Sequential safety monitoring using observational data: A comparison of methods appropriate for newly-licensed vaccines in children

Kelly G. Stratton

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2012

Committee:
Jennifer C. Nelson
Andrea J. Cook

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Sequential safety monitoring using observational data:
A comparison of methods appropriate for
newly-licensed vaccines in children

Kelly G. Stratton

Chair of the Supervisory Committee:

Sequential safety monitoring of newly-licensed vaccines is a national health priority and is routinely conducted using observational data. However, applying sequential methods in observational settings where the adverse events (AEs) being monitored are rare is relatively new and much is unknown about method performance. In this thesis we use simulations and an example application to compare four existing group sequential (GS) methods: two that use regression to control for confounding (GS Generalized Estimating Equations and GS Lan-DeMets-Regression) and two that use one-to-one matching to control for confounding (GS Likelihood Ratio Test and GS Lan-DeMets-Matching). We simulated five sites and varied the amount of confounding by site, sample sizes of the sites, and prevalence of the AE under surveillance. We also applied the methods to data from a recent Vaccine Safety Datalink study. In the simulations, the matched methods were less powerful, were slower to detect true safety signals, and experienced more implementation difficulties with rare AEs compared to the regression methods. Across different confounding by site scenarios, differences in power to detect a safety signal depend on how evenly the AEs were distributed across sites, the amount of statistical information at each site, and the direction of the relationships between site and exposure or between site and AEs. In the data application, both regression methods successfully detected a safety signal under a variety of testing frequencies, while neither matched method detected a safety signal for any of the testing fre-

quency options we used. The differences in power and time-to-surveillance-end between the matched and regression methods are largely explained by the reduction in data contributing to the test statistic for the matched methods (i.e., reduction to discordant matched pairs). Additionally, lack of newly accrued statistical information between analyses required us to skip some analysis times for the matched methods. Our results indicate that the choice of sequential method, particularly the choice of strategy for confounder control, is critical in rare AE observational safety monitoring settings, and further study is needed to better understand and optimize method performance.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to all of the faculty, staff, and students in the Biostatistics Department. I hope to continue to be surrounded by such intelligent, yet down-to-earth and friendly people. I am particularly grateful to my thesis advisers, Jennifer Nelson and Andrea Cook, both of whom provided immense guidance, mentorship, encouragement, and enthusiasm over the past year. Our conversations always left me excited to get back to work.

DEDICATION

To my husband, Casey, for all of his support over the past three years and for putting up with a sometimes very stressed wife.

To my parents, Paul and Julie, for their encouragement. To my siblings, Rhea, David, and Laura, for making me smile all the times I needed to be reminded of life outside of graduate school.

To Irene for listening. To Jessica for understanding my mental states without me having to explain.

Chapter 1

INTRODUCTION

1.1 Importance and setting

Rapid post-licensure safety surveillance has become a national public health priority [1], and safety monitoring of newly-licensed vaccines is now routinely conducted as part of the Centers for Disease Control and Prevention's (CDC's) multi-site Vaccine Safety Datalink (VSD) project [15, 21] as well as the Food and Drug Administration's (FDA's) Mini-Sentinel pilot [10]. This type of surveillance aims to quickly identify or rule out what is often called a "safety signal," a quantitative indication that recipients of a newly-licensed vaccine have an increased risk of an adverse event (AE) compared to those who receive an alternative or no vaccine. Such monitoring within the VSD has prompted changes to national vaccine policy [30] as well as reassured the public of the safety of particular vaccines [25, 36, 50, 11, 26, 33]. To accomplish rapid detection of safety signals, sequential monitoring methods are used so that statistical analyses can be performed as new vaccine uptake occurs, instead of waiting until a pre-specified end of study time point to analyze the data all at once. Post-licensure safety surveillance is an important component of a vaccine's development because, while some safety information is known about a vaccine from clinical trials prior to its licensure, these trials are limited with respect to the number and diversity of participants, as well as the duration of follow-up.

In this thesis we compare a selection of sequential statistical methods that were either designed for prospective observational vaccine safety surveillance or have been used for sequential monitoring in clinical trials. The general setting for this work is sequential safety monitoring of newly-licensed vaccines, where the data to be analyzed derive from electronic health care information that is collected by managed care organizations during routine

patient medical care visits. There are several key features of this setting to note. First, vaccine exposure is limited to a single instance, as opposed to medications that are taken daily or interventions such as pacemakers, where exposure is chronic. Methods appropriate for safety surveillance in chronic exposure settings are available but not described here. Second, surveillance is performed in order to detect AEs that are both rare and occur acutely after vaccination. Third, the data are captured from multiple sites, and patient privacy is protected by using a distributed data format in which only summarized data (as opposed to individual-level data) from each site are combined centrally for analysis. Fourth, the rate of uptake of the newly-licensed vaccine generally increases over time, and at the beginning of monitoring the number of subjects exposed to the new vaccine can be small. Last, the data used for surveillance are observational in nature. Specifically, they are routinely collected for clinical and billing purposes by health care providers and, as such, not based on an experiment where subjects are randomized to receive the new vaccine or not. This lack of randomization makes necessary the selection of a comparator group that minimizes the influence of confounding. Options for comparator groups that have been used in practice include historical, self, and concurrent controls. Each choice of comparator has advantages and limitations, requires different methods, and deserves study. In this thesis, we consider methods designed for concurrent comparators who receive a similar vaccine (that is already widely used) during the same time period in which uptake of the new vaccine occurs. Concurrent comparators have been used in both the FDA's Mini Sentinel pilot and in the CDC's VSD safety monitoring.

1.2 Thesis aims

Applying sequential methods in observational settings where outcomes are rare is relatively new. Hence, there is much to learn about how the available methods perform in this setting, particularly with regard to incorporating confounder control. This work aims to add to knowledge about currently-available methods for sequential observational safety monitoring for rare AEs and increase awareness of their strengths and limitations by conducting a simulation study to quantify empirical information about their statistical performance characteristics. Specifically, we:

1. Compute and compare type I error rates, power, and the time-to-surveillance-end between sequential methods that use either regression-adjustment or one-to-one exposure matching to control for confounding.

2. Explore the impact that confounding by health care site has on these performance characteristics by varying the:

   (a) Sample size at each health care site.

   (b) Strength of confounding by health care site.

3. Investigate the following practical issues related to method implementation when monitoring for rare AEs:

   (a) Performing sequential tests when little statistical information is available.

   (b) Matching when AE prevalence is low.

We begin in Chapter 2 by providing a brief overview of sequential monitoring methods used both in randomized clinical trials (RCTs) and in observational settings. In Chapter 3 we detail the selected sequential monitoring methods implemented in this thesis, which use either a likelihood ratio test (LRT), a generalized estimating equations (GEE) approach with a score test, or a standardized normal test statistic compared to a sequential stopping boundary. We describe, in Chapters 4 and 5, respectively, the methods and results of the simulations performed to compare these approaches. In Chapter 5 we also apply the methods to a subset of observational safety monitoring data from the VSD's surveillance of a combination diphtheria, tetanus toxoids, acellular pertussis, inactivated poliovirus, and Haemophilus b conjugate (DTaP-IPV-Hib) vaccine [25]. Finally, in Chapter 6 we discuss the performance of the methods in the simulations and application, the impact of confounding by site, the limitations of our approach to evaluate these methods, and potential areas for future work.

Chapter 2

OVERVIEW OF SEQUENTIAL MONITORING METHODS

2.1 Sequential monitoring

Sequential monitoring has been used for decades in RCTs as a way to monitor the results of an intervention as subjects, and therefore data, accrue. This allows investigators to stop the trial early if, prior to the pre-designated end of the study, there is sufficient confidence that the intervention is harmful (in which case continuing to administer the intervention would be unethical), or that the intervention is beneficial (in which case not making the treatment widely available as quickly as possible would be unethical). In addition to the ethical motivations for using sequential monitoring, the expected sample size is smaller with a sequential monitoring procedure than without, so there is an increase in the cost-effectiveness of trials; less time and money will be dedicated to a trial if it is stopped early. Sequential monitoring can either be performed after a group of subjects accrues (group sequential monitoring) or after each individual subject accrues (continuous sequential monitoring). In RCTs, a group sequential monitoring procedure with one or two interim analyses and a final analysis at the end of the trial is common. There is an extensive body of literature as well as comprehensive reviews that describe sequential monitoring methods in RCTs [3, 28, 49, 23, 24]. Here, we describe the basic use and statistical implications of group sequential monitoring in RCTs, discuss additional components that must be considered when using them or adapting them for use in the observational safety setting, and briefly review the selected methods that are relevant to this thesis.

2.1.1 Repeated testing

When the same statistical hypothesis is tested repeatedly, a correction for multiple comparisons must be done in order to maintain the overall type I error rate at the desired level. Consider a scenario where the same hypothesis test is performed T pre-specified times, with

more data being accrued over time. To maintain the overall type I error rate we could simply adjust the critical values at each test with a Bonferroni correction or other adjustment for multiple comparisons. Now consider sequential testing, which can be viewed as a special case of repeated testing. Again we pre-specify T analysis times. In this scenario, however, we will cancel all remaining tests if, at one of the intermediate analyses, we find sufficient evidence either for or against the intervention. Here, simply adjusting for multiple comparisons will not maintain the overall type I error rate because in order to perform the second test, we must fail to reject the null hypothesis at the first test. In order to perform the third test, we must fail to reject the null hypothesis at the first and second tests, and so on. In other words, this multiple testing approach is conditional on not having stopped the trial at the previous analyses. Conditioning in this way causes the distribution of the test statistic to be skewed, and this skewness must be accounted for when computing the rejection boundary in order to correctly control the overall type I error rate.

2.1.2 Specifying a stopping boundary

For each test in a sequentially-monitored study we compare the observed test statistic, which is computed based on all data accrued up to each analysis time, to a rejection boundary, called the stopping boundary. Selection of a stopping boundary is guided by scientific, ethical, and statistical criteria, and involves specifying the frequency of testing, the boundary shape over time, the overall type I error rate to be maintained, and the total sample size, which is often based on the amount of power desired at the end of the study. The boundaries are formulated in advance of testing based on these specifications, although sometimes adjustments to the stopping boundary values are made at each analysis time if unplanned circumstances related to the logistics of carrying out the study arise [22]. Various methods for formulating sequential testing boundaries have been proposed, and two common approaches are $\alpha-$spending [32] and unifying family [29]. In the $\alpha-$spending approach, the total amount of type I error, or α, to be spent during the study is divided between each of the pre-specified analysis times. The unifying family literature offers a procedure to generate a wide spectrum of sequential testing boundaries that are based on the test statistic

of interest rather than on $\alpha-$spending. This procedure involves defining the shape of the boundary on the test statistic scale and then iteratively searching for a boundary until one that results in the desired overall type I error rate is found. This literature also shows that stopping rules designed on the standardized test statistic scale are easily transformed to rules on the $\alpha-$spending scale.

Different sequential testing boundaries have different properties and thus are considered appropriate for use in different contexts. In RCTs, common sequential testing designs involve testing on a bi-annual or quarterly basis using a Pocock boundary [41], an O'Brien-Fleming boundary [40], or general power family boundary function [23]. For instance, in a study to detect efficacy, where overall power is of highest priority (since we do not want an intervention to be incorrectly deemed ineffective), the O'Brien-Fleming boundary may be desired [40]. In the safety evaluation setting, where detecting a safety signal as quickly as possible is a primary goal, a Pocock boundary (which is lower at earlier time points compared to an O'Brien-Fleming boundary) has been frequently used [31, 36]. Another sequential design characteristic that is often varied depending on the context of the study is the frequency of testing. A higher testing frequency than is typically used in RCTs (e.g., weekly testing) has been used in the safety monitoring setting, particularly for vaccines [37, 31, 46]. This is considered justified since timeliness, even at the expense of a slight reduction in statistical power, is a priority in safety monitoring. This allows us to detect safety signals as quickly as possible and avoid any potential harm to future vaccine recipients, who often are from vulnerable populations such as healthy infants and children. Moreover, in observational surveillance it is usually feasible for safety monitoring to continue for a longer period of time in order to gain more statistical power since the cost of additional data capture within the health care databases that are commonly used for safety monitoring is minimal. However, if the reduction in power from more frequent testing is substantial, then signal detection may actually take longer than it would with less frequent looks. These power and timeliness trade-offs are important issues to examine when designing a study.

One can think about the choice between the Pocock and O'Brien-Fleming boundaries in the following context. The Pocock boundary was developed in the $\alpha-$spending framework, and has the property that, when plotted on the standardized test statistic scale, it is ap-

proximately flat over time. If we instead use the unifying family approach for computing the stopping boundary, we can specify a boundary that we call 'Pocock-like,'which is exactly flat over time on the scale of a standardized test statistic. The O'Brien-Fleming boundary decreases proportionally to the square root of the total amount of statistical information spent up to time t (i.e., decreases $\approx \sqrt{N_t/N}$ where N_t is the sample size up to time t and N is the final sample size) on the standardized test statistic scale. Thus, the Pocock and Pocock-like boundaries compared to the O'Brien-Fleming boundary spend more α (relative to the amount of statistical information) at earlier decision times, and less α at later decision times. In the safety monitoring setting, where rapid detection of an elevated number of AEs is often of highest priority, this lack of early conservatism can be a desirable property as it increases the power to detect a safety signal early in monitoring. The trade-off for this earlier spending of type I error with a Pocock or Pocock-like boundary is that to achieve comparable power, a larger maximum sample size is required compared to that when using an O'Brien-Fleming boundary.

2.2 Observational settings

Sequential monitoring for rare safety outcomes in the observational setting is relatively new. In this setting, where randomization is absent, sequential monitoring requires additional consideration in order to ensure valid statistical inference. A primary concern when conducting sequential monitoring in an observational setting is confounding. A confounder is a covariate that is associated with both the treatment (in our setting, the newly-licensed vaccine) and the outcome (a rare AE). In RCTs, the randomization of subjects into treatment groups allows us to assume that the treatment groups do not differ systematically in the expected values of both the measured and unmeasured baseline covariates. Since observational studies lack random allocation to treatment groups, some form of adjustment for the potential differences in measured baseline covariates should be made in order to avoid confounding bias in the estimated treatment effect. (Note that there is no way to adjust for unmeasured confounders.)

When conducting pediatric observational vaccine safety surveillance across multiple health plan databases, it has been common to adjust for age, sex, and site as potential

confounders. Depending on the health plan site, use of a new vaccine over a pre-existing vaccine may be more or less likely. For instance, with emerging exposure uptake, initial usage of a new vaccine may be lower at larger sites than at smaller sites. In this case children enrolled at a smaller site would be more likely to receive the new vaccine than children who visit a larger site, at least during the beginning portion of safety monitoring. In addition, physicians at different sites may have different AE coding practices; that is, they may differ in the way that they record AEs using International Classification of Diseases, 9^{th} Revision (ICD-9) codes. Thus, children enrolled at certain sites may be recorded to have experienced fewer (or more) fevers, for example, than they experienced in actuality or compared with children at other sites. These differences in vaccine uptake and in AE coding practices by site mean that site is often associated with both the probability of exposure to the vaccine and the probability of outcome, making it a confounder for which statistical analyses must be adjusted.

In traditional (i.e., non-sequential) observational studies, adjustment for confounders is standard practice and has generally been accomplished by using one of the following approaches: regression adjustment for the measured confounders, matching subjects on the measured confounders or on propensity scores, stratifying subjects into groups defined by different categories of the measured confounders, or employing inverse probability of treatment weighting where the weights are a function of the measured baseline confounders. However, most existing sequential methods were developed in the context of RCTs and thus have not historically focused on incorporating confounder control. Theory suggests that similar confounding control techniques, such as regression [27], can be used in sequential settings, but these approaches have not been well studied in observational contexts in practice or with rare events. Recently, as the framework for rapid post-licensure safety surveillance has emerged, sequential methods with confounding adjustment have been proposed and applied in observational safety monitoring. A brief history of selected sequential monitoring methods developed for and used in this setting follows.

2.3 Methods for sequential monitoring

One approach recently used within the VSD [14, 20, 9] is the maximized sequential probability ratio test (MaxSPRT), detailed by Kulldorff [31]. This method uses a generalized LRT statistic [46] with continuous sequential testing of composite hypotheses and incorporates adjustment for potential confounders. It was developed as an extension of Wald's sequential probability ratio test (SPRT) [47] for continuous testing of a simple null against a simple alternative hypothesis. MaxSPRT further assumes a flat continuous boundary on the LRT scale, but incorporates an additional analysis time at the end of the study.

The MaxSPRT method has two versions, which allow for different choices of comparator groups to control for confounding. The Binomial MaxSPRT is for the two sample setting (i.e., for independent concurrent controls, who receive a comparable vaccine in the same time frame as the newly-licensed vaccine is administered) while the Poisson MaxSPRT is for the one sample setting (i.e., for comparisons to historical controls, from whom expected AE rates, which are assumed known, are calculated). This thesis focuses on the concurrent control setting and thus on use of the Binomial MaxSPRT. To control for confounding, the Binomial MaxSPRT uses exposure matching with a fixed matching ratio. Since the method involves continuous testing (which is often implemented in practice as weekly testing), matching is conducted among the subset of newly accrued subjects each week. As a relatively limited number of subjects may accrue between weekly tests, use of fixed ratio matching can be difficult. Matching can be further complicated by the strictness of the matching criteria on other confounders (e.g., age). This is an example of the "bias-variance" trade-off, where matching too strictly can result in a loss of subjects and thus a loss of power, but matching too loosely can introduce bias.

The MaxSPRT approach can be viewed as part of a broader class of sequential generalized LRTs that employs continuous testing and has been detailed by Shih [46]. Zhao et. al [52] also recently described and implemented a similar generalized LRT method on a group sequential testing basis. This method, which we call a group sequential likelihood ratio test (GS LRT), has been evaluated for use within the FDA's Mini-Sentinel pilot [2]. GS LRT is more broadly applicable than MaxSPRT because, in addition to allowing greater flexibility

in testing frequency (non-continuous testing is an option), it was designed to accommodate various boundary shapes over time using a unifying family boundary approach.

Two other methods that were designed specifically for the observational safety monitoring setting and offer additional options for confounder control have also been proposed. One of these is the conditional sequential sampling procedure (CSSP) [34]. CSSP is an $\alpha-$spending approach that uses stratification to adjust for confounders and is designed for chronic exposure and rare AE settings. Another recently developed method, group sequential generalized estimating equations (GS GEE) [18], allows analysis-based confounder adjustment using regression. GS GEE uses estimating equations theory to compute a score test statistic. This method was designed for use with rare AEs and does not rely on asymptotic assumptions for boundary formulation. Instead, it uses a unifying family approach to compute the stopping boundary. Computing the test statistic and boundary in this way makes this method robust in the rare event setting, although also more computationally intensive, relative to other available methods.

A final method worth considering for use in observational sequential safety monitoring was developed by Lan and DeMets and has primarily been used in RCTs [32]. Lan and DeMets' approach compares a standardized test statistic to a stopping boundary that is formulated using an $\alpha-$spending approach. Two important assumptions on which this method relies are the asymptotic normality of the test statistic and the accrual of data in independent increments. This method was an improvement over existing RCT sequential testing methods at that time because it did not require prior specification of the number and timing of the analyses, it involved a relatively simple boundary calculation, and it relied on a well-defined asymptotic distribution. Although this method was designed for use in RCTs, it can accommodate confounder control, for instance as shown by Jennison and Turnbull [27]. We refer to the Lan and DeMets method with confounder control as group sequential Lan-DeMets (GS LD). GS LD is appealing for its simplicity and its flexible confounding control, but it warrants further investigation in the observational safety setting because monitoring for rare AEs and highly frequent testing may violate the large sample and independent increments assumptions on which the method relies.

As evidenced by the methods described in this section, sequential testing for observa-

tional safety surveillance is relatively new. A variety of approaches with different methods for confounder adjustment are now available. While there has been some investigation into the performance of these methods [18, 46, 38, 37, 31, 16], there remains much to be learned about their comparative performance, both in simulated settings and in practice. In particular, little is known about how the use of matching to control for confounding compares to an analysis-based regression adjustment approach and how the strength of confounding by health plan site affects method performance. In this thesis, we examine the following group sequential methods for observational safety monitoring of pediatric vaccines: GS LRT using one-to-one exposure matching to control for confounding; GS GEE using regression adjustment to control for confounding; GS LD-R using regression adjustment to control for confounding; and GS LD-M using one-to-one exposure matching to control for confounding. In the next chapter we more explicitly define each of these approaches.

Chapter 3

SEQUENTIAL MONITORING METHODS COMPARED IN THIS WORK

We begin by defining the notation used throughout this thesis, and then we describe each of the sequential monitoring methods that we compare. The general characteristics of each method are shown in Table 3.1.

Table 3.1: Overview of the basic characteristics of the four group sequential methods we compare for use in observational safety monitoring.

Method	Type of confounder control	Boundary formulation approach	Test statistic
GS LRT	1:1 exposure matching	Unifying boundary	Likelihood ratio
GS GEE	Regression adjustment	Unifying boundary	Score
GS LD-R	Regression adjustment	$\alpha-$spending	Score
GS LD-M	1:1 exposure matching	$\alpha-$spending	Wald

3.1 Notation

We denote the sequential analysis times by $t = 1, 2, \ldots, T$ and the individuals (who receive either the newly-licensed vaccine or a comparator vaccine) up to time t by $i = 1, 2, \ldots, n_t$. Exposure to the newly-licensed vaccine for subject i is given by $X_i = 1$ and lack of exposure by $X_i = 0$. The AE outcome is denoted $Y_i = 1$ if subject i experienced the pre-specified AE and $Y_i = 0$ otherwise. Baseline confounders (i.e., confounders that have fixed values, known prior to vaccination) for subject i are given by $\mathbf{Z_i}$. When monitoring vaccine-related safety outcomes in children using concurrent comparators, three confounders commonly considered are age, sex, and site, thus $\mathbf{Z_i} = \{Z_{age,i}, Z_{sex,i}, Z_{site,i}\}$.

The hypothesis of interest in our setting is based on the relative risk (RR) of a pre-specified AE conditional on $\mathbf{Z}$, comparing a child vaccinated with the newly-licensed vaccine

($X_i = 1$) to a child not vaccinated with the newly-licensed vaccine ($X_i = 0$), namely

$$H_0 \quad : \quad RR = 1$$
$$H_A \quad : \quad RR > 1.$$

To test this hypothesis, a stopping boundary, $c(t)$, and test statistic, $S(t)$, are computed according to the sequential monitoring method used. The hypothesis is tested at each pre-specified analysis time, t, and if $S(t)$ exceeds $c(t)$, then H_0 is rejected. Otherwise, we continue to the subsequent analysis time. This process is repeated until we detect a safety signal or until we reach the pre-defined end of study. The end of the study can be defined based on calendar time (e.g., stop monitoring after two calendar years have passed), sample size (e.g., stop monitoring after accruing 10,000 subjects), or statistical information (e.g., stop monitoring after observing a certain number of AEs). The choice of which metric to use depends largely on the context of the monitoring and the decision-makers involved. We use calendar time because this is a common choice used in the VSD. At each analysis time, more subjects (and potentially more instances of the AE) are available for analysis, so in theory more statistical information is available as time goes on. The function $c(t)$ is designed to maintain the overall type I error rate by adjusting for the repeated testing and accounting for the skewed distribution of the test statistic that arises from conditioning on whether or not earlier test statistics exceeded the stopping boundary.

In simulations performed in this thesis, we conduct quarterly analyses and hold the shape of the stopping boundary constant (i.e., flat or nearly flat on the scale of the standardized test statistic) across all methods. The test statistic, the method used to formulate the stopping boundary ($\alpha-$spending or unifying family), and the method of confounder control all vary according to the sequential monitoring method used (Table 3.1).

3.2 Group sequential likelihood ratio test (GS LRT)

The first method we examine is GS LRT, which is a modification of Binomial MaxSPRT [31] in several ways. Both GS LRT and Binomial MaxSPRT control for confounding using fixed ratio exposure matching. However, the likelihood used by GS LRT is the standard

conditional logistic regression likelihood, which conditions on the matched pairs [12]. This is a slight modification from the Binomial MaxSPRT, which instead conditions on the total number of AEs. The Binomial MaxSPRT likelihood is reasonable in the rare AE setting and is valid under the assumption that only a single AE is observed per matched set. For the one-to-one matching scenario and the rare AE setting, the two approaches are equivalent. In GS LRT's conditional logistic regression likelihood, only the discordant matched sets (i.e., the pairs with either an unexposed AE or an exposed AE, but not both) contribute statistical information. In the one-to-one matching scenario this reduces the data at time t to the k_t discordant, or informative, matched pairs. The data structure for the one-to-one conditional logistic regression approach assumes that k_t is the number of informative matched pairs at time t, $Y_{s,j} = 1$ if the j^{th} $(j = 1, 2)$ subject belonging to informative matched pair s experiences the AE and $Y_{s,j} = 0$ otherwise. $X_{s,j}$ indicates exposure to the vaccine. The conditional logistic likelihood is the following

$$L_c(\beta(t)) = \prod_{s=1}^{k_t} \frac{\exp(\sum_{j=1}^{2} Y_{s,j} X_{s,j} \beta)}{1 + \exp(\beta)}.$$

Assuming this partial conditional likelihood, the maximum likelihood estimate of $\beta(t)$ is

$$\hat{\beta(t)} = log\left(\frac{\sum_{s=1}^{k_t} Y_{s,j} X_{s,j}}{k_t - \sum_{s=1}^{k_t} Y_{s,j} X_{s,j}}\right),$$

which is, among discordant pairs, the log of the total number of exposed subjects who experience the AE divided by the total number of unexposed subjects who experience the AE.

The second modification of GS LRT from the Binomial MaxSPRT is that GS LRT allows group sequential monitoring in addition to continuous sequential monitoring. (Note that continuous sequential monitoring can be thought of as a special case of group sequential monitoring, where the 'group' is an individual.) Finally, GS LRT permits more flexible boundary options. In this paper, we use a Pocock-like boundary, which is flat on the scale of the standardized test statistic. Further, we restrict our attention to one-to-one exposure

matching and do not investigate more flexible matching techniques. The specific matching procedure we use when implementing GS LRT is described in more detail in Section 3.5, and descriptions of more general matching approaches (such as one-to-m or m-to-n matching) are available from Kulldorff [31] and from Cook [18]. We focus on one-to-one matching because it is simple and commonly used in practice. This straightforward setting allows us to explore the basic performance characteristics and practicalities of implementing matching, the latter of which will apply to any matching scheme. The GS LRT test statistic at time t uses the standard conditional logistic likelihood, assumes a one-sided alternative hypothesis ($H_A : RR > 1$), and is given by the following:

$$S_{LRT}(t) = \begin{cases} log\left(\frac{L_c(\hat{\beta(t)})}{L_c(0)}\right), & \text{if } \sum_{s=1}^{k_t}\sum_{j=1}^{2} X_{s,j}Y_{s,j} > 0.5\sum_{s=1}^{k_t}\sum_{j=1}^{2} Y_{s,j} \\ 0, & \text{otherwise} \end{cases}.$$

Defined in this manner, the log likelihood ratio (LLR) test only rejects the null hypothesis of no difference in AE risk between the exposed and unexposed subjects if we observe more AEs among those vaccinated with the vaccine of interest than expected. Using a one-to-one matching ratio, this means that in order to observe a safety signal, at least half of the observed AEs must be among the exposed.

To formulate the stopping boundary under H_0 we use the unifying family approach, a Pocock-like boundary shape (i.e., a flat boundary $c(t) = c$ where c is constant across all analysis times), and the following simulation approach:

1. For a given dataset, at each analysis time, t, match the subjects one-to-one and retain only the discordant matched pairs. This yields an informative matched pairs dataset $(X_{1,1}, Y_{1,1}, X_{1,2}, Y_{1,2}, \ldots, X_{k_t,1}, Y_{k_t,1}, X_{k_t,2}, Y_{k_t,2})$.

2. Permute the exposure values within each informative matched pair $(X_{s,1}, X_{s,2})$ to create $(X_{s,1}^{(p)}, X_{s,2}^{(p)})$, while fixing the outcomes. This forms a set of P $(p = 1, \ldots, P)$ permuted datasets under the null hypothesis.

3. For each analysis time and each permuted dataset, calculate $S_{LRT}^{(p)}(t)$.

4. Finally, calculate $S_{LRT}^{(p)} = \sup_t S_{LRT}^{(p)}(t)$. Define the critical value, c, as the $(1-\alpha)^{\text{th}}$ percentile of these P permuted realizations of S_{LRT} under H_0.

This simulation approach holds the overall type I error rate at α by setting c equal to the value where, at most, $(1-\alpha) * P$ permuted datasets reject H_0 and signal at any of the time points. Once this stopping boundary, c, has been computed using these simulation steps, we can apply it in practice. Specifically, to assess whether the observed data has signaled, at each analysis time we compare the test statistic, $S_{LRT}(t)$, to c and stop monitoring either at the earliest t such that $S_{LRT}(t) > c$, or at the final analysis time.

3.3 Group sequential generalized estimated equations (GS GEE)

GS GEE is another group sequential method that was developed for, and has been used in, the observational safety surveillance setting [18]. GS GEE is a flexible method that can be applied to monitor chronic or single time exposures for binary, continuous, or count outcomes. The general method is described in [18], but in this thesis we focus on the case with a single point in time exposure and a constant probability of AEs occurring within an acute risk window following exposure (i.e., the vaccine setting). When defining the test statistic, this method takes advantage of the theory of generalized estimating equations [35, 51] and the generalized score statistic [43]. Hence only the mean model must be correctly specified in order to provide consistent, asymptotically normal estimators of the regression parameters. We first assume the following mean model under H_0:

$$g(E[Y_i(t)]) = g(\mu_i) = \beta_0 + \beta_z \mathbf{Z_i}, (i = 1, \ldots, n_t),$$

where $g(\cdot)$ is the logit link function. The marginal variance under H_0, which depends on μ_i, is given by $Var(Y_i|X_i, \mathbf{Z_i}) = \mu_i(1-\mu_i)$. We then specify the family from which the data come, which is binomial in our binary AE case, and compute the generalized score statistic [43], $S_{GEE}(t)$. Confounding control is accomplished via regression adjustment for the individual confounders.

To compute the stopping boundary, we use the same unifying family approach as the GS LRT. However, the data we use for GS GEE are very different. For this method we

use the entire dataset at each analysis time, $(X_1, Y_1, \cdots, X_{n_t}, Y_{n_t})$ where n_t is the number of subjects accrued between the $(t-1)^{th}$ and the t^{th} analysis times (rather than just the informative matched pairs used by GS LRT). We derive the boundary by permuting the exposure data across subjects within analysis time points (as opposed to within matched pairs within analysis times points as for GS LRT), and calculate the generalized score statistic (instead of the LLR test statistic used in GS LRT). Similarly to GS LRT, we set $c(t) = c$ to be the value where at most $(1-\alpha) * P$ permuted datasets reject H_0 and signal at any of the time points. At each analysis time, we compare $S_{GEE}(t)$ to c, and continue monitoring as long as $S_{GEE}(t)$ does not exceed c.

3.4 Group sequential Lan-DeMets (GS LD)

Lan and DeMets' $\alpha-$spending group sequential approaches [32] have primarily been used in RCTs, but here we adapt a specific $\alpha-$spending procedure for use in observational safety monitoring, as done previously by Cook [18]. In particular, we incorporate control for confounding in the methods we call GS LD-R, which uses regression to control for confounding, and GS LD-M, which uses one-to-one exposure matching to control for confounding. In these approaches, the test statistics are directly formulated (either via regression or matching) to control for confounding. In GS LD-R we use the same generalized score statistic as in GS GEE, so that these two methods may be compared fairly. In GS LD-M we use the same conditional logistic regression model as in GS LRT, but instead of using a LRT we use the more commonly used Wald test statistic, $S_{LD-M}(t) = \frac{\beta_x}{\sqrt{Var(\beta_x)}}$, from the conditional logistic regression model where β_x and $Var(\beta_x)$ are estimated using standard maximum likelihood estimation techniques. Note that for both GS LD-R and GS LD-M, the test statistics are standardized.

To specify the shape of the stopping boundary, we use an $\alpha-$spending approach. At sequential testing time point t, the cumulative type I error rate is given by the monotonic increasing function $\alpha(t)$, where

$$\begin{aligned} \alpha(t) &= \text{cumulative type I error spent at the } t^{th} \text{ analysis time, and} \\ 0 &< \alpha(1) = \ldots = \alpha(t) = \ldots = \alpha(T) = \alpha, \end{aligned}$$

where α is the total type I error we wish to spend over the course of the study. Any function that preserves the family-wise error rate can be used, and common choices include the Pocock boundary function, $\alpha(t) = \frac{\alpha}{T} log(1 + (e - 1) * t)$, and the O'Brien-Fleming boundary function, $\alpha(t) = 2 - 2\phi\left(\frac{z_{1-\alpha/2}}{\sqrt{t}}\right)$. We use a Pocock boundary, which is nearly flat on the scale of the standardized test statistic. Based on a given error spending function, Lan and DeMets [32] developed a method that constructs a conditional (on not having stopped yet) asymptotic stopping boundary to which a standardized normal test statistic, in our case $S_{LD-R}(t)$ or $S_{LD-M}(t)$, can be compared. To compute the stopping boundary, we use the *ldbounds* package [13] in R. The boundary values, $c(t)$, for this method only vary with the number of looks (the fewer the looks the lower the boundary values); thus GS LD-R and GS LD-M use the same stopping boundary as long as the same number of analyses are performed. If $S_{LD}(t)$ exceeds $c(t)$ monitoring is stopped, otherwise data collection continues until the next analysis time or the pre-specified end of study.

3.5 Propensity score matching

When matching in the rare AE setting, estimation can be especially difficult if there are many confounders requiring control. One way to proceed in this situation is to summarize the confounders with a single score, such as the propensity score (PS), and to use that score to control for all of the confounders. In light of this motivation for using the PS, and given its recent popularity for confounding control [19, 6, 45, 4, 5], we implement one-to-one exposure matching on the PS in both the GS LRT and GS LD-M methods. To estimate the PS, a standard approach is to fit a logistic regression model, $logit(P(X_i|\mathbf{Z}_i)) = \beta_{x,z}\mathbf{Z}_i$, and define the estimated PS as the predicted probabilities from the model. In our case, the observed covariates used to obtain the estimated PS are the observed confounder values $\mathbf{z}_i$, yielding

$$\hat{PS}(\mathbf{z}_i) = \frac{\exp(\hat{\beta}_{x,z}\mathbf{z}_i)}{1 + \exp(\hat{\beta}_{x,z}\mathbf{z}_i)}.$$

While there are various ways to use the PS to control for the measured confounders, including as an adjustment variable in a regression model or as inverse probability of treatment weights [18], we focus on one-to-one matching based on the PS. Controlling for confounding

via PS matching instead of jointly matching on individual covariates is advantageous because the use of the PS allows many confounding variables to be summarized into a single variable upon which matches can be formed. Particularly with multiple confounders, it can be difficult to form matches based on the covariate values, which can lead to discarding much of the data (due to the inability to find matches for some subjects). For instance, the total number of strata formed by two sexes, two age groups, and five sites is 20. In the rare AE setting it is likely that some of these strata will not contain subjects who experience an AE, and the loss of subjects (for any reason) is especially troublesome when AEs are rare. When matching, any strata not containing subjects who experienced an AE are excluded from the analysis because these strata do not contain any statistical information. Thus, the fewer individual covariates we match on, the less statistical information we stand to lose. The downside to matching on fewer individual covariates is that confounding control may not be as tight. Another issue is that PS matching may match subjects who are dissimilar in terms of their actual baseline covariates; for instance a match may form between a 6 week old male and an 18 month old female if they have a similar propensity to be vaccinated. When deciding to match on the PS, we must weigh all these considerations.

In our implementation of PS matching, at each analysis time we estimate the PS only for the new subjects who have just entered the study and then match these newly exposed subjects to newly unexposed subjects from the same site. There are innumerable ways to form matches based on the PS, and we implement a matching scheme that aligns with recent published studies [4]. Namely, we match exposed to unexposed subjects enrolled at the same health plan site whose PSs are within 0.025 of each other [8]. For simplicity, we match without replacement because matching with replacement means that the same unexposed subject could be matched to multiple exposed subjects, and when this occurs, it is necessary to account for the lack of independence between the pairs [5]. The R package *Matching* [44] is used to form matches within the data. In order to implement the matched methods to their best ability and to follow good statistical practice of analysis following design, we account for the matching in our analyses for both GS LRT and GS LD-M by conditioning on the matched pairs when computing the test statistic via conditional logistic regression. For GS LRT we further incorporate matching when calculating the sequential

testing boundaries by permuting exposures within matched pairs.

Chapter 4

SIMULATION STUDY METHODS

We performed a simulation study to compare the performance characteristics of the four sequential methods (GS LRT, GS GEE, GS LD-R, and GS LD-M) described in Chapter 3 in settings varying the exposure prevalence, AE prevalence, amount of confounding by site, and sample size at each site. Our primary aim was to compare the methods that use matching to control for confounding to the regression-based adjustment methods. Matched methods are being used in this setting in practice but they have not been systematically evaluated for use in sequential observational safety surveillance where the AEs being monitored are often rare. We compare the methods based on type I error rate, power, and time-to-surveillance-end. For a given simulation setting we performed 1,000 simulations and used the same data across the methods. The type I error rate for a simulation set was computed as the mean number of simulations that stopped monitoring early under the null hypothesis. Time-to-surveillance-end was defined across the 1,000 simulations as the average number of days until we either detected a safety signal or reached the pre-defined end of study, whichever occurred first. Power was computed as the mean number of simulations under the alternative hypothesis that stopped monitoring early. Simulations were performed using R 2.14.2 [42].

4.1 Generating data

For each simulation, we generated a dataset with $n = 10,000$ subjects and included three confounders: age group (0-1 year or >1 year), sex, and site, which was generated from a multinomial distribution. Table 4.1 shows the precise distributions and definitions for each of these confounders. We specified $T = 8$ equally-spaced analyses conducted over a two year period, held the type I error rate at 0.05, and used a stopping boundary that was flat or nearly flat on the standardized test statistic scale (i.e., Pocock-like or Pocock boundaries). We define $\bar{\mathbf{Z}}$ as the average confounder distribution across all confounders.

The average probability of an AE for the unexposed subjects given the average confounder distribution for the entire population (both exposed and unexposed subjects), $P(Y|X = 0, \bar{\mathbf{Z}})$, was set either to 0.05 or 0.01. For the simulations where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$, we set the $RR(Y|X, \mathbf{Z}) = 1.3$ and for the simulations where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, we set the $RR(Y|X, \mathbf{Z}) = 2$. We used different RRs for the settings where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ versus 0.01 so that we would achieve and be able to evaluate power in a desirable range (0.80 - 0.90) for both settings. Finally, we set the average probability of exposure, $P(X)$, to 0.25. This meant that subjects within some sites or age groups may have had an exposure prevalence greater or less than 0.25, but that the average exposure prevalence across all confounders was 0.25. This value was chosen because (1) it is realistic in practice for a newly-licensed vaccine in the initial stages of uptake in a population, (2) using a larger value when doing exposure matching could result in the loss of some of the exposed individuals (since some strata could have an exposure prevalence of greater than 50%), and (3) using a smaller value would result in a smaller matched sample because we are matching one exposed individual to one unexposed individual. A smaller matched sample would be problematic because this would likely mean fewer informative matches and thus a reduction in power, and we did not want to penalize the matched methods in this way. In choosing $P(X) = 0.25$, our primary goal was to design a simulated data set that was realistic based on previous observational safety monitoring studies, but also fair when considering the resulting size of the matched sample. We expected to see a loss of power for the matched methods, and so wanted to make them as powerful as possible, given realistic practical constraints, when comparing them with the regression methods.

For each subject, the categorical confounders $\mathbf{Z_i} = [Z_{age,i}, Z_{sex,i}, Z_{site1,i}, Z_{site2,i}, Z_{site3,i}, Z_{site4,i}, Z_{site5,i}]$ were defined according to the distributional assumptions shown in Table 4.1. For each simulation, all subjects had the same RRs of exposure and AE given sex and age. Technically, sex was generated as a precision variable and not a confounder because we set the RR for exposure comparing boys to girls to be one. This is reasonable to assume because exposure to a new vaccine is typically not differential between boys and girls within the same health care site. However, adjustment for sex is common in this type of monitoring, and so we included it in our work to reflect this practice. The RR

of AE may plausibly differ by sex, thus we chose a modestly elevated $RR(Y|X, Z_{sex} = 1)$ of 1.2 among vaccinated boys. We specified the risk of exposure to be higher for children between 0-1 year of age than for children older than one year, $RR(X|Z_{age}) = 0.5$, because many infant vaccines are given as part of a series, with three doses administered prior to one year of age and a single booster after one year. We also specified a higher risk of AE for older children, $RR(Y|X, Z_{age}) = 1.5$ since children who have already received several vaccinations often have increased reactogenicity to later doses of the same vaccine because their bodies recognize some of the vaccine agents as foreign (i.e., immune memory after primary vaccination).

Confounding and heterogeneity by site can be complex in observational safety monitoring settings like the VSD. Thus, a key goal of our simulation study was to explore method performance under a variety of site sample size and confounding scenarios. Specifically, we conducted simulations with five health care sites because this was computationally tractable and a reasonable number based on previous VSD studies. We performed three sets of simulations with equal-sized sites (e.g., the sites had approximately equal sample sizes) where all the sites had equal exposure prevalence. However, since the health care sites involved in safety monitoring often vary dramatically in terms of size, we focused most of our simulation scenarios on unequal-sized sites (two large and three small sites). For site, $\mathbf{Z_{site}} = \{Z_{site1}, Z_{site2}, Z_{site3}, Z_{site4}, Z_{site5}\}$ different combinations of $RR(X|\mathbf{Z_{site}})$ and $RR(Y|X, \mathbf{Z_{site}})$ were used for different simulation sets. For a given simulation set we selected, from Table 4.1, a set of RRs from the exposure column for site and a set of RRs from the AE column for site, and we describe this in detail below. We used RRs of magnitude 0.5 or 2 depending on the direction of the relationship (lower or higher prevalence of AE compared to the reference site, respectively) to keep the magnitude of the confounding consistent across scenarios.

The exposure model used in the simulations was the following:

$$\begin{aligned} log[P(X_i|\mathbf{Z_i})] &= \beta_{0,X} + RR(X|Z_{sex})Z_{sex,i} + RR(X|Z_{age})Z_{age,i} + RR(X|Z_{site2})Z_{site2,i} \\ &+ RR(X|Z_{site3})Z_{site3,i} + RR(X|Z_{site4})Z_{site4,i} + RR(X|Z_{site5})Z_{site5,i}, \end{aligned}$$

Table 4.1: Distributions and RRs for the confounders in the simulation study

	Strength of association between confounder and:	
	Exposure	AE
Confounder definition ($\mathbf{Z}$)	RR($X|\mathbf{Z}$)	RR($Y|X, \mathbf{Z}$)
Sex[a]: male or female P(male) $\sim$ Bernoulli(.5)	1	1.2
Age: 0-1 yrs or $>$ 1 yr P($>$ 1 yr) $\sim$ Bernoulli(.25)	0.5	1.5
5 equal-sized sites M[b](.2, .2, .2, .2, .2) Site 1 (reference) Site 2 = 0 or 1 Site 3 = 0 or 1 Site 4 = 0 or 1 Site 5 = 0 or 1	Equal exposure prevalence at all sites Site 1 (reference) Site 2: 1 Site 3: 1 Site 4: 1 Site 5: 1	Large sites have lower AE prevalence Site 1 (reference) Site 2: 1 Site 3: 2 Site 4: 2 Site 5: 2
2 large and 3 small sites M(.425, .425, .05, .05, .05) Site 1 (reference) Site 2 = 0 or 1 Site 3 = 0 or 1 Site 4 = 0 or 1 Site 5 = 0 or 1	Large sites have lower exposure prevalence Site 1 (reference) Site 2: 1 Site 3: 2 Site 4: 2 Site 5: 2	Large sites have higher AE prevalence Site 1 (reference) Site 2: 1 Site 3: .5 Site 4: .5 Site 5: .5
	1 large site has higher exposure prevalence Site 1 (reference) Site 2: 2 Site 3: 1 Site 4: 1 Site 5: 1	1 large site has lower AE prevalence Site 1 (reference) Site 2: .5 Site 3: 1 Site 4: 1 Site 5: 1
	1 large site has lower exposure prevalence Site 1 (reference) Site 2: .5 Site 3: 1 Site 4: 1 Site 5: 1	

[a]Sex is not a confounder because it is not related to both exposure and outcome, however we adjust for sex in our analyses because this is common practice in observational safety monitoring.

[b]Multinomial

where $\beta_{0,X}$ was defined to hold the overall prevalence of X at 0.25 for the average confounder distribution (e.g. $Z_{sex,i}$=$\bar{Z}_{sex}$ is the proportion of males and similarly for the other confounders). The outcome model was:

$$\begin{aligned} log[P(Y_i|X_i, \mathbf{Z_i})] &= \beta_{0,Y} + RR(Y|Z_{sex})Z_{sex,i} + RR(Y|Z_{age})Z_{age,i} + RR(Y|Z_{site2})Z_{site2,i} \\ &+ RR(Y|Z_{site3})Z_{site3,i} + RR(Y|Z_{site4})Z_{site4,i} + RR(Y|Z_{site5})Z_{site5,i} \\ &+ \beta_X X_i, \end{aligned}$$

where $\beta_{0,Y}$ was defined to hold the AE prevalence among the unexposed subjects, for the average confounder distribution, at either 0.01 or 0.05 depending upon the outcome prevalence of interest.

Since we held the average prevalence of exposure constant at 0.25, increasing the exposure to the vaccine at one site necessarily decreased the exposure to the vaccine at other sites. For the simulations with equal-sized sites, we assumed that there was an equal probability of exposure to the newly-licensed vaccine at each site, $RR(X|\mathbf{Z_{site}}) = 1$ for sites 2, 3, 4, and 5 compared to site 1. For the RR of AE given exposure and site, $RR(Y|X, \mathbf{Z_{site}})$, in these equal-sized site settings we chose from three site-AE scenarios based on trends seen in previous VSD studies, where: two sites have lower AE prevalence than the other three sites; two sites have higher AE prevalence than the other three sites; and, one site has lower AE prevalence than the 4 other sites. Since site was not related to vaccine exposure in these three simulation sets, it was not a confounder and therefore adjustment for it could hinder the performance of the methods. For the simulations with two large and three small sites, we considered three different site-exposure scenarios and three different site-AE scenarios based on trends seen in previous VSD studies. The site-exposure scenarios included: both large sites have lower exposure prevalence than the small sites; one large site has higher exposure prevalence than the other four sites; and, one large site has lower exposure prevalence than the other four sites. The site-AE scenarios included: both large sites have lower AE prevalence than the small sites; both large sites have higher AE prevalence than the small sites; and, one large site has lower AE prevalence than the 4 other sites.

4.2 Impact of rare AEs on implementation of regression methods

Application of all the methods in the rare AE setting was sometimes complicated by lack of observed AEs. In order to compute the test statistics for GS GEE and GS LD-R, we needed to observe some minimum number of AEs to have enough statistical information. Since at each analysis time we computed the test statistic based on all data accrued up to that time, if there were enough AEs at the first scheduled analysis time, then there were enough AEs at all remaining analysis times. When AEs were rarer, for instance when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, there were not always enough observed AEs at the first analysis time to compute the test statistic. One way to handle this situation is to delay the first analysis until an adequate amount of statistical information accrues [37], and this is the approach we took in our simulations. The procedure was implemented as follows: After the data are generated for a given simulation data set, make sure there is at least one exposed and one unexposed AE case at the first analysis time. If not, then postpone the first analysis time until the second scheduled analysis time. If the combined data from the originally planned first two analysis times do not contain at least one exposed and one unexposed AE case, then postpone again, and so on. Since adjusting the analysis times in this way meant that we performed fewer tests than originally planned, we re-computed the critical values to reflect the new reduced number of tests performed under this implementation scheme that skipped analyses if there was insufficient statistical information. Analyzing the data fewer times lowers the resulting stopping boundary, since with fewer tests there are fewer occasions to make a type I error. Once this initial data check has been performed and satisfied, the regression methods can then be applied.

4.3 Impact of rare AEs on implementation of matched methods

Additional complications arose when we matched the data. For GS LRT and GS LD-M, the data first went through the same procedure, described in Section 4.2, as the regression approaches. Next, the data were matched as described in Section 3.5. When using one-to-one matching for confounder adjustment, statistical information corresponds to informative pairs (e.g., pairs with either an unexposed AE or an exposed AE, but not both). These

informative pairs are the data used to compute the test statistic at each analysis time. Thus, if no additional informative pairs have accrued at a particular analysis time since the prior analysis, then computing a test statistic does not make sense (since it will have the same value as the test statistic from the previous analysis time). To handle this situation, after matching the data at a given analysis time, we checked to make sure that the new data at that analysis time contained at least one new informative pair. If not, then we skipped that test. Thus, we sometimes skipped an analysis in the middle of monitoring. For instance, even if there were informative pairs at the first planned analysis, if no new informative pairs accrued between the first analysis time and the second planned analysis time, then we skipped the second planned analysis and checked again for informative pairs at the third planned analysis time.

For both GS LRT, and GS LD-M we adjusted the critical boundary to reflect *any* skipped analyses (due to either of the reasons described in this section or Section 4.2). Since we are in the rare event setting and anticipated that some analyses would need to be skipped due to matching, we tracked this metric.

Chapter 5

RESULTS

To assess the performance characteristics of GS GEE, GS LD-R, GS LRT, and GS LD-M, we performed a simulation study where we varied the exposure prevalence, AE prevalence, amount of confounding by site, and the sample size at each site. In addition, we applied the four methods to a subset of data from the VSD's monitoring of a DTaP-IPV-Hib vaccine, and varied the frequency of the group sequential testing. Specifically, we used quarterly, monthly, and weekly testing schedules where the first test was delayed for 12 months.

5.1 Simulation Study Results

5.1.1 Practical issues related to method implementation

As one comparison metric of the regression versus matched methods, we considered the number of subjects contributing information to each of the test statistics. For the regression methods, all 10,000 simulated subjects were used to compute the test statistics for each simulation. For the matched methods, we computed the average number of subjects matched ($n_{matched}$) and the average number of subjects belonging to informative matches (n_{inform}). Since only the informative matches were used to compute the test statistics for the matched methods, these values provided insight into the performance of GS LRT and GS LD-M. Table 5.1 shows the number of subjects matched and the number of subjects belonging to informative matches under the alternative hypotheses for both $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ and $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$. As expected, when the probability of an AE was lower, there were fewer informative matches, despite the fact that the total number of subjects matched was similar for the two AE settings. Specifically, when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$, 11% of the matched subjects belonged to informative matches, and when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, only 3% of matched subjects belonged to informative matches. The percentages of subjects belonging to informative matches reflect the overall exposure prevalence and $P(Y|X = 0, \bar{\mathbf{Z}})$.

Table 5.1: The average number of subjects matched ($n_{matched}$) and average number of subjects belonging to informative matches (n_{inform}) among n=10,000 simulated subjects, where $P(Y|X=0,\bar{\mathbf{Z}})=0.05$ (left two columns) and $P(Y|X=0,\bar{\mathbf{Z}})=0.01$ (right two columns), under their respective alternative hypotheses, $H_A: RR(Y|X=1,\mathbf{Z})=1.3$ and $H_A: RR(Y|X=1,\mathbf{Z})=2$.

Exposure scenario	AE scenarios	$P(Y\|X=0,\bar{\mathbf{Z}})=0.05$		$P(Y\|X=0,\bar{\mathbf{Z}})=0.01$	
		$n_{matched}$[a]	n_{inform}[b]	$n_{matched}$	n_{inform}
Equal exposure prevalence	2 sites have lower AE prevalence	5202.0	573.2	5197.3	157.4
	2 sites have higher AE prevalence	5200.9	574.3	5202.0	158.8
	1 site has lower AE prevalence	5205.7	564.5	5205.9	154.1
Large sites have lower exposure prevalence	Large sites have lower AE prevalence	5139.4	592.3	5134.3	163.6
	Large sites have higher AE prevalence	5134.5	**534.6**[c]	5136.6	**146.1**
	1 large site has lower prevalence AEs	5137.7	581.1	5140.5	160.6
1 large site has higher exposure prevalence	Large sites have lower AE prevalence	5517.6	578.3	5511.6	157.4
	Large sites have higher AE prevalence	5518.1	608.5	5520.1	167.2
	1 large site has lower prevalence AEs	5514.4	545.4	5512.9	149.6
1 large site has lower exposure prevalence	Large sites have lower AE prevalence	5472.0	611.5	5479.3	168.2
	Large sites have higher AE prevalence	5469.6	580.3	5478.5	157.7
	1 large site has lower AE prevalence	5475.9	**659.1**	5470.3	**182.2**

[a]Number of subjects matched

[b]Number of subjects belonging to informative matches

[c]Bold numbers highlight the fewest and most subjects, on average, that belonged to informative matched pairs

There were no clear trends across site-exposure or site-AE scenarios in terms of the number of subjects belonging to informative matches. The fewest subjects belonging to informative matches (shown in bold in Table 5.1: $n_{inform} = 534.61$ when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ and $n_{inform} = 146.05$ when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$), were found in the simulation setting where the large sites had lower exposure prevalence and higher AE prevalence. The maximum number of subjects belonging to informative matches (shown in bold in Table 5.1: $n_{inform} = 659.14$ when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ and $n_{inform} = 182.182$ when $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$), were found in the simulation setting where one large site had both lower exposure prevalence and lower AE prevalence.

As described in Sections 4.2 and 4.3, lack of enough statistical information at the first analysis or the lack of enough *new* statistical information at subsequent analyses resulted in one or more of the pre-specified tests being skipped. In the setting where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$, no analysis times were skipped (for either the regression or the matched methods) because there was enough statistical information at each analysis time in all the simulations to compute the test statistics. For simulations wither lower AE prevalence where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, Table 5.2 shows the number of simulations under the null hypothesis where at least one analysis time was skipped, as well as the average number of analyses skipped. For the simulations under the alternative hypothesis, fewer analysis times were skipped (not shown) because there were more observed AEs and thus more informative matches in each simulation. In general, there were fewer simulations that had to skip analysis times for regression methods than for the matched methods, which was expected since the regression methods only skipped at the outset while the matched methods could also skip intermediate analyses.

There were no clear trends across site-exposure or site-AE scenarios in terms of the number of simulations that skipped for either the regression or matched methods. In fact, the simulation settings with the most and least skipped analyses overall (shown in bold in Table 5.2) were from the same site-exposure scenario, where the large sites had lower exposure prevalence. When the large sites had lower AE prevalence, the fewest simulations skipped analyses, 12, and when the large sites had higher AE prevalence the, the most simulations skipped analyses, 44.

Table 5.2: For simulations where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, the number of simulations under the null hypothesis in which it was necessary to skip at least one analysis time, and the average number of analysis times skipped given that at least one was skipped.

		Regression methods		Matched methods	
Exposure scenario	AE scenario	# of simulations with ≥ 1 skip	Mean # of skipped analyses	# of simulations with ≥ 1 skip	Mean # of skipped analyses
Equal exposure prevalence	2 sites have lower AE prevalence	36	1.08	42	1.10
	2 sites have higher AE prevalence	26	1.04	29	1.03
	1 site has lower AE prevalence	43	1.00	57	1.00
Large sites have lower exposure prevalence	Large sites have lower AE prevalence	**12**[a]	1.00	20	1.00
	Large sites have higher AE prevalence	**44**	1.02	64	1.02
	1 large site has lower AE prevalence	20	1.00	24	1.04
1 large site has higher exposure prevalence	Large sites have lower AE prevalence	27	1.00	36	1.03
	Large sites have higher AE prevalence	31	1.03	36	1.03
	1 large site has lower prevalence AEs	39	1.05	49	1.04
1 large site has lower exposure prevalence	Large sites have lower AE prevalence	23	1.04	29	1.03
	Large sites have higher AE prevalence	25	1.04	31	1.03
	1 large site has lower AE prevalence	20	1.10	22	1.09

[a]Bold number highlight the simulations with the most and least skipped analyses overall

5.1.2 Method performance

Tables 5.3 and 5.4 show the type I error rates for the simulations with equal- and unequal-sized sites, respectively. In general, all methods held the overall type I error rate close to the desired 0.05 level. We note that across all simulation scenarios, the type I error rate ranged from 0.017 to 0.065 for all methods. In general, GS LD-R was the most conservative (e.g., had the lowest type I error rate), followed by GS LD-M, GS LRT, and GS GEE. For some simulations, the type I error rate varied more across the methods, such as for the simulations with unequal-sized sites and $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$ where both large sites had lower exposure prevalence and lower AE prevalence than the small sites (shown in bold in Table 5.4). In this scenario the type I error rates ranged from 0.031 for GS LD-R to 0.059 for GS GEE.

Table 5.3: Type I error rates for the simulation sets with equal-sized sites.

		Regression methods				Matched methods			
		GS GEE $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LD-R $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LRT $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LD-M $P(Y\|X=0,\bar{\mathbf{Z}})$	
Exposure scenario	AE scenarios	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
Equal exposure prevalence	2 lg sites have lower AE prevalence	0.047	0.054	0.037	0.029	0.041	0.047	0.039	0.038
	2 lg sites have higher AE prevalence	0.050	0.057	0.035	0.030	0.045	0.048	0.044	0.038
	1 site has lower AE prevalence	0.050	0.039	0.039	0.017	0.056	0.045	0.053	0.031

Table 5.4: Type I error rates for the simulation sets with unequal-sized sites.

		Regression methods				Matched methods			
		GS GEE $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LD-R $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LRT $P(Y\|X=0,\bar{\mathbf{Z}})$		GS LD-M $P(Y\|X=0,\bar{\mathbf{Z}})$	
Exposure scenarios	AE scenarios	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
Both large sites have lower exposure prevalence	Both large sites have lower AE prevalence	0.046	**0.059**[a]	0.036	**0.031**	0.053	**0.054**	0.056	**0.041**
	Both large sites have higher AE prevalence	0.043	0.044	0.033	0.020	0.043	0.035	0.037	0.026
	1 large site has lower AE prevalence	0.050	0.045	0.038	0.027	0.051	0.040	0.045	0.024
1 large site has higher exposure prevalence	Both large sites have lower AE prevalence	0.043	0.051	0.035	0.026	0.047	0.037	0.043	0.031
	Both large sites have higher AE prevalence	0.058	0.055	0.046	0.025	0.045	0.044	0.036	0.029
	1 large site has lower AE prevalence	0.045	0.043	0.037	0.022	0.050	0.047	0.052	0.037
1 large site has lower exposure prevalence	Both large sites have lower AE prevalence	0.030	0.047	0.025	0.025	0.039	0.038	0.031	0.029
	Both large sites have higher AE prevalence	0.065	0.038	0.044	0.026	0.052	0.039	0.044	0.027
	1 large site has lower AE prevalence	0.062	0.050	0.047	0.030	0.065	0.051	0.053	0.035

[a]Bold numbers highlight the varied type I error rates across the methods for certain simulation settings

Tables 5.5 and 5.6 show the power and time-to-surveillance-end for the simulations with equal-sized sites where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ and $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, respectively. When $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$, power to detect a RR of 1.3 ranged from 0.638 to 0.831 across methods and site scenarios. When $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, power to detect a RR of 2 was relatively high and ranged from 0.816 to 0.965 across methods and site scenarios. Tables 5.7 and 5.8 show the power and time-to-surveillance-end for the simulations with unequal sized sites (i.e., two large and three small sites) where $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$ and $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, respectively. When $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.05$, power to detect a RR of 1.3 ranged from 0.624 to 0.847. When $P(Y|X = 0, \bar{\mathbf{Z}}) = 0.01$, power to detect a RR of 2 was again relatively high and ranged from 0.820 to 0.976.

The regression methods were consistently and substantially more powerful than the matched methods, about 0.08 to 0.20 times more powerful, depending on how rare the AE prevalence was and on the site-exposure and site-AE scenario. In both AE prevalence settings, the two regression methods had similar power, although GS GEE showed slightly higher power compared to GS LD-R. Similarly, the two matched methods had comparable power, although GS LD-M was slightly higher in power than GS LRT. In general across all simulation sets, GS GEE was the most powerful method, followed by GS LD-R, GS LD-M, and GS LRT. Across all methods and all simulation scenarios, time-to-surveillance-end ranged from 312 days to 525 days. Time-to-surveillance-end was shortest for GS GEE, followed by GS LD-R, GS LRT, and GS LD-M.

Among the four exposure scenarios, the highest power was generally observed when either one or both large sites had lower exposure prevalence. Lowest power was observed when one large site had high exposure prevalence. Perhaps surprisingly, power was *not* the highest for the scenarios with equal-sized sites, although in these simulations adjustment for site was made even though it was not a confounder. Power across site-AE scenarios was generally quite similar. In all cases, the site-AE scenario where the large sites had higher AE prevalence (middle rows for each set of site-AE scenarios, shown in bold in Tables 5.5, 5.6, 5.7, and 5.8) differed compared to the other two site-AE scenarios (top and bottom rows for each set of site-AE scenarios). For instance, in Table 5.8, when one large site had higher exposure prevalence and the large sites had higher AE prevalence, GS LD-M

had power of 0.872, whereas for the other two site-AE scenarios, this method had power of 0.839 and 0.844. We also see this trend in Table 5.7 for GS GEE, when the large sites had lower exposure prevalence. Here, the power for the site-AE scenario where the large sites had higher AE prevalence was 0.796, compared with 0.847 and 0.835 for the other two site-AE scenarios. We also note that in both Tables 5.7 and 5.8, the most powerful site-AE scenarios within each site-exposure scenario where those that had AE prevalence in the same direction as the exposure prevalence. For instance, in Table 5.7, when large sites had lower exposure prevalence (top site-exposure scenario), the highest power was achieved by all methods when the large sites also had lower AE prevalence. Similarly, when one large site had lower exposure prevalence (bottom site-exposure scenario), the highest power was achieved when the large sites had lower AE prevalence. We note that in some cases, power was relatively high ($> 90\%$) and this made it more difficult to determine which methods or scenarios were the most powerful.

Table 5.5: Power and time-to-surveillance-end for simulations with equal-sized sites and $RR(Y|X=1,\mathbf{Z})=1.3$ and $P(Y|X=0,\bar{\mathbf{Z}})=0.05$.

Exposure scenario	AE scenarios	Regression methods				Matched methods			
		GS GEE		GS LD-R		GS LRT		GS LD-M	
		Power	Time[a]	Power	Time	Power	Time	Power	Time
Equal exposure prevalence	2 sites have lower AE prevalence	0.831	415.746	0.814	452.271	0.644	504.867	0.671	510.985
	2 sites have higher AE prevalence	**0.820**[b]	429.717	**0.799**	459.758	**0.638**	508.519	**0.652**	519.020
	1 site has lower prevalence AEs	0.811	428.164	0.795	460.854	0.641	504.228	0.660	511.533

[a]Time-to-surveillance-end (in days)

[b]Bold numbers highlight the trend among site-AE scenarios that, when large sites had higher AE prevalence, the power differed compared to the other two site-AE scenarios

Table 5.6: Power and time-to-surveillance-end for simulations with equal-sized sites and $RR(Y|X=1,\mathbf{Z})=2$ and $P(Y|X=0,\bar{\mathbf{Z}})=0.01$.

		Regression methods				Matched methods			
		GS GEE		GS LD-R		GS LRT		GS LD-M	
Exposure scenario	AE scenarios	Power	Time[a]	Power	Time	Power	Time	Power	Time
Equal exposure prevalence	2 sites have lower AE prevalence	0.956	327.903	0.926	384.334	0.816	426.247	0.838	448.436
	2 sites have higher AE prevalence	**0.965**[b]	323.429	**0.938**	380.956	**0.839**	418.668	**0.846**	443.961
	1 site has lower prevalence AEs	0.954	333.447	0.932	394.640	0.828	430.238	0.843	459.002

[a]Time-to-surveillance-end (in days)

[b]Bold numbers highlight the trend among site-AE scenarios that, when large sites had higher AE prevalence, the power differed compared to the other two site-AE scenarios

Table 5.7: Power and time-to-surveillance-end for simulations with unequal-sized sites and $RR(Y|X=1,\mathbf{Z})=1.3$ and $P(Y|X=0,\bar{\mathbf{Z}})=0.05$.

		Regression methods				Matched methods			
		GS GEE		GS LD-R		GS LRT		GS LD-M	
Exposure scenario	AE scenarios	Power	Time[a]	Power	Time	Power	Time	Power	Time
Large sites have lower exposure prevalence	Large sites have lower AE prevalence	0.847	424.147	0.827	451.723	0.695	493.088	0.706	501.397
	Large sites have higher AE prevalence	**0.796**[b]	437.204	**0.776**	466.790	**0.627**	513.176	**0.643**	525.138
	1 large site has lower prevalence AEs	0.835	428.895	0.827	459.119	0.670	512.628	0.689	518.838
1 large site has higher exposure prevalence	Large sites have lower AE prevalence	0.797	439.213	0.780	467.794	0.637	512.994	0.656	520.938
	Large sites have higher AE prevalence	**0.821**	442.318	**0.801**	466.242	**0.677**	501.854	**0.702**	508.154
	1 large site has lower AE prevalence	0.772	448.527	0.770	472.907	0.624	518.290	0.642	522.125
1 large site has lower exposure prevalence	Large sites have lower AE prevalence	0.804	431.086	0.795	457.567	0.670	504.045	0.686	511.441
	Large sites have higher AE prevalence	**0.823**	422.412	**0.818**	451.723	**0.663**	489.161	**0.683**	494.822
	1 large site has lower AE prevalence	0.841	406.980	0.834	431.817	0.735	460.763	0.753	471.720

[a]Time-to-surveillance-end (in days)

[b]Bold numbers highlight the trend among site-AE scenarios that, when large sites had higher AE prevalence, the power differed compared to the other two site-AE scenarios

Table 5.8: Power and time-to-surveillance-end for simulations with unequal-sized sites and $RR(Y|X=1,\mathbf{Z})=2$ and $P(Y|X=0,\bar{\mathbf{Z}})=0.01$.

		Regression methods				Matched methods			
		GS GEE		GS LD-R		GS LRT		GS LD-M	
Exposure scenario	AE scenarios	Power	Time[a]	Power	Time	Power	Time	Power	Time
Large sites have lower exposure prevalence	Large sites have lower AE prevalence	0.963	311.741	0.946	362.693	0.843	408.989	0.852	438.026
	Large sites have higher AE prevalence	**0.955**[b]	339.369	**0.940**	399.283	**0.820**	435.521	**0.830**	463.92
	1 Large site has lower prevalence AEs	0.967	306.732	0.951	366.372	0.861	402.714	0.870	430.578
1 large site has higher exposure prevalence	Large sites have lower AE prevalence	0.951	340.700	0.927	397.157	0.828	428.360	0.839	454.293
	Large sites have higher AE prevalence	**0.955**	326.247	**0.936**	376.351	**0.866**	405.61	**0.872**	432.913
	1 large site has lower prevalence AEs	0.947	342.344	0.920	397.966	0.845	426.351	0.844	457.306
1 large site has lower exposure prevalence	Large sites have lower AE prevalence	0.952	323.703	0.943	371.211	0.862	408.636	0.867	433.839
	Large sites have higher AE prevalence	**0.958**	326.951	**0.941**	387.308	**0.845**	406.406	**0.853**	439.552
	1 large site has lower prevalence AEs	0.976	308.023	0.948	362.093	0.882	390.844	0.887	417.598

[a]Time-to-surveillance-end (in days)

[b]Bold numbers highlight the trend among site-AE scenarios that, when large sites had higher AE prevalence, the power differed compared to the other two site-AE scenarios

5.2 Application of the methods to DTaP-IPV-Hib vaccine data

To further illustrate and compare the sequential methods described in this work, we applied them to data from a prior VSD study of the safety of a combination DTaP-IPV-Hib vaccine (trade name Pentacel; Sanofi Pasteur, Swiftwater, Pennsylvania) [25, 39]. VSD surveillance of this vaccine began in 2008, following its licensure. Children six weeks up to three years of age from seven different medical care organization sites were monitored for the occurrence of seven pre-specified AEs (including medically-attended fever, seizure, and serious allergic reactions). Eleven group sequential analysis times were specified based on the number of accrued doses of DTaP-IPV-Hib vaccine. The sequential tests compared children vaccinated with DTaP-IPV-Hib vaccine to a group of historical comparators vaccinated with a different DTaP-containing vaccine during the four years prior to the licensure of the combination DTaP-IPV-Hib vaccine. No safety signals were found during the surveillance of DTaP-IPV-Hib vaccine using this historical comparator group.

Secondarily, concurrent comparator data, which included recipients of a different DTaP-containing vaccine during the same period of time as the new combination DTaP-IPV-Hib vaccine was received by VSD study subjects, were also collected and analyzed. No safety signals were found using this data either. However, in the subpopulation of children 12 to 35 months old, exploratory analyses showed an elevated RR for medically-attended fever among the children receiving DTaP-IPV-Hib versus the comparator vaccine after adjusting for sex, site, and age group (OR = 1.75; CI = 1.38, 2.22) [39]. We re-analyzed this same subpopulation, which consisted of 248,923 children who received a DTaP-containing vaccine, 19,264 of whom received the new DTaP-IPV-Hib combination vaccine, to compare the four methods described in Chapter 3. We implemented all four methods using three different sequential designs. After one year of monitoring, there were only two cases of medically-attended fever among DTaP-IPV-Hib vaccine recipients. Thus, in each design the first test was delayed for 12 months due to the relatively slow uptake of the vaccine and the low prevalence of medically-attended fever. Thereafter, we performed analyses on either a quarterly, monthly, or weekly basis.

Tables 5.9, 5.10, and 5.11 show the results of our quarterly, monthly, and weekly analyses,

respectively. In each of these three designs, neither of the matched methods, GS LRT nor GS LD-M, detected a safety signal, and monitoring stopped at the pre-specified end of study (week 125). Both regression methods, GS LD-R and GS GEE, stopped monitoring early in all design scenarios. GS LD-R detected a safety signal sooner than GS GEE in each design. In the case of quarterly testing, GS LD-R detected a safety signal at week 78 and GS GEE at week 91. In the case of monthly testing, GS LD-R detected a safety signal at week 80 and GS GEE at week 88. In the case of weekly testing, GS LD-R detected a safety signal at week 78 and GS GEE at week 86.

Tables 5.9, 5.10, and 5.11 also show that the matched methods used only a small portion of the data compared to the regression methods, and as a result did not retain all of the observed AEs (either for DTaP-IPV-Hib or for comparative DTaP vaccines) in the set of informative matched pairs. For instance, in week 65 for quarterly testing, the informatively matched data set contained 13 of the 16 total observed DTaP-IPV-Hib AEs and 10 of the 280 total observed DTaP AEs, and by week 125, the informatively matched data set contained only 69 of the 112 total observed DTaP-IPV-Hib AEs and 55 of the 528 total observed DTaP AEs. When testing on a quarterly and monthly basis, the matched methods did not skip any analysis times. However, when we increased the testing frequency to weekly, these two methods skipped 17 of the 74 pre-specified analyses due to lack of additional informative pairs.

Table 5.9: (Quarterly testing schedule) Results of applying the four methods to a subset of data from the monitoring of a combination DTaP-IPV-Hib vaccine, where the first analysis time occurred after 12 months and testing was performed on a quarterly basis. Abbreviations: Match = matched subjects, Inform = informatively matched subjects, OR = odds ratio, Adj = adjusted

		DTaP-IPV-Hib recipients			AEs among subjects receiving			
					DTaP-IPV-Hib		DTaP	
Week	N	Total	Match	Inform	Total	Inform	Total	Inform
52	109851	1448	1448	7	2	2	235	5
65	135198	4263	3978	23	16	13	280	10
78[a]	**158236**	**7147**	**6393**	**40**	**35**	**21**	**329**	**19**
91	**182707**	**10460**	**9190**	**58**	**55**	**31**	**375**	**27**
104	209205	13901	12159	80	77	45	441	35
117	235248	17346	15199	110	97	61	491	49
125	248923	19264	16878	124	112	69	528	55

						Safety Signal?			
Week	Matched OR	Adj OR	S_{GEE}, S_{LD-R}	S_{LRT}	S_{LD-M}	GS GEE	GS LD-R	GS LRT	GS LD-M
52	0.40	0.52	-1.28	0.00	-1.10				
65	1.30	1.47	1.21	0.20	0.62				
78	**1.11**	**1.72**	**2.35**	**0.05**	**0.32**		**signal**		
91	**1.15**	**1.75**	**2.96**	**0.14**	**0.52**	**signal**			
104	1.29	1.74	3.43	0.63	1.12				
117	1.24	1.73	2.96	0.66	1.14				
125	1.25	1.75	4.10	0.79	1.25			end	end

[a]Bold numbers indicate analysis times at which one of the methods detected a safety signal

Table 5.10: (Monthly testing schedule) Results of applying the four methods to a subset of data from the monitoring of a combination DTaP-IPV-Hib vaccine, where the first analysis time occurred after 12 months and testing was performed on a monthly basis. Abbreviations: Match = matched subjects, Inform = informatively matched subjects, OR = odds ratio, Adj = adjusted

		DTaP-IPV-Hib recipients			AEs among subjects receiving			
					DTaP-IPV-Hib		DTaP	
Week	N	Total	Match	Inform	Total	Inform	Total	Inform
52	109851	1448	1448	7	2	2	235	5
56	118139	2349	2300	14	8	8	245	6
60	126280	3229	3072	20	12	12	265	8
64	133288	4054	3786	24	15	14	276	10
68	139730	4831	4468	26	17	16	287	10
72	146774	5714	5217	27	23	17	298	10
76	154281	6688	6006	29	30	18	317	11
80[a]	**162178**	**7648**	**6802**	**38**	**37**	**22**	**338**	**16**
84	169689	8636	7647	43	41	25	352	18
88	**177552**	**9726**	**8581**	**52**	**50**	**30**	**363**	**22**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
116	233069	17061	14942	106	94	60	486	46
120	240203	18059	15806	113	103	66	501	47
125	248923	19264	16872	122	112	71	528	51

						Safety Signal?			
Week	OR	Adj OR	S_{GEE}, S_{LD-R}	S_{LRT}	S_{LD-M}	GS GEE	GS LD-R	GS LRT	GS LD-M
52	0.40	0.52	-1.28	0.00	-1.10				
56	1.33	1.36	0.71	0.14	0.53				
60	1.50	1.48	1.04	0.40	0.89				
64	1.40	1.46	1.14	0.33	0.81				
68	1.60	1.39	1.08	0.70	1.17				
72	1.70	1.55	1.60	0.92	1.33				
76	1.64	1.62	2.02	0.85	1.29				
80	**1.38**	**1.65**	**2.27**	**0.48**	**0.97**		**signal**		
84	1.39	1.61	2.28	0.57	1.06				
88	**1.36**	**1.73**	**2.78**	**0.62**	**1.10**	**signal**			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
116	1.30	1.69	3.58	0.93	1.36				
120	1.40	1.77	4.01	1.60	1.78				
125	1.39	1.75	4.10	1.65	1.80			end	end

[a]Bold numbers indicate analysis times at which one of the methods detected a safety signal

Table 5.11: (Weekly testing schedule) Results of applying the four methods to a subset of data from the monitoring of a combination DTaP-IPV-Hib vaccine, where the first analysis time occurred after 12 months and testing was performed on a weekly basis. Abbreviations: Match = matched subjects, Inform = informatively matched subjects, OR = odds ratio, Adj = adjusted

		DTaP-IPV-Hib recipients			AEs among subjects receiving			
					DTaP-IPV-Hib		DTaP	
Week	N	Total	Match	Inform	Total	Inform	Total	Inform
52	109851	1448	1448	7	2	2	235	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
76	154281	6688	5979	(skip)	30	(skip)	317	(skip)
77	156255	6938	6186	35	32	20	322	15
78[a]	**158236**	**7147**	**6365**	**38**	**35**	**22**	**329**	**16**
79	160195	7403	6584	40	36	22	334	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
85	171616	8865	7839	(skip)	41	(skip)	356	(skip)
86	**173610**	**9145**	**8072**	**49**	**45**	**27**	**358**	**22**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
124	247074	18988	16573	112	112	61	523	51
125	248923	19264	16818	113	112	61	528	52

						Safety Signal?			
Week	OR	Adj OR	S_{GEE}, S_{LD-R}	S_{LRT}	S_{LD-M}	GS GEE	GS LD-R	GS LRT	GS LD-M
52	0.40	0.52	-1.28	0.00	-1.1				
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
76	(skip)	1.62	2.02	(skip)	(skip)				
77	1.33	1.65	2.12	0.36	0.84				
78	**1.38**	**1.72**	**2.35**	**0.48**	**0.97**		**signal**		
79	1.22	1.66	2.27	0.20	0.63				
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
85	(skip)	1.59	2.21	(skip)	(skip)				
86	**1.23**	**1.69**	**2.56**	**0.26**	**0.71**	**signal**			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
124	1.20	1.78	4.22	0.45	0.94				
125	1.17	1.75	4.10	0.36	0.85			end	end

[a]Bold numbers indicate analysis times at which one of the methods detected a safety signal

Chapter 6

DISCUSSION

In this thesis we performed a simulation study and an applied data analysis to compare four methods for observational group sequential safety monitoring of rare AEs. Two of the methods accounted for confounding using regression and two used one-to-one exposure matching on the PS. We showed via simulation that the regression methods were more powerful and stopped monitoring sooner than the matched methods. Overall, GS GEE consistently had the highest power, had the shortest time to study end, and held the type I error rate better than the other three methods. In the application of the methods to a subset of data from the VSD's prior study that monitored the safety of a DTaP-IPV-Hib vaccine, both the regression methods detected safety signals relatively early in monitoring when testing was done on a either a quarterly, monthly, or weekly basis after an initial delay of 12 months. In contrast, neither of the matched methods detected a safety signal in any of the sequential designs we used. These results highlight a number of aspects worth consideration. First, the approach chosen for confounder control impacts the amount of data that is used to compute the test statistic, which in turn affects the power and time-to-surveillance-end. Specifically, restricting data to a matched sample when AEs are rare often yields relatively few informative pairs, which lowers power. Second, the strength and type of confounding by site, as quantified by different exposure and AE prevalences at each site where monitoring occurs, impacts the power and time-to-surveillance-end of each method. Third, as AEs become rarer it becomes increasingly important to consider, during the design phase of safety monitoring, the practicalities of method implementation.

When comparing the regression methods to the matched methods, we found that the regression methods had an advantage. Overall, GS GEE was the most powerful method, followed by GS LD-R, GS LD-M, and GS LRT. The loss of power for the matched methods is likely due to the fact that one-to-one exposure matching on the PS and on site reduced

the (informative) sample size drastically. Such a reduction was not surprising since AEs were rare and it is only the informative pairs (i.e., those discordant on AE status) that are used to compute the test statistics for the matched methods. This reduction in sample size explains the loss of power for GS LRT and GS LD-M compared to GS GEE and GS LD-R. The loss of power for the GS LD methods is also likely due in part to the violation of the asymptotic and independent increments assumptions on which these methods rely. Also, among regression methods we found that GS LD-R tended to be more conservative than GS GEE. Similarly, among matched methods, GS LD-M tended to be more conservative than GS LRT. This behavior has been seen and discussed previously [16]. That the GS LD methods were outperformed by GS GEE and GS LRT is not surprising, since both GS GEE and GS LRT use stopping boundaries that are designed for the rare AE setting and are thus more robust in this way.

In addition to lower power, the matched methods had other limitations. As the application demonstrated, infrequent testing benefited matched methods because matching occurs within analysis times. As testing occurs more frequently, there are fewer new subjects at each analysis who are available for matching, and in the application this caused a number of analysis times to be skipped (e.g., when the testing frequency was as high as weekly). Thus, the desired testing frequency should be considered when choosing how to control for confounding in observational sequential safety monitoring. In particular, if frequent testing is a high priority, then the use of matching of any form to control for confounding may not be an ideal option. In our simulations we tested on a quarterly basis, which is rather infrequent for observational sequential safety monitoring standards. Further simulation study to explore how testing frequency affects method performance is needed to better understand the limit on testing frequency for matched methods. Another implementation choice is the type of matching. We performed one-to-one exposure matching on site and the PS, where three covariates contributed to the PS. Matching jointly on individual covariates (e.g., gender, age group, and site) instead of on the PS could potentially provide tighter confounding control and may perform similarly or better, and this warrants further study. Finally, we chose a simple and practical matching approach (one-to-one matching) that has been used in safety monitoring within the VSD and is similar to methods being proposed for use in the

Mini Sentinel pilot. More sophisticated matching approaches, such as one-to-m or variable ratio matching, may be worth studying. However, we note that it has been shown that for non-sequential testing, increasing the number of controls in a matched set does not substantially increase the power of the test [7], and we may see similar trends in the sequential setting with rare AEs. Inherent in any matching scheme are many specific choices that must be made in order to form matches. These choices impact the performance of the sequential monitoring methods used and deserve more study.

The impact of confounding by site on method performance was relatively comparable across methods, but complex. The most powerful and fastest detection site-AE scenarios varied depending on the site-exposure scenario. This demonstrates the intricate influence that the distributions of exposure and AE prevalences at different sites can have on the power and time-to-surveillance-end. For the equal-sized site scenarios (where the relationship between site and exposure was held constant), we can think about the differences in power between the site-AE scenarios in terms of how evenly the AEs were spread between the sites. For instance, when $P(Y|X = 0, \bar{Z}) = 0.05$, the set of simulations where two sites had lower AE prevalence than the other three was the most powerful likely because AEs were most evenly spread across the sites compared with other settings. When $P(Y|X = 0, \bar{Z}) = 0.01$, the power for each method across the site-AE scenarios was similarly high and thus more difficult to determine which site-AE settings were the most powerful. Further simulation studies that evaluate settings with slightly lower power would be helpful to better elucidate trends. In both AE prevalence settings, the scenarios with the shortest time-to-surveillance-end and most power generally coincided.

For the unequal-sized site scenarios, the most powerful methods again tended to be the ones with shortest time-to-surveillance-end. Here, the differences in power can be understood first in terms of the amount of statistical information available at a given site (especially at the small sites), and second, in terms of the direction of the relationships between site and exposure and between site and AEs. In both AE prevalence settings, the site-exposure scenario that was least powerful on average was the one where one large site had higher exposure prevalence (the other two site-exposure scenarios had comparable power to each other). This is reasonable because, when one large site had higher exposure prevalence,

this inherently took exposed subjects away from the other four sites, which destabilizes the other sites, particularly the small ones. The most powerful site-AE scenarios depended on the site-exposure scenario. For instance, in the more prevalent AE setting where one large site had higher exposure prevalence, the set of simulations where one large site had higher AE prevalence was the most powerful out of the three site-AE scenarios because both exposure and outcome were being concentrated in the sites. Both of the other site-AE scenarios (in this particular site-exposure scenario) had two sites that differed in terms of AE prevalence and they achieved similar power. However, the least powerful of the site-AE scenarios within this same site-exposure scenario where one large site had higher exposure prevalence was the one where the exposure and AE prevalences acted in opposite directions at one of the large sites (exposure prevalence increased while the AE prevalence decreased). This potentially resulted in a reduction of statistical information for the matched methods and difficulties in fitting the regression models for the regression methods, resulting in a loss of power and increase in time-to-surveillance-end. We saw similar trends in the other site-exposure scenarios.

Both the simulation study and the application highlighted important aspects of implementing sequential monitoring methods that are important to consider, particularly when matching in a setting where exposure is emerging and where AEs are rare. In the simulation study, we found that when AEs were rarer, lack of statistical information required us to skip analyses for both the regression and matched methods at the outset of surveillance. Since we match subjects who accrue in the same time period, the matched methods skipped additionally at intermediate analyses and thus skipped more looks than the regression methods. In addition, in the application we saw that as the testing frequency increased to weekly, the matched methods began to skip analyses. This is an inherent challenge for any method that uses matching, since subjects are matched within analysis times, and suggests that matching may not be an appropriate choice for confounder control when uptake is slow, AEs are rare, and testing is highly frequent. Furthermore, matching restricted the informative sample size considerably and this resulted in a loss of power compared to the regression methods. We saw this in our simulation study where the number of people belonging to informative matches was between 3 and 11% of the total number of people who were matched. In our

analyses of concurrent control data from the VSD's DTaP-IPV-Hib safety study, the regression methods both stopped surveillance early regardless of testing frequency (e.g., weekly, monthly, or quarterly), while the matched methods did not stop early in any of the testing frequency situations. One possible reason for this is that when doing sequential monitoring, emerging exposure uptake, such as with the DTaP-IPV-Hib vaccine, greatly reduces the amount of informative data (unexposed participant data in particular) early in monitoring and yields an overall reduction in power across all analyses. Emerging exposure uptake has less of an impact on the regression methods.

Based on our simulations and application, matched methods may not be a reliable choice for observational safety monitoring of rare AEs. While we have begun to elucidate important aspects of observational sequential safety monitoring using regression versus matching to control for confounding, many areas would benefit from further study. This includes further study of confounding scenarios at equal-sized sites, the frequency of testing, the matching ratio, and the criteria used for matching (e.g., matching on the PS or jointly on individual covariates). Another matched approach potentially worth considering is the use of a case-control matched design instead of exposure matching. This type of matching would ensure retention in our matched data set of all matched subjects who experienced an AE and is standardly employed in non-sequential observational studies of rare AEs. Thus it could also make good sense in a sequential safety monitoring setting where AEs are rare. Case-control matching could be implemented similarly to the approaches used in this thesis using conditional logistic regression. As we have described, there is still much to be learned about the performance of sequential methods for observational safety surveillance in the rare AE setting.

REFERENCES

[1] Food and drug administration amendments act (fdaaa) of 2007, 2007.

[2] *FDA's Mini-Sentinel Program to evaluate the safety of marketed medical products*, 2012.

[3] P Armitage, CK McPherson, and BC Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society A*, 132(2):235–244, 1969.

[4] PC Austin. Propensity-score matching in the cardiovascular surgery literature from 2004-2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, 134:1128–1135, 2007.

[5] PC Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27:2037–2049, 2008.

[6] PC Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, pages 661–677, 2009.

[7] PC Austin. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*, 172(9), 2010.

[8] PC Austin. Optimal caliper widths for propensity-score matching when estimated differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10:150–161, 2011.

[9] J Baggs, J Gee, E Lewis, and et al. The vaccine safety datalink: a model for monitoring immunization safety. *Pediatrics*, 127:S45–S53, 2011.

[10] RE Behrman, JS Benner, JS Brown, M McClellan, J Woodcock, and R Platt. Developing the sentinel system - a national resource for evidence development. *New England Journal of Medicine*, 364:498–499, 2011.

[11] EA Belongia, S Irving, IM Shui, and et al. on behalf of the Vaccine Safety Datalink Investigation Group. Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. *Pediatr Infect Dis J*, 29(1):1–5, 2010.

[12] N Breslow. Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, 38:661–672, 1982.

[13] C Casper and OA Perez. *Lan-DeMets method for group sequential boundaries*. R Foundation for Statistical Computing, 2006.

[14] RT Chen, F DeStefano, RL Davis, and et al. The vaccine safety datalink: immunization research in health maintenance organizations in the usa. *Bulletin of the World Health Orgnization*, 78(2):186–194, 2000.

[15] RT Chen, F DeStefano, RL Davis, LA Jackson, Thompson RS, JP Mullooly, SB Black, HR Shinefield, CM Vadheim, JI Ward, and SM Marcy. The vaccine safety datalink: immunization research in health maintenance organizations in the usa. *Bulletin of the World Health Organization*, 78:186–194, 2000.

[16] AJ Cook, Wellman R, JC Nelson, LA Jackson, and RC Tiwari. Group sequential method for observational data incorporating confouning with application in postmarketing vaccine safety datalink (vsd) project. *Biometrics (submitted)*, 000:000, 2012.

[17] AJ Cook, R Wellman, JC Nelson, LA Jackson, and RC Tiwari. Group sequential method for observational data incorporating confounding with application in postmarketing vaccine safety datalink (vsd) project. *Biometrics (submitted)*, 000:000–000, 2012.

[18] AJ Cook, RD Wellman, T Marsh, L Li, S Heckbert, P Heagerty, RC Tiwari, and JC Nelson. Statistical approaches to group sequential monitoring of post-marketing safety surveillance data: Current state of the art for use in the mini-sentinel surveillance project. *Pharmacoepi and Drug Safety*, 21:72–81, 2012.

[19] RB D'Agostino. Tutorial in biostatistics: Propensity score methods for bias reducation in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–2281, 1998.

[20] RL Davis, M Kolczak, E Lewis, and et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*, pages 16:336–341, 2005.

[21] F DeStefano. The vaccine safety datalink project. *Pharmacoepidemiol Drug Safety*, 10(5):403–406, 2001.

[22] SS Emerson. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics*, 59:770–777, 2003.

[23] SS Emerson and TF Fleming. Symmetric group sequential test designs. *Biometrics*, 45:905–923, 1989.

[24] SS Emerson, JM Kittelson, and DL Gillen. Frequentist evaluation of group sequential clinical trials. *Statistics in Medicine*, 26:5047–5080, 2007.

[25] Centers for Disease Control and Prevention. Licensure of a diptheria and tetanus toxoids and acellular pertussis adsorbed, inactivated poliovirus, and haemophilus b conjugate vaccine and guidance for use in infants and children. *MMWR Morb Mortal Wkly Rep*, 57:1079–1080, 2008.

[26] J Gee, A Naleway, I Shui, and et al. Monitoring the safety of quadrivalent human papillomavirus vaccine: Findings from the vaccine safety datalink. *Vaccine*, 29(46):8279–8284, 2011.

[27] C Jennison and BW Turnbull. Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

[28] C Jennison and BW Turnbull. *Group sequential methods with applications to clinical trials*. Boca Raton, FL: CRC Press, 2000.

[29] JM Kittelson and SS Emerson. A unifying family of group sequential test designs. *Biometrics*, 55(3):874–882, 1999.

[30] NP Klein, B Fireman, W Yih, and et al. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. *Pediatrics*, 126:e1–e8, 2010.

[31] M Kulldorff, RL Davis, M Kolczak, and et al. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*, 30:58–78, 2011.

[32] KKG Lan and DL DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[33] GM Lee, SK Greene, ES Weintraub, and et al. on behalf of the Vaccine Safety Datalink Project. H1n1 and seasonal influenza vaccine safety using prospective weekly surveillance in the vaccine safety datalink project. *Am J Prev Med*, 41(2):121–128, 2011.

[34] L Li. A conditional sequential sampling procedure for drug safety surveillance. *Statistics in Medicine*, 28(25):124–138, 2009.

[35] K-Y Liang and SL Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 78:13–22, 1986.

[36] TA Lieu, M Kulldorff, RL Davis, and et al. Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care*, 45 (10 Suppl 2):S89–S95, 2007.

[37] JC Nelson, AJ Cook, O Yu, and et al. Challenges in the design and analysis of sequentially-monitored postmarket safety surveillance evaluations using electronic observational health care data. *Pharmacoepi Drug Safety*, 21:62–71, 2012.

[38] JC Nelson, AJ Cook, O Yu, S Zhao, LA Jackson, and BM Psaty. Methods for observational post-licensure medical product safety surveillance. *Statistical Methods in Medical Research*, 2011.

[39] JC Nelson, O Yu, C Dominguez, AJ Cook, D Peterson, SK Greene, K Yih, MF Daley, SJ Jacobsen, NP Klein, E Weintraub, and LA Jackson. Adapting surveillance: Results of a pentavalent combination dtap-ipv-hib (pentacel) vaccine safety study. *000*, 000:000, 2012.

[40] PC O'Brien and TR Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[41] SJ Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[42] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[43] A Rotnitzky and NP Jewell. Hypothesis-testing of regression parameters in semiparametric generalized linear-models for cluster correlated data. *Biometrika*, 77(3):485–497, 1990.

[44] JS Sekhon. Multivariate and propensity score matching with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52, 2011.

[45] BR Shah, A Laupacis, JE Hux, and PC Austin. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58:550–559, 2005.

[46] MC Shih, TL Lai, JF Heyse, and J Chen. Sequential generalized likelihood ratio tests for vaccine safety evaluation. *Statistics in Medicine*, 29:2698–2708, 2010.

[47] A Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[48] SK Wang and AA Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43(1):193–199, 1987.

[49] J Whitehead. *The design and analysis of sequential clinical trials*. New York: Wiley, 1997.

[50] WK Yih, JD Nordin, M Kulldorff, and et al. An assessment of the safety of adolescent and adult tetanus-diphtheria-acellular pertussis (tdap) vaccine, using near real-time surveillance for adverse events in the vaccine safety datalink. *Vaccine*, 27:4257–4262, 2009.

[51] SL Zeger and K-Y Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

[52] S Zhao, AJ Cook, LA Jackson, and JC Nelson. Statistical performance of group sequential methods for observational post-licensure medical product safety surveillance: a simulation study. *000*, 000:000, 2012.

Appendix A

R CODE FOR GS LRT METHOD

```
##############################################################################
### CODE FOR THE GROUP SEQUENTIAL LIKELIHOOD RATIO TEST (GS LRT) FOR ###
### GROUP SEQUENTIAL POST-LICENSURE SAFETY SURVEILLANCE              ###
###                                                                  ###
### By: Kelly Stratton                                               ###
### Date: May 2012                                                   ###
##############################################################################

##############################################################################
### Function that wraps the GSLRT method together
  # Called by: simulation_code
  # Inputs:
      # mydata <- result of call to gen_data, dataframe (n by 6), with
             # columns for the analysis time, sex, age, site, outcome,
             # eandxposure status
      # perms <- number of permutations to generate for getting dist'n
             # of LRTs
      # alpha <- type I error rate, overall
      # d <- number of looks
      # matchtype <- "exact" or "propensity"
      # looks <- the numbers of the looks (e.g. 1, 2, 4, 6, 8 if we
             # skipped analysis times 3, 5, and 7)
  # Outputs:
      # SampleSize <- sample size after matching
```

```
      # c <- signaling threshold
      # t1e <- sum(llr.max>c)/perms = percentage of llr.max that exceed
            # the 95th percentile of llr.max
      # ObservedLLR <- llr statistic for the actual data (not permuted)
      # stop.trial <- logical vector (1 by d) of whether the test
            # statistic at that analysis time exceeded c
do_one_gslrt <- function(mydata.matched, perms, alpha, d=NULL, looks){
  # permute the data
  mydata.permed <- perm_data(mydata.matched, looks, perms)

  # GET C GIVEN ALPHA
  d <- length(unique(mydata.matched$i.look))
  temp <- get_c(mydata.permed, alpha, d)
  c <- temp[1]
  t1e <- temp[2]
  # get llr for actual data, compare to c
  actual.data <- mydata.matched[, 1:7]
  colnames(actual.data) <- c("stratum", "i.look", "sex", "age", "site",
      "y", "x")
  actual.llr <- get_llr_matching(actual.data,
      d = length(unique(actual.data$i.look)))

  stop.trial <- (actual.llr > c); #stop.trial
  # if any of the stop.trial's are NA's (which happens if we skipped a
       # look), then change the NA's to FALSE
  if(sum(is.na(stop.trial)) > 0){
    stop.trial[which(is.na(stop.trial))] <- FALSE
  }

  if(any(stop.trial == TRUE)){     # we stopped early
```

```
    lookstop <- min(which(stop.trial == TRUE))
    daystop <- lookstop * 365.25 * 2 / d
    stop.early <- TRUE
  }
  if(!any(stop.trial == TRUE)){   # we did not stop early
    daystop <- 365.25 * 2
    lookstop <- d
    stop.early <- FALSE
  }

  samplesize <- dim(mydata.matched)[1]
  obs.stat <- actual.llr[lookstop]

  gslrt.out <- c(samplesize, as.numeric(c), obs.stat, stop.early,
      daystop, t1e)
  names(gslrt.out) <- c("sample.size", "c", "obs.stat.at.stop",
      "stop.early.YN", "stop.time", "t1e")
  return(gslrt.out)
}

##########################################################################
### Function that gets all the permutations
  # Called by: do_one_gslrt
  # Inputs:
      # mydata.new <- data to permute (informative pairs only)
      # perms <- number of permutations
      # looks <- the timing of the looks
  # Output:
      # dat <- dataframe (n by 6+number of permutations), dat[, 1:6] =
          # mydata.new, dat[, 7:end] = columns of permuted exposures
```

```
perm_data <- function(mydata.new, looks, perms){
  # get permutations (within each match)
  dat <- mydata.new
  for(k in 1:perms){
    permed.exposure <- NULL
    for(i in looks){
      temp.data <- mydata.new[mydata.new$i.look == i, ]
      # re-number the stratum column (if we skipped looks, then multiple
         # pairs can have the same stratum indices, which causes an error
         # with the sampling command below)
      n.pairs <- dim(temp.data)[1] / 2
      temp.stratum <- rep(1:n.pairs, each = 2)
      temp.data$stratum <- temp.stratum
      test <- by(temp.data$exposed, temp.data$stratum, sample, size = 2,
          replace = FALSE , prob = c(.5, .5))
      temp <- NULL
      for(j in 1:length(test)){
        temp <- c(temp, t(as.data.frame(test[j])))
      }
      permed.exposure <- c(permed.exposure, temp)
    }
    dat <- as.data.frame(cbind(dat, permed.exposure))
  }
  return(dat) # an (n.informative.subjects) by (perms+17) data frame
}


#########################################################################
### Function that computes the signaling threshold, c, based on the
  #    desired alpha level
```

```
  # Called by: do_one_gslrt
  # Inputs:
      # mydata <- data frame
      # alpha <- desired type I error
      # d <- number of looks
  # Outputs:
      # c <- signaling threshold
      # t1e <- sum(llr.max>c)/perms = percentage of llr.max that exceed
          # the 95th percentile of llr.max
      #       [should be 0.05]
get_c <- function(mydata.permed, alpha, d){
  n <- dim(mydata.permed)[1]
  perms <- dim(mydata.permed)[2] - 17 # subracting 17 for the exposure,
        # outcome, confounder (3), stratum, look, and n.matched cols
  x.permed <- mydata.permed[,18:(perms+17)] # add 17 to get all perms
  y <- mydata.permed$outcome

  #Initialize
  llr <- array(NA, dim=c(perms, 8)) # matrix to hold the log likelihood
     # ratios for each perm at each look

  for(i in 1:perms){  # look at one perm at a time
    ithdata <- as.data.frame(cbind(mydata.permed[,1:4], y, x.permed[, i]))
    names(ithdata) <- c("stratum", "i.look", "sex", "age", "outcome",
        "exposed")

    colnames(ithdata) <- c("stratum", "i.look", "sex", "age", "y", "x")
    d <- length(unique(ithdata$i.look))
    temp <- get_llr_matching(ithdata, d)
    llr[i,] <- c(temp, rep(-1, times = 8 - length(temp)))
```

```
  }
  # get the max of the LLRs, then take 95th percentile to get c
  llr.max <- apply(llr, 1, max, na.rm=TRUE)
  c <- quantile(llr.max, .95, na.rm=TRUE)
  t1e <- sum((llr.max > c), na.rm=TRUE) / perms

  return(c(c, t1e))
}

#########################################################################
### Function that computes the LLR for the current exposure and outcome
  #    when the data are matched
  # Called by: get_c, in do_one_gslrt
  # Inputs:
      # current.data <- data including all looks for informative matches
         # only, for the current perm that we're getting the llr for
      # d <- number of looks
  # Output:
      # llr <- vector (1 by d) of llr test statistics, one for each look
get_llr_matching <- function(current.data, d){
  n <- length(current.data$x)
  llr <- array(NA, dim = c(1, d))
  looks <- unique(current.data$i.look)
  for(i in looks){
    current.look <- current.data[current.data$i.look <= i, ]
    k <- dim(current.look)[1] / 2 # num pairs w/single event
    # num pairs for which event is among the exposed person
    sumx <- sum(current.look$x == 1 & current.look$y == 1)
    xbar <- sumx / k
```

```
    if(xbar > .5){  # to make it a one-sided test
      llr[i] <- k * (xbar * log(xbar) + (1-xbar) * log(1-xbar) + log(2))
    }
    else{llr[i] <- 1}
  }
  return(llr)
}
```

Appendix B

R CODE FOR GS GEE METHOD

```
#############################################################################
### CODE FOR THE GROUP SEQUENTIAL GENERALIZED ESTIMATING EQUATIONS   ###
###    (GS GEE) METHOD FOR (COVARIATE ADJUSTED)                      ###
###     GROUP SEQUENTIAL POST-LICENSURE SAFETY SURVEILLANCE          ###
### By: Andrea J Cook                                  ###
#############################################################################

#############################################################################
### Function to run the GS EE Approach
  # Called by: method_wrapper.r
  # Inputs:
      # mydata <- result of call to gen_data, dataframe (n by 6), with
           # columns for the analysis time, sex, age, site, outcome,
           # and exposure status
      # alpha <- total type I error desired to spend across all looks
      # delta <- shape parameter for the boundary (delta=.5 is Pocock)
                # (delta=0 is Fleming)
      # nsim <- Number of simulations to calculate the boundary
                # (use at least 1000, but higher the better)
  # Outputs:
      # gsee.out <- a vector with the samplesize , c, ObservedScT (at
          # signal or end of study), stop (0=no, 1=yes),
          # stoptime=sigTime (at or end of study)
do_one_gsee <- function(mydata,delta=0.5, alpha=0.05, nsim=1500){
```

```
Y <- mydata[, 5]
X <- mydata[, 6]
Z <- mydata[, c(2:3)]
#Add Site as categorical "dummy" variables
for(s in 1:4)
{
Z <- cbind(Z, as.numeric(mydata[, 4]==s))
}
S <- mydata[, 1]
ATime <- unique(S)
Y <- as.matrix(Y[order(S)])
X <- X[order(S)]
Z <- as.matrix(Z)
Z <- Z[order(S), ]
S <- S[order(S)]
Zint <- cbind(1, Z)

# Create a vector with amount of sample size observed up to time t
Ncum <- NULL
for(T in 1:length(ATime))
{
Ncum <- c(Ncum,sum(S <= ATime[T]))
}

betaZ <- NULL
for(T in 1:length(ATime))
{
betaZ <- cbind(betaZ, glm(Y[1:Ncum[T]] ~ Z[1:Ncum[T], ],
        family=binomial)$coef)
}
```

```
#Skip looks that you could not estimate beta
bfound <- apply(is.na(betaZ), 2, sum)  # 0's if no problems,
                    # 1's if beta couldn't be estimated at that look
betaZ <- betaZ[, bfound==0]
ATime <- ATime[bfound==0]
#propinfo<-propinfo[bfound==0] (don't need this since pocock)
Ncum <- Ncum[bfound==0]

  # Get critical boundary values
bd <- seqEECovAdjUnif.boundary.bin(Y=Y, X=X, Z=Z, S=S, Ncum=Ncum,
      delta=delta, betaZ=betaZ, alpha=alpha, nsim=nsim)
sig <- 0
sigtime <- max(ATime)
ScT <- NULL

boundScTt <- NULL
boundESt <- NULL
ATimet <- NULL  # analysis time
sigt <- NULL    # signal time
T <- 1
while(T <= length(ATime))
{
  # compute the "current" score test statistic
ScTcur <- ScTt.bin(Y=Y[1:Ncum[T]], X=X[1:Ncum[T]], Z=Z[1:Ncum[T],],
        betaZ=betaZ[,T])

    # put in into a vector that will contain all the score statistics
    # (one for ea look)
ScT <- c(ScT, ScTcur)
```

```
boundScTt <- c(boundScTt, bd[T])
ATimet <- c(ATimet, ATime[T])
# compare the score statistic to the critical boundary
sigt <- c(sigt, as.numeric(ScTcur > boundScTt[T]))
if(ScTcur > boundScTt[T])   # if there's a signal
{
sig <- 1
sigtime <- ATime[T]
T <- length(ATime)+1
}
T <- T+1
 }

  d <- max(mydata$i.look)
  if(sig == 1){
    stoptime = sigtime * 365.25 * 2 / d
  }
  if(sig == 0){
    stoptime = 365.25 * 2 # Kelly edited this line
  }
gsee.out <-c (sample.size=length(Y), c=bd[1],
      obs.stat.at.stop=ScT[length(ScT)], stop.early.YN=sig,
      stop.time=stoptime)
return(gsee.out)
}

#############################################################################
### Function to calculate score statistic
  # Called by: do_one_gsee
  # Inputs:
```

```
      # X <- a vector indicating if on exposure of interest
      # Y <- a vector or matrix of outcomes for look t
      # Z <- a matrix of confounders for X at look t
  # Outputs:
      # ScT <- the value of the Score Statistic at each look statistic
ScTt.bin <- function(Y, X, Z, betaZ=NULL)  ## Case 1: Single Exposure Time
{
Y <- as.matrix(Y)
Zint <- cbind(1, Z)

if(is.null(betaZ))
{
betaZ <- glm(Y ~ Z, family=binomial)$coef
}
n <- length(Y)
mui <- as.vector(exp(Zint %*% betaZ) / (1 + exp(Zint %*% betaZ)))
ei <- as.vector(Y - mui)
XZ <- cbind(X, Zint)
Ui <- XZ * ei
U <-t (rep(1, length(Y))) %*% Ui
W <- ginv(t(XZ) %*% (XZ * mui * (1 - mui)))
V <- W %*% (t(Ui) %*% Ui) %*% W
# Standardized Score Statistic
ScT <- U %*% W[, 1] / sqrt(V[1, 1])
ScT[(is.na(t(X) %*% ei))] <- 0

return(ScT)
}

#########################################################################
```

```
### Function to find boundary on the score statistic given unifying
  #     boundary function
  # Called by: seqEECovAdjUnif.boundary.bin
  # Inputs:
      # Ncum <- a vector of the cumulative number of people to be
            # observed at each look
      # X <- a vector of indicator of exposure data observed at each
            # look up to current look
      # Z <- a vector of confounder data observed at each look up to
            # current look
      # S <- a vector of the start times observed up to current look
      # betaZ <- a vector of confounder coefs under Ho at each look
  # Outputs:
      # maxstatY <- a vector of the critical values at each look t
do.Cr.bin <- function(Ncum, Y, X, Z, S, delta=0.5, betaZ)
{
T <- length(Ncum)
N <- Ncum[T]
Ncum0 <- c(1, Ncum)

ScTY <- NULL
Xcur <- NULL
for(i in 1:T)
{
Xcur <- c(Xcur, sample(X[Ncum0[i]:Ncum0[i + 1]], Ncum0[i + 1] -
Ncum0[i] + 1, replace = F))
ScTY <- c(ScTY, ScTt.bin(Y=Y[1:Ncum[i]], X=Xcur[1:Ncum[i]],
Z=Z[1:Ncum[i], ], betaZ=betaZ[, i]))
}
```

```
maxstatY <- max(ScTY, na.rm=T)
return(maxstatY)
}

#############################################################################
### Function to obtain critical boundary values
  # Called by: do_one_gsee
  # Inputs:
      # Y <- a vector of observed outcomes up to time t
      # X <- a vector of the current indicator of exposure data observed
           # up to time t
      # Z <- a vector of the confounders observed up to time t
      # S <- a vector of the start times observed up to time t
      # Ncum <- a vector of the cumulative number of people to be
           # observed at each look
      # betaZ <- a vector of confounder coefficients under Ho at ea look
      # nsim <- Number of simulations to calculate the boundary
  # Outputs:
      # maxstatY <- a vector of the critical values at each look t
seqEECovAdjUnif.boundary.bin <- function(Y, X, Z, S, Ncum, delta=0.5,
    betaZ, alpha=0.05, nsim=1000)
{
temp <- replicate(nsim, do.Cr.bin(Ncum=Ncum, Y=Y, X=X, Z=Z, S=S,
      delta, betaZ=betaZ))

qtail <- quantile(temp, 1 - (alpha), na.rm=T, type=9)
return(CV = rep(qtail, length(Ncum)))
}
```

Appendix C

R CODE FOR GS LD METHODS

```
#########################################################################
### CODE FOR THE GROUP SEQUENTIAL LAN-DEMETS (GS LD) METHOD FOR      ###
### GROUP SEQUENTIAL POST-LICENSURE SAFETY SURVEILLANCE              ###
###                                                                  ###
### By: Kelly Stratton                                               ###
### Date: March 2012                                                 ###
#########################################################################

# Relies on the libraries:
#library(survival)
#library(ldbounds)

#########################################################################
### Function that outputs all the things we want from GSLD methods at
  #    the end of the simulation
  # Called by: method_wrapper
  # Inputs:
      # mydata <- result of call to gen_data, dataframe (n by 6), with
          # columns for the analysis time, sex, age, site, outcome, and
          # exposure status
      # z <- standardized test statistic, the result of either
          # do_one_gsld_regr or do_one_gsld_match
      # possible.bounds <- list of ldbounds for different looks numbers
          # (1st element is for 1 look, 2nd element is for 2 looks, ...,
```

```
          # 8th element is for 8 looks)
  # Outputs:
      # gsld.out <- vector with: sample size at end of surveillance,
          # boundary value at end of surveillance, observed test statistic
          # at end of surveillance, whether or not surveillance was stopped
          # early (prior to 2 years), the time (in days) at which
          # surveillance ended
do_one_gsld <- function(mydata, z, samplesize=NULL, possible.bounds){
 n.in.looks <- by(mydata$i.look, mydata$i.look, length)
 looks <- length(unique(mydata$i.look))
 myldbounds <- possible.bounds[[looks]]

 stop.trial <- (z > myldbounds)
  if(any(stop.trial)){ # stopping early
    lookstop <- min(which(stop.trial))
    daystop <- (lookstop/length(n.in.looks))*365.25*2
    stop.early <- TRUE
  }
  if(!any(stop.trial)){ # not stopping early
    lookstop <- length(stop.trial)
    daystop <- 365.25*2
    stop.early <- FALSE
  }
  # test statistic value at last look done (i.e. when we stopped)
  obs.stat <- z[lookstop]
  # critical boundary value at last look done (i.e. when we stopped)
  c <- myldbounds[lookstop]

  gsld.out <- c(samplesize, c, obs.stat, stop.early, daystop)
  names(gsld.out) <- c("sample.size", "c", "obs.stat.at.stop",
```

```
      "stop.early.YN", "stop.time")
  return(gsld.out)
}

#############################################################################
### Function that outputs the standardized (score) test statistic for
  #   regression-based confounding control
  # Called by: method_wrapper in the function call for do_one_gsld
  # Inputs:
      # mydata = result of call to gen_data, dataframe (n by 6), with
          # columns for the analysis time, sex, age, site, outcome, and
          # exposure status
  # Outputs:
      # gsld.out = standardized (score) test statistic
do_one_gsld_regr <- function(mydata){
 looks <- unique(mydata$i.look)
 z <- rep(NA, length(unique(looks)))
 ind <- 1
 for(i in looks){
  Y <- mydata[mydata$i.look <= i, 5]  # outcome
  X <- mydata[mydata$i.look <= i, 6]  # exposed
  Z <- mydata[mydata$i.look <= i, c(2:3)]   # sex & age
  #Add Site
  for(s in 1:4){  # there are 5 sites total
    Z <- cbind(Z, as.numeric(mydata[mydata$i.look <= i, 4] == s))
  }

  z[ind] <- ScTt.bin(Y, X, as.matrix(Z))
  ind <- ind+1
 }
```

```
 return(z)
}

#############################################################################
### Function that outputs the standardized (wald) test statistic for
  #      matching-based confounding control
  # Called by: method_wrapper in the function call for do_one_gsld
  # Inputs:
      # mydata = result of call to gen_data, dataframe (n by 6), with
          # columns for the analysis time, sex, age, site, outcome, and
          # exposure status
  # Outputs:
      # gsld.out = standardized (Wald) test statistic
do_one_gsld_match <- function(mydata.matched){
  looks <- unique(mydata.matched$i.look)
  z <- rep(0, length(looks))
  ind <- 1
  for(i in looks){
    # re-number the strata, since each look starts counting at "1"
    mydata.matched.subset <- mydata.matched[mydata.matched$i.look <= i, ]
    stratum.new <- rep(1:(dim(mydata.matched.subset)[1] / 2), each=2)
    mydata.matched.subset$stratum <- stratum.new
    mymodel <- clogit(outcome ~ exposed + strata(stratum),
        data=mydata.matched.subset, method="exact")
    z[ind] <- summary(mymodel)$coefficients[4]
    ind = ind+1
  }
  return(z)
}
```