	2	0.612	0.585	0.579	418.1	406.0	408.0	535.2	536.3	539.4
5%	1.00	0.00	2	2	0.5	0.5	0.057	0.043	0.050	391.6	380.9	397.8	701.3	705.4	703.9
			2	2	1	1	0.058	0.046	0.048	374.0	367.8	403.1	699.9	703.8	704.8
			2	2	2	2	**0.068**	**0.031**	**0.035**	398.4	397.7	385.7	698.1	710.0	708.3
	1.25	0.0125	2	2	0.5	0.5	0.701	0.644	0.641	392.4	400.8	399.3	490.3	514.4	514.4
			2	2	1	1	0.780	0.696	0.696	387.3	397.8	401.9	460.5	495.7	498.6
			2	2	2	2	0.825	0.698	0.703	368.9	386.8	386.1	430.4	487.4	485.3
	1.50	0.025	2	2	0.5	0.5	0.995	0.991	0.991	254.9	272.9	272.0	257.2	276.9	276.0
			2	2	1	1	0.999	0.996	0.996	243.6	261.0	262.4	244.1	262.8	264.2
			2	2	2	2	1.000	0.997	0.996	228.7	257.1	255.6	228.7	258.5	257.5

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05

Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 6a. Power, days to rejection, and days to study end across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Sequential; *Frequency:* First look at 180 days, quarterly thereafter; *Duration:* Two years; *Probability of Exposure*: 50%; *Outcome confounding strength:* OR of 2 for all variables; *Site distribution*: Mini-Sentinel

P(Y)	**Effect Size**		**Confounders**				**Power**			**Time to Detection**			**Time to Study End**		
			OR (X|Z,S)												
	RR	RD	Z1	Z2	S1	S2	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
1%	1.00	0.00	2	2	0.5	0.5	0.055	0.059	0.061	353.5	343.2	346.7	699.8	697.8	697.2
			2	2	1	1	**0.067**	**0.075**	**0.078**	349.3	331.2	339.2	695.2	690.8	690.3
			2	2	2	2	**0.071**	**0.080**	**0.070**	377.7	367.9	367.7	695.7	691.8	695.3
	1.25	0.0025	2	2	0.5	0.5	0.249	0.244	0.248	392.5	373.6	376.0	638.5	635.5	634.7
			2	2	1	1	0.247	0.256	0.255	383.7	380.7	377.6	636.9	633.2	632.7
			2	2	2	2	0.243	0.253	0.250	403.7	384.5	384.1	643.1	635.1	636.0
	1.50	0.005	2	2	0.5	0.5	0.616	0.608	0.613	405.3	399.7	395.4	526.1	525.2	521.0
			2	2	1	1	0.613	0.606	0.606	394.5	384.7	385.8	520.5	516.8	517.5
			2	2	2	2	0.619	0.623	0.625	398.7	386.1	390.0	521.1	512.0	513.7
5%	1.00	0.00	2	2	0.5	0.5	0.052	0.059	0.054	360.0	376.8	365.0	701.3	699.8	700.8
			2	2	1	1	0.058	0.062	0.060	344.5	345.5	360.0	698.2	696.8	698.4
			2	2	2	2	**0.070**	**0.065**	0.064	378.0	386.3	398.0	696.1	698.3	699.4
	1.25	0.0125	2	2	0.5	0.5	0.772	0.754	0.752	381.0	384.2	375.2	458.3	466.8	460.7
			2	2	1	1	0.724	0.722	0.716	385.4	384.8	384.3	477.7	478.0	479.6
			2	2	2	2	0.741	0.717	0.717	380.0	374.4	375.9	468.1	472.2	473.3
	1.50	0.025	2	2	0.5	0.5	0.998	0.998	0.996	238.7	243.9	243.5	239.7	244.9	245.4
			2	2	1	1	0.999	0.999	0.999	240.5	244.6	245.4	241.0	245.1	245.9
			2	2	2	2	0.999	0.998	0.998	238.2	243.5	243.1	238.7	244.4	244.1

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05

Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

Table 6b. Power, days to rejection, and days to study end across varying strength of confounding and probability of outcome for the case of a binary (sex), a continuous (age) and a categorical confounder (site) with a binary outcome.
Analysis type: Sequential; *Frequency:* First look at 180 days, quarterly thereafter; *Duration:* Two years; *Probability of Exposure*: 25%; *Outcome confounding strength:* OR of 2 for all variables; *Site distribution*: Mini-Sentinel

P(Y)	**Effect Size**		**Confounders**				**Power**			**Time to Detection**			**Time to Study End**		
			OR (X|Z,S)				GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S	GLM	IPTW	IPTW_S
	RR	RD	Z1	Z2	S1	S2									
1%	1.00	0.00	2	2	0.5	0.5	**0.035**	0.041	0.046	447.4	386.3	387.4	710.5	706.3	704.7
			2	2	1	1	0.041	0.049	0.047	443.4	404.1	404.0	708.7	704.5	705.2
			2	2	2	2	0.040	0.048	0.041	373.5	375.0	384.1	706.1	703.4	706.2
	1.25	0.0025	2	2	0.5	0.5	0.215	0.228	0.231	439.5	432.6	428.2	659.7	654.5	652.6
			2	2	1	1	0.222	0.224	0.232	441.5	427.9	423.6	658.2	654.6	651.2
			2	2	2	2	0.240	0.237	0.231	435.0	425.7	422.7	651.6	650.3	651.3
	1.50	0.005	2	2	0.5	0.5	0.558	0.574	0.576	413.2	402.6	402.3	548.8	537.8	537.0
			2	2	1	1	0.576	0.560	0.560	442.7	418.7	417.5	560.3	551.3	550.6
			2	2	2	2	0.576	0.559	0.557	426.1	422.3	418.7	550.7	553.6	552.2
5%	1.00	0.00	2	2	0.5	0.5	0.052	**0.034**	0.038	394.6	420.9	419.2	703.1	709.8	708.6
			2	2	1	1	**0.078**	0.046	0.044	455.8	428.5	425.5	699.4	706.6	707.0
			2	2	2	2	0.064	0.051	0.051	392.3	391.8	397.1	699.0	703.3	703.5
	1.25	0.0125	2	2	0.5	0.5	0.747	0.679	0.674	391.3	399.8	401.1	474.5	502.6	505.1
			2	2	1	1	0.779	0.674	0.672	392.5	404.5	405.7	464.9	507.3	508.8
			2	2	2	2	0.780	0.668	0.671	378.0	391.3	391.3	453.2	500.4	499.4
	1.50	0.025	2	2	0.5	0.5	0.999	0.992	0.994	245.4	262.9	264.4	245.9	266.6	267.1
			2	2	1	1	0.998	0.992	0.991	245.1	261.5	262.1	246.1	265.1	266.2
			2	2	2	2	0.999	0.995	0.996	240.2	265.0	264.8	240.7	267.3	266.6

* Bold indicates outside +/- 1.5% of the expected type I error of 0.05

Abbreviations: X=exposure; Y=outcome event; Z=confounders; S=site; RR=relative risk; RD=risk difference; OR=odds ratio; GLM=generalized linear model with adjustment for confounding using data pooled across sites; IPTW=inverse probability weighted estimator using data pooled across sites; IPTW_s=stratified inverse probability weighted estimator

VIII. REFERENCES

1. Austin PC. Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. International Journal of Biostatistics 2009;5.
2. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine 1998;17:2265-81.
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. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550-60.
5. 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.
6. Li L, Kulldorff M, Nelson J, Cook A. A Propensity Score-Enhanced Sequential Analytic Method for Comparative Drug Safety Surveillance. 2011:45-62.
7. 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.
8. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology 2008;168:656-64.
9. Hernan MA, Alonso A, Logan R, et al. Observational Studies Analyzed Like Randomized Experiments An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. Epidemiology 2008;19:766-79.
10. Lumley T. Complex Surveys: A Guide to Analysis Using R. Hoboken, New Jersey: John Wiley and Sons; 2010.
11. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken, NJ: John Wiley and Sons Inc; 2002.
12. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Statistics in Medicine 2010;29:2137-48.
13. Forbes A, Shortreed S. Inverse probability weighted estimation of the marginal odds ratio: correspondence regarding 'The performance of different propensity score methods for estimating marginal odds ratios'. Stat Med 2008;27:5556-9; author reply 60-3.
14. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007;26:3078-94.
15. Savu A, Liu Q, Yasui Y. Estimation of relative risk and prevalence ratio. Stat Med 2010;29:2269-81.
16. Rotnitzky A, Jewell NP. Hypothesis-testing of regression parameters in semiparametric generalized linear-models for cluster correlated data. Biometrika 1990;77:485-97.
17. Lan KKG, Demets DL. Discrete Sequential Boundaries for Clinical-Trials. Biometrika 1983;70:659-63.
18. 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. In Preparation 2011.
19. 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.

20. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics 2011;127 Suppl 1:S54-64.

21. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. Stat Med 2007;26:5047-80.

22. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics 1999;55:874-82.