Statistics and Its Interface Volume 5 (2012) 381–390

Statistical performance of group sequential methods for observational post-licensure medical product safety surveillance: A simulation study

Shanshan Zhao, Andrea Cook, Lisa Jackson and Jennifer Nelson*

In order to improve post-licensure drug and vaccine safety surveillance, new systems are being developed that prospectively monitor observational health care data from large health plans. Continuous sequential testing has been proposed in this setting to facilitate rapid detection, but group sequential methods commonly used in randomized clinical trials (RCTs) have received less consideration. We propose a group sequential approach tailored for safety and to account for complications like confounding that arise in this non-randomized setting and thus have not been previously examined in RCTs. For comparability with prior continuous monitoring applications, we use a likelihood ratio statistic and historical controls. We compute sequential boundaries using Monte Carlo simulation and show how they can accommodate unequal between-test sample sizes and changes in confounder distributions among accruing subjects over time. We evaluate via simulation the performance of this approach across sequential designs suited for safety and not previously addressed by simulation studies evaluating RCT boundaries. Such designs include much higher frequency testing and designs that employ early conservatism followed by frequent testing. Contrary to prior RCT simulations, we found major differences in the average time-to-surveillance-end and overall power. We apply this methodology to safety data on a new pediatric combination vaccine.

Keywords and phrases: Medical product safety, Observational study, Post-licensure surveillance, Sequential testing.

1. INTRODUCTION

Pre-licensure clinical studies and traditional post-licensure safety surveillance systems often do not provide a complete safety profile for a newly licensed medical product. In fact, gaps in this safety evidence base have led to the approval and widespread use of some medical products that were later found to be unsafe and removed from the market. A well-known example is the drug rofecoxib (Brand Names: Vioxx or Ceoxx), a nonsteroidal antiflammatory medication prescribed to treat osteoarthritis and acute pain conditions. Rofecoxib was licensed by the US Food and Drug Administration (FDA) in 1999, prescribed to over 80 million people worldwide, and then removed from the market in 2004 after a colon cancer prevention trial designed to examine rofecoxib as a chemopreventative agent detected a doubling of cardiovascular risk among rofecoxib users [20]. High profile withdrawals, like that of rofecoxib, have increased public concern about the health risks of approved drugs and led to Congressional legislation and a national strategy aimed to detect safety problems in a more timely manner before large populations are exposed [3, 21].

One statistical methodology that has gained considerable traction for use in post-licensure safety surveillance is sequential testing. For example, the Centers for Disease Control and Prevention (CDC) Vaccine Safety Datalink (VSD) is a national project that weekly links and updates administrative patient information on demographics, immunizations, and adverse event diagnoses assigned during outpatient, emergency department, and hospital visits from eight managed care organizations (MCOs) covering 8.8 million people (3% of the U.S. population) [2]. With these data, the VSD has pioneered the use of a sequential monitoring approach called the maximized sequential probability ratio test (MaxSPRT) [10] to conduct near real-time surveillance for targeted safety outcomes of interest for newly licensed vaccines since 2005 [4, 6, 9, 13, 23]. MaxSPRT is a continuous sequential test based on the likelihood ratio statistic. Continuous monitoring using other sequential generalized likelihood ratio tests has also been recommended for use in post-licensure safety settings [19]. In the VSD, MaxSPRT has had demonstrated successes, including the identification of an increased risk of seizure among infants after receipt of the newly licensed measles, mumps, rubella, and varicella (MMRV) combination vaccine compared to age-similar infants who received separate injections of measles, mumps, and rubella (MMR) vaccine and varicella vaccine. This finding led to changes to the Advisory Committee on Immunization Practices' (ACIP) national policy recommendations [9].

Although continuous sequential testing appears promising and can facilitate early detection, such highly frequent testing may not be feasible or desirable in many practical circumstances. When more periodic interim testing is

*Corresponding author.

desired, group sequential methods can offer a flexible and powerful approach. And while group sequential testing is an established approach in randomized clinical trial (RCT) settings, it has received little consideration in the post-licensure safety monitoring arena. In RCTs, the number of sequential tests performed during a study is generally small (e.g., half-yearly over a 2-year study period), and the main study objective is to evaluate efficacy, which often implies use of a conservative stopping boundary (e.g., an O'Brien-Fleming boundary which is higher at early tests) [15]. In contrast, in observational post-licensure safety surveillance, researchers and policy-makers have advocated for much more frequent testing and a larger total number of tests (e.g., weekly over a 2-year study period). Further, the primary goal is to monitor rare and serious adverse events, which may motivate the use of a lower boundary at earlier tests to avoid missing a critical safety concern. Thus, although the statistical performance of many group sequential designs has been well-studied in RCT settings with efficacy endpoints, the performance of designs geared to address safety hypotheses in observational settings has not been systematically evaluated. For example, Pocock et al. (1982) [18] found there is little statistical advantage in undertaking more than five interim analyses during the course of a trial, since there is very little extra reduction in ASN (average sample number) in going from a five-group to a 20-group design. However, they did not consider the wide spectrum of designs that might be considered important in a safety setting, such as a design that employs early conservatism and then follows with very highly frequent testing. Providing such information could help researchers begin to develop a formal statistical framework for optimally designing group sequentially-monitored observational safety studies.

In this paper, we propose a group sequential approach tailored for safety and modified to account for the many complications that arise in this non-randomized post-licensure setting [14], complications that do not exist and thus have not been previously examined in RCTs. We conduct a simulation study that compares the statistical performance of these safety-customized group sequential designs with the performance of the currently used continuous sequential test MaxSPRT. Specifically, in Section 2 we briefly review existing sequential methods used both in RCTs and in observational studies, and we then describe how to extend the use of group sequential methods to an observational setting. This involves addressing a major challenge present in observational studies and not in RCTs: confounding. In Section 3, we conduct a simulation study to compare the statistical power and the time-to-detection of a safety problem of several continuous and group sequential stopping boundaries, including custom-designed choices that are specifically geared to address safety objectives. In Section 4 we apply and compare a variety of safety-tailored sequential designs using vaccine safety data from the VSD. We conclude in Section 5 with a discussion.

2. GROUP SEQUENTIAL TESTING METHODS

In both RCTs and observational studies it is desirable to examine the data over time as information accumulates instead of waiting until the end of the pre-defined study period to conduct formal statistical analyses. Sequential testing methods have been proposed to allow such interim analyses while taking into account the multiple testing issue by explicitly holding a desired Type I error across all tests. These methods can be broadly categorized into continuous and group sequential methods. In this manuscript we focus on group sequential testing and treat continuous testing as a special case, where the group is one observation. In the following sections we briefly summarize some existing group sequential methods that have been applied in both randomized and observational study designs with an emphasis on observational post-licensure surveillance applications.

2.1 Group sequential data and testing framework

Group sequential testing methods analyze data after a new group of observations enters the study at specific time points t ($t = 1, \ldots, T$), where T is the total number of tests, n_t is the sample size accrued between tests $t-1$ and t, and $N_t = \sum_{k=1}^{t} n_k$ is the cumulative sample size observed up to time t. In post-licensure safety surveillance, the purpose is to test for elevated rates of adverse outcomes among recipients of a drug or vaccine of interest relative to an expected rate. Expected rates may be estimated using concurrent control, historical control, or self-control information. To enhance comparability with previous post-licensure continuous monitoring applications [10, 19] and without loss of generality, we focus on a sequential likelihood ratio test applied in a historical control setting. Similar methods and conclusions can be generalized for other test statistics (e.g., a relative risk or risk difference) and settings (e.g., designs using concurrent controls). Further, we investigate a single adverse event occurrence, but we address the issue of multiple events in Section 5.

Assume that at each time point t, each subject's data consists of a binary outcome, Y_i, defined as 1 if subject i ($i = 1, \ldots, N_t$) has the adverse outcome of interest and 0 otherwise, and a vector of confounders, $\mathbf{Z}_i$, such as age, gender, and comorbidities. We further assume that the statistic $Y(t) = \sum_{i=1}^{N_t} Y_i$ has the following conditional distribution,

$$Y(t)|\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t} \sim Poisson(\beta\mu_t(\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t}))$$

where $Y(t)$ is the observed number of adverse outcomes up to time t, $\mu_t(\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t})$ is the expected number estimated from historical control data and the observed confounders of interest, $\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t}$, and β is the relative risk of Y for the vaccine or drug recipients compared to historical controls.

We will discuss how to estimate $\mu_t(\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t})$ from historical control information later. A Poisson distribution is used because adverse events monitored in post-licensure safety studies are usually rare, and we are making comparisons to expected rates among historical controls. In other settings, different distributions appropriate to the data could be considered, such as a binomial distribution when observed rates are compared to rates among concurrent controls [14]. The hypotheses of interest are $H_0 : \beta = 1$ versus $H_A : \beta \geq 1$. We use the log likelihood ratio statistic, which has been previously used by the MaxSPRT method in observational post-licensure surveillance [10]:

$$(1) \qquad LLR_t = \log\left(\max_{\beta \geq 1} \frac{e^{-\beta\mu_t}(\beta\mu_t)^{Y(t)}/Y(t)!}{e^{-\mu_t}(\mu_t)^{Y(t)}/Y(t)!} \right) = \log\left(\max_{\beta \geq 1} \frac{e^{-\beta\mu_t}(\beta\mu_t)^{Y(t)}}{e^{-\mu_t}(\mu_t)^{Y(t)}} \right).$$

β is estimated as $\max(1, Y(t)/\mu_t(\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_t}))$ using the maximum likelihood estimation. Hence,

$$(2) \quad LLR_t = \begin{cases} 0 & \text{if } Y(t) \leq \mu_t \\ \mu_t - Y(t) + Y(t)\log(\frac{Y(t)}{\mu_t}) & \text{if } Y(t) > \mu_t \end{cases}.$$

Note that the sequential probability ratio test (SPRT, [22]) was originally proposed for a simple hypothesis, $H_A : \beta = \beta_0$, and the MaxSPRT extends the SPRT with a one-sided composite hypothesis, $H_A : \beta \geq 1$.

2.2 Sequential testing boundary

In this manuscript, we use a unifying family of boundaries [5, 8, 16], which includes the Pocock [17] and O'Brien-Fleming [15] boundaries as special cases. We have chosen this flexible boundary framework to facilitate comparisons of a range of standard boundaries used in RCTs and to enable the construction of other boundaries specifically customized to the post-licensure surveillance setting.

The boundary is defined as $b(t) = au(t)$, where $u(t)$ is a function dependent on t and is from the unifying boundary family. On the Z-statistic (i.e., standard normal distribution) scale the Pocock boundary is constant, and the O'Brien-Fleming boundary is proportional to $\sqrt{N_T/N_t}$. Since $2LLR(t) \sim \chi^2_1$, its square root $\sqrt{2LLR(t)}$ is a standard Z-statistic. We use $u(t) = (N_T/N_t)^{1-2\Delta}$ on the scale of the log likelihood ratio statistic, where $\Delta \in [0, 1/2]$ is a parameter controlling the shape of the boundary, and the critical value a is chosen to hold the overall Type I error at the desired level α. For LLR(t), a Pocock boundary is given by $\Delta = 1/2$ ($u(t) = 1$), while $\Delta = 0$ ($u(t) = N_T/N_t$) gives the O'Brien-Fleming boundary. There are multiple methods available to solve for the critical value a. We used the following simulation framework [8] rather than using a normal approximation to solve for a [1], as this approximation may not work well in the rare event setting:

Step 1: Simulate data under H_0 and expect μ_t adverse events (i.e., $Y(t) \sim Poisson(\mu_t)$ for $(t = 1, \ldots, T)$).

Step 2: Compute LLR_t, $t = 1, \ldots, T$ on the simulated dataset.

Step 3: Compare LLR_t with $au(t)$, $t = 1, \ldots, T$. If for any $t \in [1, T]$, $LLR_t \geq au(t)$, set the rejection indicator $S_j = 1$; otherwise $S_j = 0$.

This process is repeated a large number of times (N_{sim}). The estimated Type I error α for this boundary is calculated across the simulated datasets as $\hat{\alpha} = \sum_{j=1}^{N_{sim}} S_j/N_{sim}$. Repeat the simulation changing the value of a until $\hat{\alpha} = \alpha$. Given the critical value a, one can compute the boundary $b(t)$, and determine its power and timeliness for different values of the relative risk, β, by repeating the simulation process with data simulated under H_A (i.e., $Y(t) \sim Poisson(\beta\mu_t)$).

This boundary formation has several advantageous features. First, by allowing the expected number of events, μ_t, to be a flexible function of time t, the boundary can properly incorporate the potential changes over time in the expected counts. Changes in expected counts may be needed over the course of the study because the observed study population may change dramatically with respect to its confounder distributions (e.g., subjects from a new study site where uptake of a new product is slower may enter during the course of the study with different demographics and thus different expected event rates). These changes in the population over time in turn affect the variance of $Y(t)$ and the Type I error α, and so they need to be appropriately accommodated by the boundaries. This represents a distinguishing and complicating feature of observational studies compared with RCTs as population characteristics are relatively well-controlled and usually stable over time in an RCT. Second, this approach does not assume that the between-test sample sizes n_t, $t = 1, 2, \ldots, T$, are equal. This is important since observational data may accrue in highly irregular quantities, or we may want to purposely manipulate the testing frequency by design to achieve specific study objectives (e.g., delay the first test to avoid false signals at early phases of the study when data are fewer and less reliable). In the next section we will discuss in more detail the design choices and considerations that are particularly relevant for post-licensure surveillance.

2.3 Sequential design choices for post-licensure safety surveillance

We have alluded to the existence of several additional challenges for sequential methods that are faced by observational surveillance studies compared to well-controlled RCTs, most notably confounding and unequal between-test sample sizes. Nelson et al. [14] provide a broader review and examination, including examples of these complications in practice for a sequentially-monitored vaccine safety study. We now address several of these issues in more depth.

First, confounding can be a major issue for observational post-licensure surveillance since characteristics of users and non-users of a new drug or vaccine can be quite different. Existing sequential methods that have been applied in the post-licensure surveillance setting handle confounding by using either matching [10] or stratification [12]. We propose to adjust for confounding through stratification in the estimation of μ_t. For example, assume that gender is a confounder, and we estimate the event rates in females and males to be λ_F and λ_M, respectively, using information from historical controls. At the t^{th} test, there are n_t^F newly observed females and n_t^M males. The estimate of μ_t controlling for gender would be $\hat{\mu}_t|(M,F) = \hat{\mu}_{t-1} + \lambda_F n_t^F + \lambda_M n_t^M$ for $t = 1, \ldots, T$ and $\hat{\mu}_0 = 0$. Through this approach, we estimate the expected number of the adverse outcomes using historical rates weighted by the observed numbers of subjects in the corresponding strata. Since the expected number is updated at each test, boundaries need to be adjusted accordingly. At each test, we first compute the expected number of events given the observed data and historical event rates of each stratum, and then we update the critical value a for current and future tests based on this information, through Monte Carlo simulation to maintain the overall Type I error. We assume the future critical values and covariate distributions are the same as those observed up to the current test. Note that we do not update critical value a for previous tests. As a result, although at a specific test t, $t = 1, 2, \ldots, T$, we assume a boundary $b(.)$ with a constant critical value a and estimate it by $\hat{a}_t$ through simulation, the estimated value $\hat{a}_t$, $t = 1, 2, \ldots, T$, can be different for different t to reflect the changing covariate distribution of new observations. Note this extends the method used in MaxSPRT [10], where the critical value a is constant and preset at the onset of the study.

Another complication for post-licensure surveillance is the occurrence of unequal between-test sample sizes over the course of the study, either by design or by the unpredictable nature of the observational data. For instance, investigators may want to delay the first test for a variety of reasons. Administratively, it may take several months after a new product is licensed to test the data collection systems and get surveillance started, which would delay the study start. Scientifically, we may want to delay the first test to avoid early signaling when sample size is relatively small and data are less stable. In addition, since multiple RCTs have already been conducted prior to licensure, to conserve power it may be desirable not to test the data until at least a comparable sample size as was observed in pre-licensure studies is obtained post-licensure. Delaying the first test can result in a much larger first between-test sample size than between subsequent tests.

Another common reason for unequal between-test sample sizes is that the rate of uptake for a new vaccine or drug may change over time. For instance, uptake might be slow at the beginning of the study and increase steadily after a couple months. If tests are planned based on calendar time (e.g., monthly), the unifying family of boundaries can account for potential between-test sample size differences. An alternative approach is to space tests evenly according to sample size. In other words, one would not test until a pre-specified number of subjects enter the study. This pre-specified number is computed at the beginning of the study to achieve the desired power under the chosen alternative hypothesis. Spacing tests evenly by sample size may result in less frequent testing (in calendar time) at the beginning of the study if initial uptake is slow. In practice, however, new data may only be available for analysis periodically (e.g., weekly rather than continuously), and so for a given analysis we may not observe the exact number of subjects between tests as planned. For example, assume that we have planned to conduct a test after every 500 subjects, and the cumulative observed weekly sample sizes for the first 6 weeks are as follows: 300, 600, 900, 1,300, 2,100, 2,600. In this case, we will perform an analysis the first week at or after the planned number of subjects are observed. Specifically, we will only test the data on week 2, 4, 5 and 6. Also notice that in this example that we skip one planned test at 1,500 subjects because of an unexpectedly large increase in sample size at week 5. Thus, since the actual numbers of subjects and tests may differ from the original plan, we propose to update the boundary at each test to account for these differences and maintain the overall Type I error. The unifying family of boundaries that we propose can be used to accomplish this.

3. SIMULATION STUDY

Drawing from the vast experience of group sequential methods used in RCTs, we have proposed a group sequential approach tailored for safety and to accommodate the special features of observational safety surveillance. In this section, we present a simulation study to evaluate our proposed sequential testing framework and to explicitly assess design choices oriented toward post-licensure surveillance studies. As we will describe in detail shortly, the general trends in performance that we observe here are not unexpected as they parallel similar tendencies seen in RCTs. The key difference in our work is that we evaluate a broader range of design options that are specifically customized to address safety in the post-licensure setting and, in particular, consider options that allow both for higher frequency testing and for more flexibility in testing frequency than usually found in a RCT setting. We also explain the importance of the patterns we observe in the context of post-licensure safety, which is not the typical orientation of a RCT.

3.1 Simulation set-up

Without loss of generality, we assumed the study period was two years (730 days) and that every subject was equally likely to have an event (i.e., no confounding). We investigated two boundary shapes (Pocock and O'Brien-Fleming boundaries) and two testing frequency scenarios (equally-spaced testing and a delayed first test followed by equally-spaced subsequent tests). The delayed first test scenario assumed that a large bolus of data was observed before the first test, and then equally-spaced tests with a smaller between-test sample size were performed thereafter ($n_t = c$ for $t = 2, \ldots, T$ and $n_1 \gg c$). We examined both a 1/4 year delay ($n_1/N_T = 1/8$) and a 1/2 year delay ($n_1/N_T = 1/4$). We varied the between-test sample size, n_t, by varying how often we tested the data from daily, weekly, monthly to quarterly. Note that MaxSPRT, which tests continuously (i.e., after each new subject enters the study), is approximated by our daily testing scenario. We chose to approximate MaxSPRT because in practical settings continuous testing is not generally feasible (e.g., the VSD actually tests on a weekly basis). In this simulation, for simplicity we assumed that subjects entered the study evenly throughout the study. Specifically, for equally-spaced testing (i.e., with no delay), daily testing implied $T = 730$ and $n_t/N_T = 1/730$, weekly implied $T = 104$ and $n_t/N_T = 1/104$, and so on. For the delayed first test design, the first test occurred either at day 91 (1/4 year delay) or at day 182 (1/2 year delay) with daily, weekly, monthly and quarterly thereafter (e.g., for a 1/2 year delay with daily subsequent testing, $n_1/N_T = 1/4$ and $n_t/N_T = 1/730$ for $t = 2, \ldots, 549$).

We varied the total expected number of adverse events μ_T from 5 to 50. Since there was no confounding and the between-test sample size was not changing within each simulation scenario, the boundaries did not have to be updated as described in Section 2.3. Note that in our historical control example, given μ_T, the proportion of observed subjects over time (N_t/N_T), rather than the between-test sample size (n_t), was the key parameter for determining boundary, power, and timeliness.

To calculate power, we first derived the boundary under the null ($\beta = 0$) with Type I error $\alpha = 0.05$ for each boundary shape, expected number of adverse events (μ_T), and testing frequency scenario. We then simulated data under different alternatives, varying the relative risk (β) from 1.5 to 2, and calculated power as the proportion of simulated datasets that crossed the boundary at least once. We also recorded for each simulation the time-to-surveillance-end, where the end could occur either because a signal was detected or because the two year end-of-study limit was reached (i.e., no signal). The time-to-surveillance-end was expressed on the scale of the expected number of adverse events under the null. We selected this measure as it is widely used in RCTs, and it characterizes timeliness while penalizing instances where a signal was failed to be detected. All results (i.e., power and time-to-surveillance-end) were based on 100,000 simulations.

It should be noted that for computational efficiency we slightly altered our simulation design as discussed in Section 2.2. Instead of simulating data at each test, as specified in step 1, we first simulated daily outcomes $Y_i \sim Poisson(\beta\mu_T/730)$ for $i = 1, \ldots, 730$ and then calculated $Y(t) = \sum_{i=1}^{N_t} Y_i$ for $t = 1, \ldots, T$. This allowed us to use the same datasets for all study designs evaluated, which reduced simulation error across study designs.

3.2 Simulation results

Power and time-to-surveillance-end are summarized in Table 1. As expected, power is higher when the adverse events are more common (i.e., μ_T increases) and for larger relative risks (i.e., β increases). For the Pocock boundary, less frequent testing yields higher power. For the O'Brien-Fleming boundary, this trend is less obvious. The O'Brien-Fleming boundary has about 5–15% higher power overall than the Pocock boundary, while the average time-to-surveillance-end is always longer for O'Brien-Fleming compared to Pocock boundary. Figure 1 shows the estimated percentage of rejections over time. The percentage increases steadily throughout the study when using a Pocock boundary. Using an O'Brien-Fleming boundary, the increase in percentage is S-shaped, with a rapid increase about 1/3 of the way through the study that catches up with the percentage for the Pocock boundary about 2/3 of the way through the study. The overall power advantage for the O'Brien-Fleming boundary is due to the fact that it saves power until the end of the study. The delayed first test designs have performance similar to those with no initial delay (Table 1). However, a longer delay yields considerably higher power when using a Pocock boundary. As the chosen boundary gets steeper, this trend is less pronounced. Delaying the first test does not necessarily delay the time-to-surveillance-end. In fact, for rare outcomes (small μ_t), delaying the first test yields a shorter average time-to-surveillance-end.

This simulation demonstrates that both the Pocock and O'Brien-Fleming boundaries have some desirable properties. On average, the Pocock boundary signals earlier, while the O'Brien-Fleming boundary has higher overall power. With the Pocock boundary, the choice of testing frequency affects the power and the timeliness more than with the O'Brien-Fleming boundary. Hence, if a Pocock boundary is chosen, then the testing frequency should be decided carefully. A delayed first test approach may also be useful to increase the overall power compared to a similarly designed study with no delay, particularly when using a Pocock boundary. Related findings are known from RCTs. Our study shows that with a much larger number of tests, which may be desirable in post-licensure observational safety surveillance, these patterns still hold. The magnitude of the differences in power, however, between less frequent (e.g., monthly) and very highly frequent (e.g., daily) testing was somewhat greater than anticipated, and this range and comparison of testing frequencies has not generally been studied in RCTs. For a safety surveillance study, where early detection of elevated risk is of high interest, a Pocock boundary may better serve the purpose. However, a delay of the first test may also be important to consider since in a rare event

Table 1. Power and average time-to-surveillance-end, varying the expected number of events, boundary shape, testing frequency, relative risk, and length of delay

RR (β)	Expected No.	Pocock $\Delta = 1/2$				O'Brien-Fleming $\Delta = 0$			
		Daily	Weekly	Monthly	Quarterly	Daily	Weekly	Monthly	Quarterly
					Power				
equally-spaced									
1.5	5	0.188	0.221	0.194	0.241	0.277	0.275	0.272	0.282
	10	0.293	0.322	0.327	0.366	0.420	0.419	0.437	0.450
	50	0.829	0.854	0.857	0.883	0.929	0.928	0.927	0.932
2	5	0.449	0.456	0.468	0.530	0.601	0.599	0.593	0.603
	10	0.704	0.734	0.738	0.779	0.832	0.831	0.842	0.851
	50	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
1/4 year delay									
1.5	5	0.237	0.238	0.233	0.239	0.276	0.275	0.270	0.279
	10	0.353	0.355	0.362	0.369	0.425	0.424	0.439	0.452
	50	0.874	0.877	0.887	0.884	0.930	0.929	0.927	0.934
2	5	0.530	0.531	0.511	0.528	0.598	0.596	0.590	0.599
	10	0.762	0.763	0.770	0.779	0.833	0.832	0.842	0.851
	50	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
1/2 year delay									
1.5	5	0.243	0.244	0.247	0.255	0.277	0.277	0.274	0.283
	10	0.370	0.367	0.376	0.380	0.424	0.423	0.437	0.449
	50	0.890	0.890	0.895	0.894	0.930	0.929	0.928	0.934
2	5	0.546	0.547	0.553	0.576	0.599	0.598	0.593	0.603
	10	0.785	0.783	0.787	0.787	0.833	0.832	0.840	0.850
	50	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
					Average Time to Surveillance End				
equally-spaced									
1.5	5	4.41	4.33	4.47	4.37	4.58	4.59	4.60	4.61
	10	8.30	8.18	8.22	8.31	8.68	8.71	8.76	8.91
	50	25.04	24.19	24.57	25.26	29.18	29.35	29.84	31.68
2	5	3.70	3.57	3.76	3.65	3.91	3.93	3.97	4.02
	10	5.88	5.70	5.79	5.98	6.58	6.64	6.74	7.04
	50	8.48	8.49	9.21	11.13	15.78	15.95	16.26	18.57
1/4 year delay									
1.5	5	4.36	4.37	4.43	4.37	3.46	3.46	3.53	3.61
	10	8.06	8.08	8.18	8.29	6.85	6.89	7.16	7.57
	50	23.60	23.76	24.81	25.16	27.54	27.59	28.22	30.32
2	5	3.57	3.58	3.70	3.65	3.92	3.93	3.99	4.04
	10	5.56	5.60	5.79	5.98	6.57	6.61	6.73	7.03
	50	9.61	9.80	11.29	11.14	15.78	15.84	16.27	18.57
1/2 year delay									
1.5	5	2.51	2.54	2.76	3.24	3.47	3.49	3.54	3.61
	10	5.06	5.13	5.38	5.63	6.82	6.85	7.16	7.55
	50	21.47	21.85	23.29	23.51	27.64	27.71	28.35	30.30
2	5	3.61	3.62	3.71	3.88	3.93	3.94	3.98	4.02
	10	5.76	5.82	5.97	6.19	6.56	6.59	6.76	7.04
	50	13.60	13.73	15.58	14.21	16.40	16.49	17.51	18.60

setting we do not expect information to accumulate very rapidly. In fact, we may not want to test until we achieve sample sizes that are comparable to those in pre-licensure studies to maximize the value added by the post-licensure evaluation and achieve power that was not possible prior to licensure. Delaying the first test can greatly increase overall power and can meaningfully shorten the time-to-surveillance-end.

4. APPLICATION: MMRV VACCINE SAFETY SURVEILLANCE

In 2005, a new combination MMRV vaccine comprised of measles, mumps, rubella and varicella components was licensed for use among children 1 to 12 years of age. This single-shot vaccine was intended to replace the need for two separate injections of the measles, mumps and rubella vac-

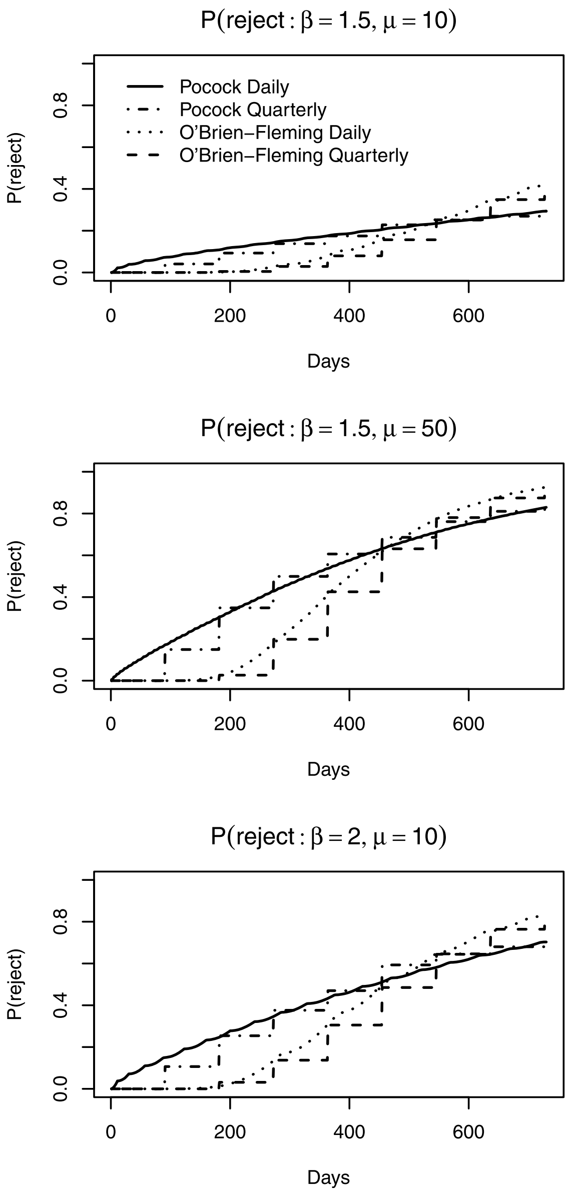

Figure 1. Percentage of rejection over time for Pocock and O'Brien-Fleming boundaries, varying the testing frequency (daily, quarterly), relative risk, and the expected number of events.

cine (MMR) and the varicella vaccine. In February 2006, the VSD began monitoring the safety of MMRV among children 1–2 years of age compared to historical recipients of MMR and varicella vaccine separately. Six adverse outcomes were separately monitored, including ataxia, meningitis and encephalitis, thrombocytopenia, febrile seizure, arthritis and allergic reactions. Data from seven sites were analyzed weekly using the Poisson-based MaxSPRT method (i.e., a continuous likelihood ratio test with a flat boundary). The total cumulative sample size N_T was 150,000 vaccines, which was chosen to achieve a specific upper limit on the expected number of adverse events [10]. We focus on the seizure data in this paper. Before the study started, each site estimated the site-specific seizure rates among a historical comparison population comprised of individuals who received MMR and varicella vaccines separately from that site during the five years prior to the introduction of the MMRV vaccine. An elevated seizure risk was detected at week 38 when the observed log likelihood ratio statistic exceeded the preset MaxSPRT rejection boundary (4.117). In this section, we present the results of a re-analysis of the seizure outcome using a group sequential framework.

In our re-analysis, we used the available total cumulative sample size $N_T = 150,000$, although in a new study, one would compute the sample size based on the desired power. We selected a Pocock boundary to improve the likelihood of detecting early signals. On average, 500 newly vaccinated subjects were observed each week of the study. Therefore we chose 500 as the between-test sample size to approximate weekly monitoring, 1,000 for bi-weekly, and 2,000 for monthly. We examined designs with equally-spaced weekly, bi-weekly and monthly tests, and also examined 1/4 year and 1/2 year delays for the first test, followed by subsequent daily, weekly, bi-weekly and monthly monitoring. All three between-test sample sizes yielded power close to 1 for a 1.5 relative risk. The results for our analyses are summarized in Table 2.

The original MaxSPRT continuous monitoring design detected a seizure signal on week 38. With (approximately) weekly monitoring, although the boundaries are reduced to around 3.5 from 4.12, the signal is still not detected until week 38. The log likelihood ratio test statistics at weeks 33 and 36 are very close to, but do not exceed, the corresponding critical values. With (approximately) bi-weekly monitoring, the boundaries further decreases to around 3.4, and a signal is detected earlier on week 33. With (approximately) monthly looks, the critical values are around 3.2, but the signal is not identified until week 39 because the monitoring frequency is reduced. When we delay the first test by a 1/2 year (sample size at first test was 5,099), the boundaries for weekly, bi-weekly and monthly all drop by about 0.1, and signals are detected at weeks 33, 36, 36, respectively. This example demonstrates that several group sequential designs yield lower boundaries, increase the power to detect elevated risks of rare adverse events, and potentially result in timelier signaling compared to a continuous sequential design. In this example, delaying the first test generally results in a shorter time-to-surveillance-end, but it does not substantially improve signal timeliness compared to designs that do not delay the first test. However, for situations where the outcome is rarer and the evidence for a safety problem appears relatively later, we expect that delaying the first test would greatly shorten the time to detection.

5. DISCUSSION

Continuous sequential testing methods have been proposed to facilitate rapid detection of safety signals in postlicensure studies [10, 19], but group sequential methods com-

Table 2. MMRV example data and group sequential testing boundaries for approximately weekly, bi-weekly, and monthly testing (between-test sample size 500, 1,000, 2,000, respectively), with a Pocock boundary, and equally-spaced testing or a delayed first test by half a year

Week	Cum. # vaccine	Cum. # event	Cum. expected # event	Observed RR(β)	Observed LLR	Boundary, continuous	Boundary, equal space			Boundary, delay 1/2 yr		
							500	1000	2000	500	1000	2000
1	48	0	0.060	0.000	0.000	4.117						
2	110	0	0.144	0.000	0.000	4.117						
3	194	0	0.257	0.000	0.000	4.117						
4	295	0	0.400	0.000	0.000	4.117						
5	402	0	0.550	0.000	0.000	4.117						
6	497	0	0.685	0.000	0.000	4.117						
7	627	0	0.865	0.000	0.000	4.117	3.576					
8	731	0	1.011	0.000	0.000	4.117						
9	865	0	1.191	0.000	0.000	4.117						
10	1011	1	1.387	0.721	0.000	4.117	3.538	3.318				
11	1177	3	1.604	1.870	0.482	4.117						
12	1353	5	1.833	2.728	1.851	4.117						
13	1514	5	2.043	2.447	1.518	4.117	3.471					
14	1686	6	2.265	2.649	2.110	4.117						
15	1845	6	2.472	2.427	1.792	4.117						
16	2021	7	2.702	2.591	2.366	4.117	3.481	3.359	3.191			
17	2147	7	2.863	2.445	2.121	4.117						
18	2306	7	3.077	2.275	1.831	4.117						
19	2505	7	3.340	2.096	1.520	4.117	3.480					
20	2710	7	3.616	1.936	1.240	4.117						
21	2920	7	3.908	1.791	0.989	4.117						
22	3127	9	4.198	2.144	2.061	4.117	3.486	3.367				
23	3462	9	4.668	1.928	1.577	4.117						
24	3819	9	5.164	1.743	1.164	4.117	3.466					
25	4236	10	5.749	1.739	1.284	4.117	3.470	3.397	3.230			
26	4648	10	6.329	1.580	0.903	4.117	3.474					
27	5099	10	6.966	1.436	0.581	4.117	3.512	3.382		3.293	3.196	3.052
28	5597	11	7.673	1.434	0.635	4.117	3.488					
29	6202	15	8.549	1.755	1.982	4.117	3.523	3.413	3.235	3.301	3.235	
30	6842	15	9.468	1.584	1.370	4.117	3.483			3.311		
31	7464	16	10.358	1.545	1.315	4.117	3.466	3.416		3.272	3.164	3.070
32	8211	18	11.435	1.574	1.602	4.117	3.471	3.447	3.200	3.293	3.190	
33	9016	23	12.604	1.825	3.438	4.117	3.452	**3.388**		**3.285**		
34	9896	24	13.920	1.724	2.994	4.117	3.524				3.196	3.055
35	10730	26	15.156	1.715	3.188	4.117	3.441		3.207		3.215	
36	11533	28	16.348	1.713	3.415	4.117	3.479				**3.208**	**3.104**
37	12345	29	17.575	1.650	3.099	4.117	3.457		3.222			
38	13168	34	18.795	1.809	4.949	**4.117**	**3.428**					
39	14063	35	20.120	1.740	4.497				**3.247**			

monly employed in RCTs have received less attention in this setting. Group sequential methods are important to consider since continuous testing may not always be feasible in practice. We have proposed a new approach for monitoring rare adverse events that uses group sequential testing methods and is tailored for post-licensure safety questions and to account for complications like confounding that arise in this non-randomized setting. The key advantage of this method is that it is more efficient and flexible than previously used continuous sequential testing methods. Although we have focused on a Poisson outcome with historical controls and likelihood ratio testing in this paper in order to facilitate comparability with previously applied methods in post-licensure safety evaluations, our method can easily be extended to other outcome types (e.g., binomial), other study designs (e.g., designs using concurrent comparisons), and other test statistics (e.g., a relative risk).

Our proposed approach uses boundaries that account for potential changes over time in the distribution of confounders in the study population and in planned between-test sample sizes, complications introduced by the lack of an experimental study and thus not previously of concern in RCT settings. Adjustment for confounding of the test statistic is handled through stratification, but development of methods that can more fully accommodate confounding (e.g., regression adjustment approach which can incorporate continuous confounders) is recommended. Additionally, in this paper we limited our focus to a single outcome per person. If there are multiple correlated outcomes, a multicomparison or joint modeling approach may need to be developed. For a relatively small number of multiple outcomes using a conservative Bonferroni correction or an extension of a multi-outcome Chi-square test developed for group sequential methods may be feasible [7]. However, for safety surveillance it may be appropriate to monitor individual outcomes separately to increase power especially when outcomes are not strongly correlated. When monitoring a large number of outcomes, then methods such as data mining that use joint modeling of all outcomes could be extended to this group sequential observational setting.

In this paper, we have used Monte Carlo simulation to determine boundary values, but a previous approach using asymptotic distribution properties of the likelihood ratio test statistic to form the boundary may be used alternatively [11]. However evaluation of this asymptotic boundary approach in the rare event setting, which is typical for postmarket surveillance, may need to be conducted.

6. ACKNOWLEDGMENTS

This study was funded through a subcontract with America's Health Insurance Plans (AHIP) under contract 200-2002-00732 from the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. We would also like to thank Nicola P. Klein, MD, PhD and Edwin Lewis of Kaiser Permanente Northern California Division of Research for kindly supplying the measles-mumps-rubella-varicella vaccine data from the Vaccine Safety Datalink project that were used to exemplify methods in this paper.

Received 17 September 2011

REFERENCES

[1] Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132** 235–244. MR0250405

[2] Baggs, J., Gee, J., Fowler, G., Weintraub, E., Lewis, E., Klein, N. P., Baxter, R., Benson, P., Jackson, L., Lieu, T., Naleway, A., Belongia, E., Glanz, J., Hambidge, S. J., Jacobsen, S. J., and Nordin, J. (2011). The vaccine safety datalink: A model for monitoring immunization safety. *Pediatrics* **127** S45–S53.

[3] Behrman, R. E., Benner, J. S., Brown, J. S., McClellan, M., Woodcock, J., and Platt, R. (2011). Developing the sentinel system — a national resource for evidence development. *New England Journal of Medicine* **364** 498–499.

[4] Belongia, E. A., Irving, S. A., Shui, I. M., Kulldorff, M., Lewis, E., Yin, R., Lieu, T. A., Weintraub, E., Yih, W. K., Li, R., Baggs, J., and Vaccine Safety Datalink Investigation Group (2010). Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. *Pediatr Infect Dis J* **29** 1–5.

[5] Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45** 905–923. MR1029609

[6] Greene, S. K., Kulldorff, M., Lewis, E. M., Li, R., Yin, R., Weintraub, E. S., Fireman, B. H., Lieu, T. A., Nordin, J. D., Glanz, J. M., Baxter, R., Jacobsen, S. J., Broder, K. R., and Lee, G. M. (2010). Near real-time surveillance for influenza vaccine safety: Proof-of-concept in the vaccine safety datalink project. *Am J Epidemiol* **171** 177–188.

[7] Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49** 741–752. MR1243490

[8] Kittelson, J. M. and Emerson, S. S. (1999). A unifying family of group sequential test designs. *Biometrics* **55** 874–882. MR1720716

[9] Klein, N. P., Fireman, B., Yih, W. K., Lewis, E., Kulldorff, M., Ray, P., Baxter, R., Hambidge, S., Nordin, J., Naleway, A., Belongia, E. A., Lieu, T., Baggs, J., Weintraub, E., and Vaccine Safety Datalink (2010). Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. *Pediatrics* **126** 1–8.

[10] Kulldorff, M., Davis, R. L., Kolczak, M., Lewis, E., Lieu, T., and Platt, R. (2011). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis* **30** 58–78. MR2770706

[11] Lai, T. L. and Shih, M.-C. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* **91** 507–528. MR2090619

[12] Li, L. (2009). A conditional sequential sampling procedure for drug safety surveillance. *Statistics in Medicine* **28** 3124–3138. MR2750410

[13] Lieu, T. A., Kulldorff, M., Davis, R. L., Lewis, E. M., Weintraub, E., Yih, K., Yin, R., Brown, J. S., Platt, R., and for the Vaccine Safety Datalink Rapid Cycle Analysis Team (2007). Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care* **45** 89–95.

[14] Nelson, J. C., Cook, A. J., Yu, O., Zhao, S., Jackson, L. A., and Psaty, B. M. (2011). Methods for observational postlicensure medical product safety surveillance. *Statistical Methods in Medical Research*.

[15] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

[16] Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stoppong in favor of the null hypothesis. *Journal of Statistical Planning and Inference* **42** 19–35. MR1309622

[17] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–200.

[18] Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38** 153–162.

[19] Shih, M.-C., Lai, T. L., Heyse, J. F., and Chen, J. (2010). Developing the sentinel system — a national resource for evidence sequential generalized likelihood ratio tests for vaccine safety evaluation. *Statistics in Medicine* **29** 2698–2708. MR2757017

[20] Topol, E. J. (2004). Failing the public health – rofecoxib, merck, and the FDA. *N Engl J Med* **351** 1707–1709.

[21] United States Code (2008). US Public Law 110-85 Food and Drug Administration Amendments Act of 2007 Accessed January 6, 2010.

[22] Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16** 117–186. MR0013275

[23] Yih, W. K., Nordin, J. D., Kulldorff, M., Lewis, E. M., Lieu, T. A., Shi, P., and Weintraub, E. (2009). An assessment of the safety of adolescent and adult tetanus-diphtheria-acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the vaccine safety datalink. *Vaccine* **27(32)** 4257–4262.

Shanshan Zhao
Department of Biostatistics
University of Washington
Seattle, WA 98195
USA
E-mail address: zhaoss@uw.edu

Andrea Cook
Biostatistics Unit
Group Health Research Institute
Seattle, WA 98101
USA
E-mail address: cook.aj@ghc.org

Lisa Jackson
Group Health Research Institute
Seattle, WA 98101
USA
E-mail address: jackson.l@ghc.org

Jennifer Nelson
Biostatistics Unit
Group Health Research Institute
Seattle, WA 98101
USA
E-mail address: nelson.jl@ghc.org