PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**(S1): 72–81
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.2320

ORIGINAL REPORT

Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the Mini-Sentinel pilot

Andrea J. Cook[1,2]*, Ram C. Tiwari[3], Robert D. Wellman[1], Susan R. Heckbert[4,5], Lingling Li[6], Patrick Heagerty[2], Tracey Marsh[5] and Jennifer C. Nelson[1,2]

[1] *Biostatistics Unit, Group Health Research Institute, Seattle, WA, USA*
[2] *Department of Biostatistics, University of Washington, Seattle, WA, USA*
[3] *Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA*
[4] *Departments of Epidemiology and Pharmacy, University of Washington, Seattle, WA, USA*
[5] *Group Health Research Institute, Seattle, WA, USA*
[6] *Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA*

ABSTRACT

Purpose This manuscript describes the current statistical methodology available for active postmarket surveillance of pre-specified safety outcomes using a prospective incident user concurrent control cohort design with existing electronic healthcare data.
Methods Motivation of the active postmarket surveillance setting is provided using the Food and Drug Administration's Mini-Sentinel Pilot as an example. Four sequential monitoring statistical methods are presented including the Lan–Demets error spending approach, a matched likelihood ratio test statistic approach with the binomial MaxSPRT as a special case, the conditional sequential sampling procedure with stratification, and a generalized estimating equation regression approach using permutation. Information on the assumptions, limitations, and advantages of each approach is provided, including how each method defines sequential monitoring boundaries, what test statistic is used, and how robust it is to settings of rare events or frequent testing.
Results A hypothetical example of how the approaches could be applied to data comparing a medical product of interest, drug A, to a concurrent control drug, drug B, is presented including providing the type of information one would have available for monitoring such drugs.
Summary We have described the current state of methodology for postmarket surveillance of pre-specified safety outcomes. We describe the limitations and advantages of the approaches while acknowledging areas for future development.

KEY WORDS—incident user cohort; observational study; postmarket; group sequential monitoring; signal refinement; surveillance

INTRODUCTION

There is a pressing public health need to monitor the safety of marketed medical products. Therapeutic and prevention products, such as vaccines, drugs, and devices, go through rigorous clinical trials evaluating efficacy and safety before being approved, but these trials are generally not of sufficient size to systematically detect rare adverse events and do not always include representation from all populations that receive them after their marketing. Therefore, the Centers for Disease Control and the Food and Drug Administration (FDA) have begun to utilize large multi-site healthcare databases to conduct postmarket surveillance evaluations for medical product safety. The FDA's Sentinel Initiative is an example of a program designed to improve the evaluation of safety across a large array of FDA-regulated medical products.[1]

This paper describes statistical methods for the evaluation of recently approved products using a prospective cohort observational design with existing electronic healthcare data for pre-specified safety

*Correspondence to: A. J. Cook, Biostatistics Unit, Group Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, USA. E-mail: cook.aj@ghc.org

outcomes. The goal of this study design is to quickly detect potential safety concerns by sequentially monitoring effect estimates multiple times throughout an evaluation. The aim is to determine whether, for a prespecified set of safety outcomes, there is an excess rate of observed events in recipients of the medical product of interest (MPI) compared with a single comparison group. The comparison group is important and can be chosen in several ways. In this manuscript, we consider a concurrent control group defined to be comparable to those taking the MPI after controlling for confounders. For example, when evaluating a new diabetes drug for safety, an appropriate comparison group could be those taking an alternative diabetes drug. However, we would need to control for site, and perhaps patient characteristics, because physicians from the various sites contributing data may exhibit differential prescribing habits, and patient characteristics may be associated with choice of diabetes drug.

This type of safety evaluation has been coined "signal refinement" because potential adverse events are predefined based upon the suggestion of a potential risk, which may result from various scenarios including, but not limited to, observation during preapproval or in adverse event reporting systems,[2] or because of known biologic reasons uncovered during the study of similar medical products. This signal refinement stage can be thought of as a preliminary step before conducting a more extensive phase IV observational study or confirmatory randomized clinical trial because existing healthcare databases, typically constructed for payment or clinical care purposes, tend to have issues such as incomplete data, data errors, and lack of information on potential confounders. There have been several examples of signal refinement studies published, but this is still a relatively new area.[3–7]

Statistical methods used to address hypotheses within postmarket safety evaluation designs must be able to detect both rare and common adverse events, control for confounding, and maintain the overall type I error across multiple tests. This manuscript describes the current state of statistical methods developed to conduct sequential analysis of prospective cohort data for medical product safety. We present four sequential methods that use different approaches to handle confounding, maintain the overall type I error, and have different statistical properties such as time to signal detection and power. Controlling for confounding is a major concern for observational safety surveillance and distinguishes it from the randomized clinical trial setting in which most sequential monitoring methods have been developed. Furthermore, when the outcomes of interest are rare, the inferential properties that hold in randomized trials, such as large sample asymptotics, may not hold in this setting and need to be assessed. We focus on methods already applied to observational safety surveillance evaluations and studies but, for comparison, also introduce one general method used in randomized clinical trials that is applicable to safety surveillance. We discuss potential limitations of these methods and conclude with discussion of the need for future work to develop methods tailored to the setting that we characterize.

METHODS

The electronic data generally captured for signal refinement by systems like Mini-Sentinel are primarily administrative and claims based, collected by health plans during the course of routine healthcare practice. Mini-Sentinel uses a distributed data system, in which individual level data, standardized using a common data model, remain at the local site. For this paper, we will assume that distributed programs summarize event and sample size counts at each site, stratified by exposure group and by confounders, and these results are then aggregated across sites for analysis. Although in some cases, analyses may be based on individual level data, more often to protect patient privacy, deidentified information is combined for central analysis, and thus, the focus of this discussion remains on aggregate data.

Data specifications and notation

We assume that accruing data will be analyzed at specific time points ($t = 1,\ldots,T$). We also assume that each individual i is either exposed to the MPI, $D_i = 1$, or not exposed, $D_i = 0$, and either has the outcome of interest occurring before the end of analysis t, $Y_i(t) = 1$, or does not $Y_i(t) = 0$. The exposure time, $E_i(t)$, denotes the cumulative exposure time prior to analysis t. It could be a single time exposure window (e.g., vaccine: $E_i(t) = 1$ for all individuals) or a chronic exposure (time on either MPI or comparator), for which assumptions of the exposure time and outcome relationship must be made (constant risk or change in risk because of exposure duration). For this manuscript, we censor a participant's exposure time at the date of disenrollment, occurrence of the outcome, or discontinuation of use of the initial prescribed treatment. In the case of discontinuation, we add a certain lag time to allow recognition of outcomes that could biologically be related to the exposure (e.g., 7 days after discontinuation

DOI: 10.1002/pds

of treatment for the outcome of seizure because outcomes more than 7 days after discontinuation are unlikely to be related to treatment). Furthermore, participants are censored if they switch exposure groups and begin taking the other medical product (i.e., an exposed individual starts taking the comparator medical product). A lag time also may be added after the date of switching exposures. These design features are consistent with the incident user cohort study design currently in common use in postmarket surveillance[8].

Furthermore, we assume that there is a set of baseline confounders, Z_i, associated with individual i, which can be composed of variables such as age, sex, site, and health conditions. When using aggregate data, these confounders often are categorized to form a set of categorical confounders, Z_i^c. For example, a continuous confounder, such as age, can be categorized into 5- or 10-year age groups. Under this data setup, confounding can be addressed by regression, stratification, or matching.

Sequential testing framework

In a signal refinement evaluation, the overall hypothesis of interest is whether there is a higher event rate for those on the MPI ($D_i = 1$) compared with the unexposed group ($D_i = 0$) after accounting for confounding and exposure time. Numerous test statistics (based on the relative risk (RR) or hazard ratio, for example) can be derived to evaluate this hypothesis, thus creating different statistical methods. The chosen hypothesis is tested at each analysis t, and if the test statistic at analysis t exceeds a pre-defined critical boundary, $c(t)$, it signals a significantly elevated rate of events at analysis t; otherwise, the study continues to the next analysis time until the pre-defined end of the evaluation. At each analysis, more new information accumulates, which may include new participants exposed and unexposed to the MPI since the last analysis, as well as more follow-up or exposure time for participants already included in the previous analysis. Different approaches to incorporating updated data induce different assumptions that need to be accounted for in the calculation of the critical boundary. The critical boundary can be chosen in numerous ways, but it must maintain the overall type I error rate across all analyses, taking into account both multiple testing and a skewed testing distribution that conditions on whether earlier test statistics exceeded the specified critical value at previous analysis times. A general review of sequential monitoring boundaries has been presented by Emerson *et al.*[9] and is beyond the scope of this paper, but we will present approaches specific to the observational surveillance setting and one general method used in randomized clinical trials that is applicable to this area.

Group sequential statistical methods

Lan–Demets group sequential approach using error spending. The first method we consider is a general group sequential method used mainly in randomized clinical trials developed by Lan and Demets[10] using an error spending approach. An error spending approach uses the concept of cumulative alpha or type I error, $\alpha(t)$, defined as the cumulative amount of type I error spent at analysis t and all previous analyses, 1,. . .,t-1. We assume that $0 < \alpha(1) \leq \cdots \leq \alpha(\mathrm{T}) = \alpha$, where α is the overall type I error to be spent across the evaluation period. The function $\alpha(t)$ can be any increasing monotonic function that preserves family-wise error, but there are several common approaches including the Pocock[11] boundary function $\alpha(\mathrm{t}) = \log(1 + (\exp(1)\text{-}1)\ \mathrm{t/T})\ \alpha$, O'Brien-Fleming[12] boundary function $\alpha(\mathrm{t}) = 2\left(1 - \Phi\left(Z_{1-\alpha/2} \Big/ \sqrt{t/T}\right)\right)$, and the general power boundary function $\alpha(\mathrm{t}) = (t/T)^p \alpha$ for $p > 0$. The most commonly used boundary function for safety evaluations has been a flat, Pocock-like, boundary on a standardized test statistic scale. This boundary spends α approximately evenly across analyses, given the test statistic is asymptotically normally distributed. Therefore, it spends more α at earlier analyses relative to later analyses, given the amount of statistical information, or sample size, observed up to time t compared with an O'Brien Fleming boundary, which is commonly used in efficacy studies. This flat boundary has been discussed as Pocock like, but a Pocock boundary when testing more frequently (quarterly or more often) is not completely flat. For further discussion of boundary shapes and statistical trade-offs between them in practice for postmarket surveillance, see Nelson *et al.*[13]

Given the error spending boundary function, Lan and Demets developed an asymptotic conditional sequential monitoring boundary for any asymptotically normal test statistic based on independent increments of data.[10] This boundary can be computed and used to compare with almost any standardized test statistic, including one that controls for confounding. For example, when interest is in an adjusted RR, $\hat{R}(t)$, or log RR, it can be estimated using Poisson regression, and a standardized test statistic can be calculated, $Zval(t) = \hat{R}(t) / \sqrt{Var(\hat{R}(\mathrm{t}))}$. The value of $Zval(t)$ can then be compared with the asymptotic conditional

DOI: 10.1002/pds

monitoring boundary developed by Lan and Demets,[10] resulting in a decision to stop if *Zval(t)* exceeds the monitoring boundary or to continue collecting additional data. This is an appealing approach because the boundary is very simple to calculate and relies on a well-defined asymptotic distribution. However, in practice with rare events and frequent testing (small amount of new information between analyses), the asymptotic properties of the boundary fail to hold. This is similar to the scenario where an exact test may be preferred to an asymptotically normal test when the sample size is small. The following methods have sought to address the shortcomings of this approach to allow for more precise statistical performance in a wider variety of settings.

Group sequential likelihood ratio test. The group sequential likelihood ratio test (LRT) approach is a method that has been used in the Vaccine Safety Data Link project to monitor vaccine safety for a single time vaccine exposure.[3,6,7,14] The approach uses exposure matching with a fixed matching ratio (1:M) to control for confounding and then computes a LRT statistic. The most commonly used method is the Binomial maxSPRT,[14] which assumes continuous monitoring (i.e., after each matched set of exposed and unexposed individuals come into the dataset, the test statistic is compared with the monitoring boundary).

Specifically, for the maxSPRT method, one creates matched exposure strata, s ($s = 1,\ldots,S$), such that each exposed individual, with $D_{s1} = 1$, is matched to one or more unexposed individuals ($D_{s2} = 0,\ldots,D_{S(M+1)} = 0$) who have the same categorical confounders, Z_i^c. Then, the log LRT statistic at each analysis, t, is the following:

$$LLR^1(t) = \log\left(\frac{\left(\frac{Y_{D=1}(t)}{Y(t)}\right)^{Y_{D=1}(t)}\left(\frac{Y_{D=0}(t)}{Y(t)}\right)^{Y_{D=0}(t)}}{\left(\frac{1}{M+1}\right)^{Y_{D=1}(t)}\left(\frac{M}{M+1}\right)^{Y_{D=0}(t)}}\right),$$

where $Y_{D=1}(t) = \sum_{s=1}^{S(t)}\sum_{j=1}^{M+1} Y_{sj}D_{sj}$ and $Y_{D=0}(t) = \sum_{s=1}^{S(t)}\sum_{j=1}^{M+1} Y_{sj}(1 - D_{sj})$ are the number of events observed among those exposed and unexposed to the MPI up to time t, respectively, and $Y(t) = Y_{D=1}(t) + Y_{D=0}(t)$ is the total number of events up to time t. Note that *S(t)* is the number of strata up to time t, which also is the number of exposed participants because we are assuming a fixed matching ratio of 1:M. This particular LRT, which conditions on the total number of events, $Y(t)$, is designed for the rare event case in which only one event is expected to be observed per exposure stratum. One can think of this LRT as comparing the observed proportion of exposed (and unexposed) events out of the total number of events to the expected proportion under the null, which is just 1/(M + 1) for the exposed participants and M/(M + 1) for the unexposed participants.

However, when events are not extremely rare, or when the probability within a stratum of more than one event occurring is not small, the assumptions of this LRT are violated, and a more general two-sample binomial likelihood ratio test statistic should be used:

$$\mathrm{LLR}^2(t) = \log\left(\frac{\left(\frac{Y_{D=1}(t)}{N_{D=1}(t)}\right)^{Y_{D=1}(t)}\left(1 - \frac{Y_{D=1}(t)}{N_{D=1}(t)}\right)^{N_{D=1}(t)-Y_{D=1}(t)}\left(\frac{Y_{D=0}(t)}{N_{D=0}(t)}\right)^{Y_{D=0}(t)}\left(1 - \frac{Y_{D=0}(t)}{N_{D=0}(t)}\right)^{N_{D=0}(t)-Y_{D=0}(t)}}{\left(\frac{Y(t)}{N(t)}\right)^{Y(t)}\left(1 - \frac{Y(t)}{N(t)}\right)^{N(t)-Y(t)}}\right),$$

where $N_{D=1}(t) = \sum_{s=1}^{S(t)}\sum_{j=1}^{M+1} D_{sj} = S(t)$ and $N_{D=0}(t) = \sum_{s=1}^{S(t)}\sum_{j=1}^{M+1}(1 - D_{sj}) = M \times S(t)$ are the number of people exposed and unexposed to the medical product up to time t, respectively, and $N(t) = N_{D=1}(t) + N_{D=0}(t)$ is the total sample size up to time t. Note that this general LRT incorporates the total sample size, unlike the binomial maxSPRT LRT that is conditional on the total number of events. For rare events, the performance of each LRT is similar. Further evaluation needs to be conducted to establish the scenarios in which each LRT has better statistical properties.

For the binomial maxSPRT, a Pocock-like boundary has been proposed, $c(t) = a$, which is a flat boundary on the log LRT statistic. One common way to solve for

DOI: 10.1002/pds

the constant, a, uses an iterative simulation approach similar to the following:

Step 1: Simulate data assuming H_o and the observed event rate while controlling for confounding (i.e., using a permutation approach: fix $Y_{s1},\ldots,Y_{sM}$ ($s=1,\ldots,S$), and permute $D_{s1},\ldots,D_{sM}$ to create $D_{s1*},\ldots,D_{sM*}$ so that you hold the exposure strata relationships and thus control for confounding).
Step 2: Calculate LLR(t) on the simulated dataset.
Step 3: If LLR(t) $\geq a$ then $Signal_k = 1$ and stop loop; otherwise, continue to next $t+1$.
Step 4: If $t = T$, then $Signal_k = 0$.

This process is repeated a large number, *Nsim*, times, and the estimated α level for the boundary is calculated as $\hat{\alpha} = \sum_{k=1}^{Nsim} Signal_k / Nsim$. One solves for a by repeating the simulation and changing a until $\hat{\alpha} = \alpha$.

This approach is a special case of the general unifying boundary approach developed by Kittleson *et al.*[15] To allow for the more general approach, define $c(t) = au(t)$ where $u(t)$ is a function dependent upon the proportion of statistical information (e.g., sample size) up to time t and is of the form $u(t) = (N(T)/N(t))^{1-2\Delta}$ where $\Delta > 0$ is a fixed parameter depending upon the design (e.g., $u(t) = 1$ is Pocock, and $u(t) = (N(T)/N(t))^{0.5}$ is O'Brien and Fleming). The same approach is used to solve iteratively for a, but the boundary $c(t)$ will now be shaped differently depending upon $u(t)$. We have named this more flexible version of the binomial maxSPRT as the group sequential LRT (GS LRT). This additional flexibility allows the method to be applied more generally, for example, within the Mini-Sentinel pilot, where data are not available as often (potentially quarterly). Furthermore, the shape of boundary can be changed to reflect the desired trade-offs appropriate to the specific safety question of interest. Because the original binomial maxSPRT used a unifying boundary type approach, we have presented it as such here, but as has been shown by others[16], the error spending approach and unifying approach are complementary, and therefore, we could have chosen an error spending approach.

A potential limitation of the GS LRT method is the fixed matching ratio. In practice, if there is a need to implement a strict matching criterion, because of the need for strong confounding control, then it can be difficult to find M unexposed matches for each exposed participant especially in the scenario of frequent monitoring. Frequent monitoring typically implies that an exposed participant should be matched to M unexposed participants within the current analysis time frame. This can lead to loss of matched strata including strata with events. When strata are lost, the results are then only generalizable to the subpopulation of the exposed population for which a matching control was found. Often, the matching criterion is then loosened, leading to less confounding control but a larger matched cohort.

Conditional sequential sampling procedure. The conditional sequential sampling procedure (CSSP)[17] was specifically developed to handle chronically used exposures, such as drugs that are taken over a period. However, the approach also is able to accommodate a single time exposure such as a vaccine. This method handles confounding using stratification and assumes that the data are aggregated.

Specifically, using categorical confounders, Z_i^c, one stratifies the entire population under evaluation (unlike GS LRT, which uses a matched sample). Then, at each analysis, t, within each confounder stratum, Z_k^S ($k=1,\ldots,K$), one calculates the exposure time, $E_{D,k}(t)$, and number of events, $Y_{D,k}(t)$ among all participants in stratum k on medical product D ($D=0$ (unexposed) or $D=1$ (exposed)) since the previous analysis t-1, where $E_{D,k}(t) = \sum_{i=1}^{N}(E_i(t) - E_i(t-1))I\left(Z_i^c = Z_k^S \text{ and } D_i = D\right)$ and $Y_{D,k}(t) = \sum_{i=1}^{N}(Y_i(t) - Y_i(t-1))I\left(Z_i^c = Z_k^S \text{ and } D_i = D\right)$. Under Ho, no relationship between exposure to the MPI and the outcome conditional on strata, the conditional distribution of $Y_{D=1,k}(t) | Y_{D=1,k}(t) + Y_{D=0,k}(t)$ is $Binomial\left(Y_{D=1,k}(t) + Y_{D=0,k}(t), \frac{E_{D=1,k}(t)}{E_{D=1,k}(t)+E_{D=0,k}(t)}\right)$, which is based on the proportion of exposure time observed for those exposed compared with the total exposure time including exposed and unexposed. Using this stratum-specific conditional distribution, one can simulate the distribution of $Y_{D=1,k}(t)$, the number of outcomes among those on the MPI within each stratum under Ho, given $Y_{D=1,k}(t) + Y_{D=0,k}(t)$.

The test statistic of interest is then the total number of adverse events observed among those exposed up to time t across all strata, $Y_{D=1}(t) = \sum_{k=1}^{K} Y_{D=1,k}(t)$. The CSSP approach uses an error spending approach in combination with the conditional stratum-specific distributions to create the sequential monitoring boundary. Specifically, it uses the following iterative simulation approach:

Step 1: Create a single realization of the following dataset of observed exposed counts under Ho for analysis t, $t=1,..,T$ as follows:

a) For all confounder strata k, simulate $\tilde{Y}_{D=1,k}(t) \sim Binomial\left(Y_k(t), \frac{E_{D=1,k}(t)}{E_{D=1,k}(t)+E_{D=0,k}(t)}\right)$ if $\tilde{Y}_k(t) > 0$ else set $\tilde{Y}_{D=1,k}(t) = 0$.
b) Calculate $\tilde{Y}_{D=1}(t) = \sum_{j=1}^{t}\sum_{k=1}^{K}\tilde{Y}_{D=1,k}(j)$ (total number of simulated exposed events at analysis t)

DOI: 10.1002/pds

Step 2: Repeat Step 1 for a large number of realizations, *Nsim*, to create a distribution of total number of exposed events at each analysis, $\tilde{Y}^{1}_{D=1}(t), \ldots, \tilde{Y}^{Nsim}_{D=1}(t)$.
Step 3: Order $\tilde{Y}^{1}_{D=1}(1), \ldots, \tilde{Y}^{Nsim}_{D=1}(1)$ from smallest to largest and if $Y_{D=1}(1) > \tilde{Y}^{(Nsim*(1-\alpha(1)))}_{D=1}(1)$ then signal at analysis t else continue.
Step 4: Set the simulated event counts that would have signaled at this analysis, $\tilde{Y}^{(Nsim(1-\alpha(t-1))+1)}_{D=1}(t-1), \ldots, \tilde{Y}^{(Nsim)}_{D=1}(t-1)$, to an extreme value, such as 1000, so that these realizations will be indicated as having past the boundary. This allows for a cumulative error spending calculation that incorporates stopping. Otherwise, keep $\tilde{Y}^{j}_{D=1}(t)$from Step 1 and repeat from 1 at next analysis, $t+1$.

Using this simulation approach explicitly incorporates the sequential monitoring stopping rules. Any form of the cumulative error spending function, $\alpha(t)$, can be assumed as discussed in the section on the Lan–Demets Group Sequential approach using error spending.

This CSSP approach is especially good when evaluating rare events, but it has limitations when there are too many strata and/or short intervals between analyses. The reason this approach breaks down is because the only informative strata are those that meet the following two criteria: (i) at least one observed event but not all participants observe an event; and (ii) both an exposed and unexposed participant. Furthermore, each analysis is treated as having separate strata because information from one analysis to the next is being treated as independent. Therefore, the true number of independent strata is$K \times T$(number of confounder strata times total number of analyses) across all analyses. So as both K and T increase, very few strata will be informative. As a result, the test statistic is less stable, which can both influence power and potentially inflate or deflate the type I error. Having a small number of informative strata also leads to results being generalizable to the informative strata population only and not to the overall population. Caution should be taken in the interpretation of the results in this high dimensional strata situation. Furthermore, this approach assumes a constant relationship between exposure duration and the probability of an event, which may not be valid. Overall, it has nice properties for the rare event case and will be applicable to post-market surveillance in settings where testing is not performed highly frequently or when too many confounder strata are required.

Group sequential estimating equation approach. The final approach we will present is an approach that controls for confounding through regression (unweighted or weighted). It can be applied to either the single exposure time or chronic exposure time settings. It has the flexibility to incorporate different exposure duration relationships, but we will focus on a constant relationship (i.e., given exposure duration, one assumes a constant rate of disease based just on exposure time). The approach uses a generalized estimating equation (GEE) framework and a score test statistic. Specifically, assume that the mean regression model under the null hypothesis, H_o, of no relationship between the MPI and the event is$g(E(Y_i(t))) = \beta_0 + \beta_Z Z_i + f_\theta(E_i(t))$, where g(.) is the mean link function; for example, the logit for a logistic model or the logarithm for a Poisson model. The exposure link function,$f_\theta(.)$, would typically be ignored for a single time exposure or specified as the logarithmic function if using a Poisson model. However, to allow for flexibility, this has been kept general.

Given the mean model, the generalized score statistic,[18] *Sc(t)*, can be calculated, with the additional specification for the family from which the data have arisen; for example, a binomial family for logistic regression and a Poisson family for a log regression model. However, a nice property of GEE when using the generalized score statistic is that it only assumes that the mean model is correctly specified.[19]

To calculate the sequential monitoring boundary, it has been proposed to use the following permutation data distribution:

Step 1: At each analysis t, simulate data by fixing $(Y_{N(t-1)+1}, Z_{N(t-1)+1}),\ldots,(Y_{N(t)}, Z_{N(t)})$ and permuting $D_{N(t-1)+1},\ldots,D_{N(t)}$ to create $D^{*}_{N(t-1)+1},\ldots,D^{*}_{N(t)}$ and calculate $\tilde{Sc}_j(t)$.
Step 2: Repeat Step 1 for a large number of realizations, *Nsim*, to create a distribution of score statistics, under H_o, at each analysis t, $\tilde{Sc}_1(t), \ldots, \tilde{Sc}_{Nsim}(t)$.

The boundary can be defined following the unifying boundary formulation as outlined for the GS LRT method or an error spending approach as outlined for GS LD method, except with this permuted dataset and score test statistic. Note that we are not directly estimating the effect of D_i because a score statistic is calculated under H_o. This allows for the test statistic to have better statistical properties, such as power, when the interest is in comparing alternative hypotheses that are closer to the null (e.g., better power relative to other methods for detecting RR = 1.5 versus RR = 3.0)[20].

The potential advantages of this approach compared with the other three approaches is that it may provide

DOI: 10.1002/pds

more flexible confounder control compared with GS LRT or CSSP, and it does not rely as heavily on the asymptotic assumptions as needed for the Lan–Demets error spending approach. However, a limitation to this approach, and any regression approach, is that it requires the first analysis to have enough events and observations to estimate the parameters of the mean regression model. This can be difficult for the extremely rare event case where the GS LRT or CSSP approaches may be preferable. As outlined by Nelson *et al.*,[13] it may be advantageous in safety surveillance to delay the first test of the data until an adequate amount of information has accrued, in which case, this method may be applicable in most commonly encountered situations. Furthermore, it requires more computational time than the well-defined asymptotically normal Lan–Demets error spending approach, so under the non-rare event case, the latter approach may be preferable for simplicity. Overall, all four approaches are applicable to the postmarket surveillance setting, and a brief summary of assumptions, limitations, and advantages is outlined in Table 1.

RESULTS

In this section, we present a hypothetical sequential monitoring application where the question of interest is as follows: Does the new drug A (the MPI) have a higher rate of myocardial infarction (MI) compared with drug B. The data are from five sites, and the confounders are age, sex, and body mass index (BMI). For deidentification, age is categorized into 5-year categories and BMI into four categories: low (BMI $18.5\,kg/m^2$), normal ($18.5\,kg/m^2 \le BMI < 25\,kg/m^2$), overweight ($25\,kg/m^2 \le BMI < 30\,kg/m^2$), and obese ($30\,kg/m^2 \le BMI$).

The surveillance evaluation is designed to sequentially monitor up to a total sample size of 10 000 participants assuming a flat, Pocock-style, boundary with the first analysis following accrual of 2500 participants and then analyses approximately every 417 participants (19 analyses) (Figure 1). This scenario is akin to a 2-year evaluation with constant accrual of 10 000 participants where the first analysis occurs after 180 days, and each subsequent analysis occurs monthly thereafter. For simplicity, the uptake of each drug is equal, and the expected percent with the event, MI, after 2 years is 5% overall. Table 2 shows an example of such a dataset.

We now apply three of the four methods discussed previously. We will not apply the *GS LRT* method because it is not applicable outside a single-time exposure setting. For the GS LD and GS EE methods, one uses the stratum-specific cumulative event data, $Y_{cum,s}=\sum_{j=1}^{t} Y_{j,S}$, and exposure time data, $Exp_{cum,s} = \sum_{j=1}^{t} Exp_{j,S}$, at each analysis (Table 2: Columns 8 and 10) and fits a Poisson regression model adjusting for age, sex, and BMI categories with $\log(Exp_{cum,s})$ as an offset term. The GS LD method then calculates the standardized Wald statistic based on the adjusted RR and compares this with the normal approximation boundary developed by Lan–Demets. The GS EE method calculates the generalized score test statistic and compares this with the permutation-derived critical boundary. The CSSP approach uses the total

Table 1. Overview of the four statistical methods sequential monitoring including potential advantages and limitations

	Exposure setting	Confounding control	Test statistic	Sequential boundary formulation	Potential advantages	Potential limitations
GS LD	Single time or chronic exposure	All: Matching, stratification, regression	Any standardized test statistic	Error spending boundary derived using a normal approximation	Easy to apply, flexible confounding control	In very rare event setting, or frequent testing, the normal approximation assumptions may not hold
GS LRT	Single time exposure	Matching with fixed matching ratio	LRT	Unifying boundary derived using permutation; potential to extend to error spending boundary	Matching provides an appealing interpretation	Information loss because of restricted sample; potential loss of exposed if matching criteria too strict or insufficient confounding control if criteria too loose
CSSP	Single time or chronic exposure	Stratification	Number of events for those on MPI	Error spending boundary derived by conditioning on number of events within strata	Works well for rare adverse events	May not maintain type I error when strata are small or if testing is frequent
GS EE	Single time or chronic exposure	Regression	Score statistic	Unifying boundary or error spending boundary derived using permutation	Flexible confounding control with few assumptions	Requires sufficient outcome data at first look to estimate the initial regression parameters

DOI: 10.1002/pds

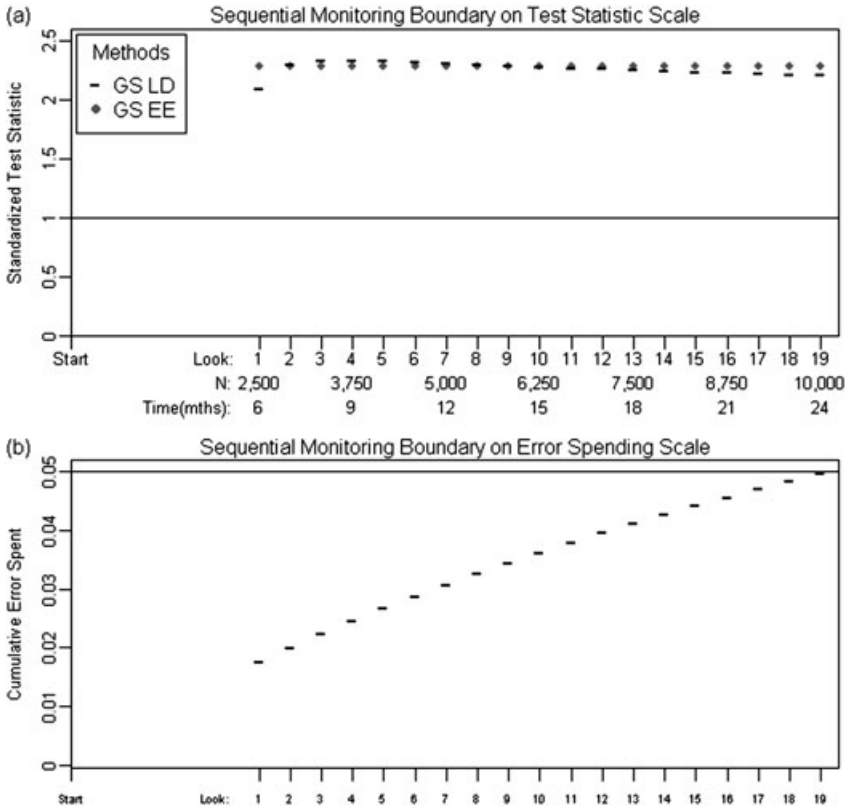

Figure 1. Sequential Monitoring boundaries for a flat, Pocock-style, boundary with a sample size of 10,000 participants with the first look after the first 2,500 participants and then approximately every 417 participants (19 looks) using a) *GS LD* and *GS EE* boundaries based on a standardized test statistic and b) *CSSP* boundary based on the error spending approach

Table 2. The structure of the aggregated data available for analysis in data systems like the Sentinel System

Look	Stratum	Site	Age	Sex	BMI	Drug	$Y_{cum,s}$[1]	$Y_{t,s}$[2]	$Exp_{cum,s}$[3]	$Exp_{t,s}$[4]
1	1	1	40–45	Male	Normal	A	2	2	250	250
1	1	1	40–45	Male	Normal	B	2	2	280	285
1	200	5	70–75	Female	Obese	A	11	11	720	720
1	200	5	70–75	Female	Obese	B	16	16	750	750
2	1	1	40–45	Male	Normal	A	2	0	320	70
2	1	1	40–45	Male	Normal	B	4	2	330	50
2	200	5	70–75	Female	Obese	A	13	2	780	60
2	200	5	70–75	Female	Obese	B	17	1	800	50

[1] $Y_{cum,s}$ is the total cumulative events observed at and before look t within each stratum s.
[2] $Y_{t,s}$ is the events observed only at look t within each stratum s.
[3] $Exp_{cum,s}$ is the total cumulative exposure time observed at and before look t within each stratum s.
[4] $Exp_{t,s}$ is the exposure time observed only at look t within each stratum s.

number of events for those on drug A, $Y_{Cum,D=A}(t) = \sum_{s=1}^{S(t)} Y_{cum,s}(t) I(D_i = A)$, as the test statistic, where $S(t)$ is the total number of strata at analysis t, and calculates an analysis-specific p-value (i.e., the probability of observing this test statistic, or one more extreme, based on the simulated distribution under the null) and compares this p-value with a Pocock error spending boundary. Figure 1 shows the different boundary shapes for the three methods.

Given these boundaries, Tables 3 and 4 provide an example of the type of monitoring summary one would create for a sequential monitoring evaluation. For this fictitious data example, the actual RR was 2, and all three methods signaled at the second analysis, but results often vary in other data settings. In this case, all methods performed equally well, and there was an indication of an elevated rate of MI for those on drug A compared with drug B even after controlling for confounding.

DOI: 10.1002/pds

Table 3. 3aExamples of monitoring data for the GS LD and GS EE methods when comparing observed test statistics with a standardized test statistic sequential boundary based on outcomes with prevalence 0.05 over the 2-year evaluation and confounding when the actual adjusted relative risk is 2

Look	Time (months)	Y_{cum}[1]	$Y_{cum,D=A}$[2]	Exp_{cum}[3] (person-days)	$Exp_{cum,D=A}$[4] (person-days)	RR_{AtoB}[5]	TestStat[6]	Test statistic boundary[7]	Signal
GS LD									
1	6	73	53	193 373	96 634	1.76	2.09	2.10	No
2	7	97	75	252 559	125 366	2.38	3.48	2.31	Yes
3	8	116	88	314 716	155 774	2.12			
4	9	143	107	379 954	187 629	2.09			
19	24	514	379	1 454 836	703 747	2.06			
GS EE									
1	6	73	53	193 373	96 634	1.76	2.16	2.28	No
2	7	97	75	252 559	125 366	2.38	3.65	2.28	Yes
3	8	116	88	314 716	155 774	2.12			
4	9	143	107	379 954	187 629	2.09			
19	24	514	379	1 454 836	703 747	2.06			

[1]Y_{cum} is the total cumulative events observed at and before look *t*.
[2]$Y_{cum,D=A}$ is the total cumulative events observed at or before look *t* for those on drug A.
[3]Exp_{cum} is the total cumulative exposure time observed at and before look *t*.
[4]$Exp_{cum,D=A}$ is the total cumulative exposure time observed at and before look *t* for those on drug A.
[5]RR_{AtoB} is the adjusted relative risk (RR) comparing drug A to drug B at each look adjusting for site, age, sex, and BMI category.
[6]TestStat is the observed test statistic calculated at each look and is the Wald-based test for GS LD and score-based test for GS EE.
[7]Test Statistic Boundary is the critical boundary in which the test statistic is compared to indicate if a given look has signaled.

Table 4. 3bExample of monitoring data for the conditional sequential sampling procedure method when comparing the estimated probability of observing number of observed outcomes in Drug group A with an error spending sequential monitoring boundary based on outcomes with prevalence of 0.05 over the 2-year evaluation and confounding

Look	Time (months)	Y_{cum}[1]	$Y_{cum,D=A}$[2]	Exp_{cum}[3] (person-days)	$Exp_{cum,D=A}$[4] (person-days)	Look *p*-value[5]	Error spending boundary[6]	Signal
CSSP								
1	6	73	53	193 373	96 634	0.020	0.017	No
2	7	97	75	252 559	125 366	0.012	0.020	Yes
3	8	116	88	314 716	155 774			
4	9	143	107	379 954	187 629			
19	24	514	379	1 454 836	703 747			

[1]Y_{cum} is the total cumulative events observed at and before look *t*.
[2]$Y_{cum,D=A}$ is the total cumulative events observed at or before look *t* for those on drug A.
[3]Exp_{cum} is the total cumulative exposure time observed at and before look *t*.
[4]$Exp_{cum,D=A}$ is the total cumulative exposure time observed at and before look *t* for those on drug A.
[5]Look p-value is the cumulative probability of observing $Y_{cum,D=A}$ or something more extreme at or before look *t*.
[6]Error spending boundary is the amount of cumulative alpha one specifies to spend at a given look. Given the error spending boundary, one computes the current *p*-value at each look, and if that current *p*-value is less than the error spending boundary, then the given look has signaled.

SUMMARY

We have presented four different group sequential monitoring approaches that are applicable to active postmarket surveillance of administrative and claims observational data. The theoretical underpinnings of each method have been described and illustrated using a hypothetical application. A formal evaluation of these four approaches still needs to be conducted to assess important statistical properties, such as delineation of scenarios in which a given method is appropriate (i.e., maintains the overall type I error and controls for confounding) or outperforms other methods. Performance often is quantified as having higher probability of signaling when a signal exists (power) or how quickly a method detects a signal (time to detection), which are clearly important quantities in safety surveillance.

There are still other methodological issues that need to be addressed. Open questions include developing better approaches to handle distributed data sources with more nuanced confounding control, extensions to the survival context for rare adverse events, and controlling for provider or facility effects. Therefore, the statistical methods presented represent a first step toward a general methodology appropriate for the signal refinement surveillance setting.

DOI: 10.1002/pds

CONFLICT OF INTEREST

The authors declare no conflict of interest.

KEY POINTS

- Active postmarket surveillance of pre-defined outcomes require sequential monitoring approaches that control the overall type I error, or false-positive rate, because of multiple testing over time.
- There are numerous sequential monitoring methods that can be applied, and these approaches differ based on the test statistic of interest, how the approach controls for confounding (stratification, matching, or regression), and how the approach derives the sequential monitoring boundary.
- There are numerous reasons that postmarket surveillance is different from the randomized control trial setting, in which most sequential monitoring methods have been developed, but key differences include the observational cohort design yielding a need for confounding control, more frequent testing because data are available more rapidly, and the interest often is in rare adverse events.
- The four approaches presented in this manuscript are the current statistical approaches being applied to the postmarket surveillance setting with appropriateness of a given approach depending upon strength of confounding control needed, frequency of testing desired, and how rare the adverse of interest is.

ACKNOWLEDGEMENT

Mini-Sentinel is funded by the Food and Drug Administration (FDA) through the Department of Health and Human Services (HHS) Contract Number HHSF223200910006I. The views expressed in this article do not necessarily represent those of the Food and Drug Administration.

REFERENCES

1. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System - A National Resource for Evidence Development. *N Engl J Med* 2011; doi: 10.1056/NEJMp1014427.
2. FDA Adverse Event Reporting System (AERS). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm (accessed 2011 01/27/2011).
3. Belongia EA, Irving SA, Shui IM, *et al*. Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. *Pediatr Infect Dis J* 2010; **29**(1): 1–5. doi: 10.1097/INF.0b013e3181af8605.
4. Brown JS, Kulldorff M, Chan KA, *et al.* Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf* 2007; **16**(12): 1275–1284. doi: 10.1002/pds.1509.
5. Brown JS, Kulldorff M, Petronis KR, *et al.* Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiol Drug Saf* 2009; **18**(3): 226–234. doi: 10.1002/pds.1706.
6. Klein NP, Fireman B, Yih WK, *et al.* Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. *Pediatrics* 2010; **126**(1): e1-e8. doi: 10.1542/peds.2010-0665.
7. Lieu TA, Kulldorff M, Davis RL, *et al.* Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care* 2007; **45**(10 Supl 2): S89–S95. doi: 10.1097/MLR.0b013e3180616c0a.
8. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf* 2010; **19**(8): 858–868. doi: 10.1002/pds.1926.
9. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. *Stat Med* 2007; **26**(28): 5047–5080. doi: 10.1002/sim.2901.
10. Lan KKG, Demets DL. Discrete Sequential Boundaries for Clinical-Trials. *Biometrika* 1983; **70**(3): 659–663.
11. Pocock SJ. Interim Analyses for randomized clinical-trials - The Group Sequential Approach. *Biometrics* 1982; **38**(1): 153–162.
12. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–556.
13. Nelson JC, Cook AJ, Yu O, *et al.* Challenges in the design and analysis of sequentially-monitored post-licensure safety surveillance studies using observational health care utilization data. *Pharmacoepidemiol Drug Saf* 2012; **21**(S1): 62–71.
14. Kulldorff M, Davis RL, Kolczakâr M, Lewis E, Lieu T, Platt R. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. *Sequential Analysis: Design Methods and Applications* 2011; **30**(1): 58–78. doi: 10.1080/07474946.2011.539924
15. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**(3): 874–882.
16. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**(4): 770–777.
17. Li LL. A conditional sequential sampling procedure for drug safety surveillance. *Stat Med* 2009; **28**(25): 3124–3138. doi: 10.1002/sim.3689.
18. Rotnitzky A, Jewell NP. Hypothesis-testing of regression parameters in semiparametric generalized linear-models for cluster correlated data. *Biometrika* 1990; **77**(3): 485–497.
19. Zeger SL, Liang KY, Albert PS. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* 1988; **44**(4): 1049–1060.
20. Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. (3rd edn), Casella G, Stephen F, Olkin I, (eds). Springer Science + Business Media, LLC: New York, NY, 2005; 545.