Evaluation of signal detection methods for use in prospective post-licensure medical product safety surveillance

FDA Sentinel Initiative Safety Signal Identification contract (contract #HHSF223200S31324P)

Jennifer Nelson
Andrea Cook
Onchee Yu

With significant contributions on Recommendations (Section 7) from:
Lisa Jackson
Bruce Psaty

March 31, 2009

Summary

The Sentinel Initiative aims to develop a national, electronic network to link data on 100 million patients from multiple existing health care data systems by 2012 in order to conduct post-licensure safety monitoring of FDA-regulated medical products. A critical component of this Initiative is determining what statistical methods can best be employed within this framework to accurately, robustly, and flexibly detect safety problems. In this report, we review existing signal detection methods for possible use in the Sentinel Initiative. We statistically and clinically describe each method, summarize prior and current uses of each method, and assess what is known about each method's robustness across different data sources and each method's flexibility to accommodate different signaling threshold levels and different types of adverse event outcomes. Recommendations for pursuing the most promising signal detection methods for use in the Sentinel Initiative are given.

FDA-2009-N-0192

RPT

1 Introduction

1.1 Context

Methods for signal detection, the ability to discriminate between signal and noise, have been used in many fields, including use in safety surveillance settings to identify elevated risks of adverse events associated with new medical products. Automated and rapid Bayesian data mining techniques[1-3] have been used to search for an unexpectedly high number of adverse drug reactions, and interactive graphical tools[3-4] have been used to visualize adverse event patterns in the FDA's spontaneous Adverse Events Reporting System (AERS). However, these methods have not been applied in a longitudinal fashion for prospective surveillance. Continuous sequential monitoring methods, such as Wald's classical sequential probability ratio test[5] (SPRT) and Kulldorff's maximized sequential probability ratio test (maxSPRT)[6] have been proposed for the rapid and prospective detection of adverse events after the introduction of a new medical product. However, such highly frequent monitoring may not be feasible or desirable in all instances. A natural methodology for prospective testing on a periodic basis is group sequential interim monitoring, and these methods have been widely used to monitor medical product efficacy and safety in clinical trials.[7-8] However, their use in active observational safety surveillance settings, in which more frequent monitoring may be desired and confounding is a potential concern, has not been considered.

Thus, although several signal detection methods have been implemented in a variety of post-licensure safety surveillance settings, a detailed statistical and clinical evaluation of these methods for use in active, longitudinal safety surveillance using linked electronic database information has not been reported. A thorough analysis of existing signal detection methods is critical to the FDA in determining what methods are most suitable for application in the Sentinel Initiative.

1.2 Scope

At the time of medical product approval, a key first step in planning post-licensure safety surveillance is to determine what type of surveillance system is most appropriate for a given medical product and adverse event pair. In the case of the Sentinel Initiative, the more specific question of interest is whether or not active surveillance of a large observational cohort is the preferred design compared to other potential choices such as a Phase IV epidemiological cohort study or a large Phase IV randomized clinical trial. If active surveillance is determined to be the preferred system, a second key step is to assess whether the relevant adverse event outcomes, medical product exposure groups, and confounders are accurately and reliably measured using the linked electronic health care utilization database sources to be used for surveillance.

Although determining when active surveillance is the optimal post-licensure system and ensuring that database information is of adequate quality to support surveillance are important, these topics are not the focus of this report. However, since the success of any

signal detection method used for active surveillance depends critically on these two factors, we conducted pilot research on these topics (including consultation with outside experts in statistics, epidemiology, and medical product safety), and present these results in Appendices A-B to supplement this report. Our recommendations (Section 7) also provide some guidance on these topics.

1.3. Purpose

The primary goal of this report is to evaluate existing signal detection methods for potential use in the Sentinel Initiative to conduct active post-licensure safety surveillance for FDA-regulated medical products. We assume that active surveillance has been determined to be the preferred post-licensure surveillance system and that the available database information is of adequate quality to support surveillance activities. Our evaluation included a systematic review of the signal detection literature to identify a wide range of possible relevant method areas, a Screening Evaluation to assess method assumptions and potential applicability to Sentinel Initiative goals, and a detailed Full Evaluation of methods with clear promise for use in the Sentinel Initiative. Our Full Evaluation involved researching published literature, facilitating teleconferences between the FDA and outside experts in vaccine and drug safety, pharmacoepidemiology, and biostatistics, and directly interviewing additional outside expert scientists in these fields.

Specifically, in this report we statistically and clinically describe each method, including each method's field of origin, statistical properties, appropriateness of assumptions, accuracy, strengths and weaknesses, and limitations in a post-licensure surveillance setting. We summarize prior and current uses of each method, including representative publications and presentations that use each method as well as first-hand input from researchers and method developers on their experience with existing methods and plans for future methodological developments. We assess what is known about each method's robustness across different data sources, including administrative and electronic medical record data, and we assess what is known about each method's flexibility to accommodate different signaling threshold levels and different types of adverse outcomes. Recommendations for pursuing the most promising existing signal detection methods for use in the Sentinel Initiative are given.

2 Continuous sequential monitoring methods

In classical hypothesis testing, a single test is performed at the end of a study to compare the evidence in support of a null (H_0) versus an alternative (H_A) research hypothesis. A hypothesis test can have two types of errors: Type 1 or false positive error (often denoted by α) which involves falsely rejecting the null when the null is true, and Type 2 or false negative error (often denoted by β) which involves failing to reject the null when the alternative is true. An individual hypothesis test is conventionally designed to ensure that the probability of a false positive error is small, typically with α=0.05.

Sequential hypothesis testing involves conducting repeated hypothesis tests at multiple points in time throughout a study as data accrue. However, if multiple hypothesis tests are

conducted, each with a false positive error rate of 0.05, then the overall false positive rate across all tests may be considerably larger than 0.05 because the probability of rejecting the null by chance increases as the number of tests performed increases. Sequential testing methods are specifically designed to control the overall false positive statistical error rate across all tests performed. Continuous sequential testing methods involve controlling the overall false positive rate when a hypothesis test is performed continuously in time as each new subject enters the study.

2.1 Sequential probability ratio test

2.1.1 Overview

The sequential probability ratio test (SPRT) is a type of sequential hypothesis test originally developed for manufacturing quality control.[5] SPRT involves comparing the amount of evidence in support of a null hypothesis (H_0) versus a simple alternative hypothesis (H_A) continuously in time after each new observation enters the study. For example, in a medical product safety setting, one might be interested in testing whether the probability of an adverse event outcome is higher by a pre-specified amount using the relative risk (RR) of an adverse event for exposed compared to unexposed individuals and following test: H_0: $RR=R_0$ (where R_0 is usually 1) versus H_A: $RR=R_A$, where R_A is a single value. Using SPRT, one tests the desired hypothesis by computing the cumulative sum of the log-likelihood ratio test (LRT) after each new subject enters the study and comparing the observed LRT statistic with pre-specified upper and lower thresholds. The larger the LRT, the more the observed data support H_A versus H_0. At each test, three decisions are possible:

1) if the *log*(LRT) is greater than a preset upper threshold (B), H_A is accepted
2) If the *log*(LRT) is less than a preset lower threshold (A), one accepts H_0
3) If the *log*(LRT) is in between A and B, monitoring continues

The thresholds A and B are selected to ensure that the overall statistical errors of the test (i.e. the Type 1 false positive error rate, α, and the Type 2 false negative error rate, β, are at pre-specified desired levels (e.g. α=0.05). A visual description of the SPRT decision rules is depicted in Figure 1. Hypothetical data are used to illustrate an example trajectory of the *log*(LRT), denoted by Λ. In this example, the *log*(LRT) exceeds the upper threshold at the 19th observation, so we accept H_A and would stop monitoring after this observation.

2.1.2 Statistical description

In statistical terms, SPRT tests a null hypothesis $H_0 : f = f_0$ versus a simple alternative hypothesis $H_A : f = f_1$ based on independent observations $X_1, X_2, \ldots$ having a common density function f. As the ith observation arrives, the log-likelihood ratio statistic Λ_i is calculated:

$$\Lambda_i = \log\left(\frac{L(H_A)}{L(H_0)}\right) = \log\left(\frac{\prod_{j=1}^{i} f_1(X_i)}{\prod_{j=1}^{i} f_0(X_i)}\right)$$

The stopping rule is as follows:

1) $\Lambda_i \geq B$: stop monitoring, and accept H_A
2) $A < \Lambda_i < B$: continue monitoring
3) $\Lambda_i \leq A$: stop monitoring, and accept H_0

where $0 < A < B < \infty$ (see Figure 1).

A and B are chosen to maintain the pre-specified Type 1 and Type 2 errors. Exact values of A and B are difficult to calculate analytically, but Wald showed an approximation to the thresholds: $A \approx \log(\frac{\beta}{1-\alpha})$, $B \approx \log(\frac{1-\beta}{\alpha})$. This approximation is based on the assumption of negligible excess of the likelihood ratio over the threshold when the test ends, which is often true in practice.

As an example, the probability of an adverse event in the general population may be equal to p_0, but vaccination may increase this risk. Let $X_1, X_2, \ldots$ be dichotomous adverse event observations from a Bernoulli distribution with probability, p, of an event. Let $p = p_0$ among unvaccinated subjects, and suppose we want to test for a 10-fold increase in this probability among vaccinated subjects. The hypotheses are thus $H_0 : p = p_0$ versus $H_A : p = p_1 = 10p_0$ or in terms of the RR we have $H_0 : RR = 1$ versus $H_A : RR = 10$ with known p_0. If we desire Type 1 error α=0.05 and statistical power= 1-β=0.8, then based on Wald's approximation the thresholds are:

$$A \approx \log(\frac{\beta}{1-\alpha}) = -1.56, \; B \approx \log(\frac{1-\beta}{\alpha}) = 2.77$$

and the log-likelihood ratio Λ_i is

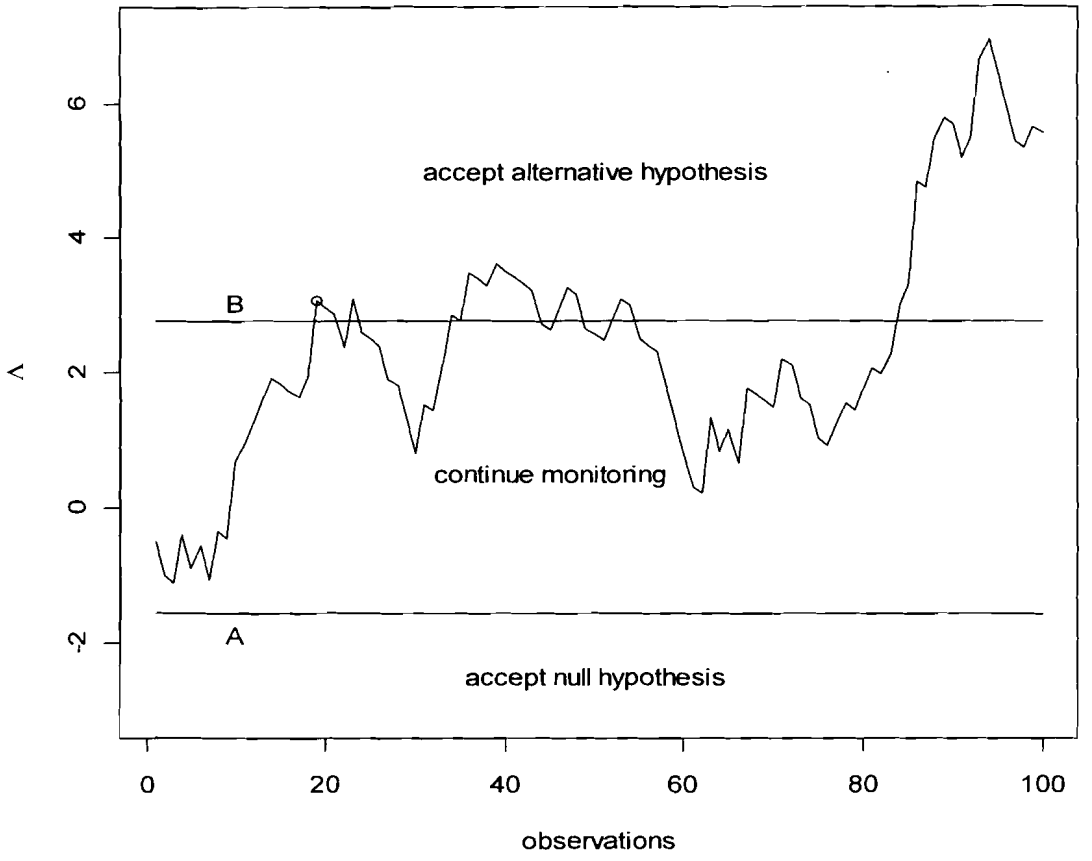
$$\Lambda_i = \log\left(\frac{\prod_{j=1}^{i} p_1^{X_i}(1-p_1)^{1-X_i}}{\prod_{j=1}^{i} p_0^{X_i}(1-p_0)^{1-X_i}}\right) = \log\left[\left(\frac{p_1}{p_0}\right)^{\sum_{j=1}^{i} X_j}\left(\frac{1-p_1}{1-p_0}\right)^{i-\sum_{j=1}^{i} X_j}\right]$$

$$= \left(\sum_{j=1}^{i} X_j\right)\log\left(\frac{p_1}{p_0}\right) + \left(i - \sum_{j=1}^{i} X_j\right)\log\left(\frac{1-p_1}{1-p_0}\right)$$

$$= \left(\sum_{j=1}^{i} X_j\right)\log 10 + \left(i - \sum_{j=1}^{i} X_j\right)\log\left(\frac{1-10p_0}{1-p_0}\right)$$

After each new subject X_i enters the study, we recompute Λ_i and compare it to A and B. We accept H_0 if $\Lambda_i \leq -1.56$, accept H_A if $\Lambda_i \geq 2.77$, and otherwise continue the study and evaluate the $(i+1)^{th}$ observation.

Figure 1: SPRT thresholds (A and B) and observed likelihood ratio statistic values (Λ) from a hypothetical example.



2.1.3 Applications

SPRT was the first method tested within the Vaccine Safety Datalink (VSD) as an approach for conducting longitudinal sequential vaccine safety monitoring.[9-10] Other applications in medicine include clinical trials[11-12] and in observational monitoring of patient outcomes post-surgery.[13-14] Adjustment for potentially confounding differences between exposed and unexposed groups, such as gender, age, or baseline risk status, has been proposed for SPRT by making stratified comparisons of the cumulative log likelihood ratios.[13]

2.1.4 Advantages

The main advantage of SPRT is that it is computationally simply and easy to implement. In addition, SPRT has been shown in some settings to detect signals earlier than chart-based detection methods, such as cumulative sum charts (CUSUM) and Shewhart p-charts.[15] The SPRT threshold plot is also a useful way to visualize the monitoring process.

2.1.5 Limitations

SPRT has several disadvantages:

- Lack of accuracy: Because SPRT tests a simple alternative hypothesis, the resulting test statistic is highly sensitive to the choice of the alternative hypothesis. SPRT may delay or fail to signal if the alternative is incorrectly specified (i.e. if the specified alternative is not near the truth). For example, for a hypothesis test based on the *RR*, if the true $RR=2$ and one specifies $RR=10$ as the alternative, then the truth ($RR=2$) is closer to the null ($RR=1$) than it is to the alternative ($RR=10$) and SPRT may not yield a signal. Conversely, if the true $RR=10$ and one specifies $RR=2$ as the alternative, then because the null ($RR=1$) and alternative ($RR=2$) are relatively similar, it takes an accumulation of more data for the LRT to distinguish between the two hypotheses and thus a longer time for SPRT to generate signal. Therefore, to ensure that SPRT produces an accurate and timely signal, one must know beforehand what the true excess risk is, which is unlikely in practice.
- Unpredictable length of surveillance: SPRT monitoring may continue indefinitely if the *log*(LRT) stays between A and B. There is no guarantee that the *log*(LRT) will either exceed B or be lower than A and thus reach a conclusion in favor of H_A or H_O.
- Designed for continuous testing:
  - Although continuous or nearly continuous weekly monitoring may be advantageous for rapid detection, such highly frequent assessments may not be feasible or desirable in all settings. For example, a complex and large linked database system may not be equipped for real-time analyses, but it might be possible to assemble and analyze such data on a less frequent basis. Further, continuous testing is not possible when outcomes require confirmation by review of medical charts, since chart-validated outcomes generally cannot be obtained in real-time. Such confirmation is not uncommon to ensure the accuracy of outcomes presumptively-identified in health care utilization databases. Interim analyses could be performed periodically, however, as batches of chart-reviewed outcomes become available. Even unconfirmed presumptively-identified adverse event data do not always accumulate in a strictly continuous fashion but rather may actually be updated on a weekly or less frequent basis.
  - When data are analyzed continuously in close-to-real-time, there is an increased probability of data quality problems that arise relating to changes in the dynamic health care utilization databases. Changes can be due to delays in the arrival of data on adverse events (e.g. due to delay in receiving hospital claims data), delays in the arrival of exposure status data, disappearing people (e.g. due to the fact that

health care organization enrollment files are only updated on a periodic basis). In the VSD's safety surveillance studies, a lag of 8 weeks is necessary to allow database information to stabilize.
  - Because SPRT thresholds are designed for continuous testing, if SPRT is applied on a less frequent basis (which may be likely in practice) it does not have optimal statistical power compared to alternative thresholds (e.g. thresholds designed for less frequent interim monitoring) that maintain the same Type 1 error.
  - Although sequential testing can largely be automated, continuous testing is inherently more expensive than testing on a less frequent ongoing basis. There are also significant costs associated with developing the infrastructure to acquire data in real-time to facilitate continuous analyses.
- Constant (flat) threshold over time: Although a flat boundary is more powerful at early time points, it also yields more Type 1 errors at early time points compared to many commonly used boundaries in clinical trials which are wider at earlier time points and narrower at later time points. Type 1 errors can be very costly in a safety surveillance setting. For example, in the VSD, all signals (true or false) based on outcomes defined using information from utilization databases are followed up with extensive confirmatory analyses (e.g. outcome validation by chart review, alternate statistical analyses using different comparison groups, control outcomes, control exposures, and scans for temporal trends in outcome occurrence). Since numerous outcomes and vaccines are being monitored within the VSD, the resources needed to confirm or refute false positive signals can therefore be quite large. A narrowing threshold over time, such as those used in some clinical trials, may be advantageous because it can prevent Type 1 errors from being made at early time points when sample sizes are small and there is greater uncertainty and variability in the data. It also retains more power at later time points when more data have accumulated.
- Lack of interpretability: Stopping thresholds for SPRT are computed on the scale of the LRT, which can be difficult to interpret. Using a more familiar statistic like the *RR*, that is on a scale which investigators naturally quantify risk differences, could improve the transparency of the threshold-setting process and facilitate a better understanding of how the thresholds will perform in practice. For instance, threshold selection could be based directly on the size of the *RR* that represents a concerning risk difference for which an investigator would want to signal a safety problem.
- Limited confounder adjustment: Accommodation of confounders is possible through stratification only, which may provide limited adjustment compared to a more continuous adjustment method if fine enough stratification cannot be achieved.

2.2 Chart-based detection methods

2.2.1 Overview

Chart-based methods, such as the cumulative sum chart (CUSUM), Shewhart charts, and exponentially-weighted moving-average control chart, derive from SPRT. In this review, we focus on the cumulative sum chart (CUSUM) method,[16] which was initially developed for industrial production monitoring problems, as a representative example of a chart-based approach. Like SPRT, CUSUM methods continuously monitor the data

sequentially over time until evidence in favor of a specific alternative hypothesis is obtained. In process control, this involves monitoring a subgroup of units taken from the production process at roughly the same time until process deterioration is detected. Thus, the primary purpose of the CUSUM methodology is to sequentially detect a *change* over time in a process from a baseline state of assumed 'control' to a specific state of process deterioration. An example of this type of problem in public health is the detection of a disease outbreak in a particular population of interest. A comprehensive review of the advantages and limitations of methods designed to sequentially detect change in process or in public health surveillance data over time like CUSUM is provided by Sonesson and Bock.[17] In this report, we focus on potential applicability of such methods in the Sentinel Initiative where the primary purpose is to detect an adverse event risk difference between populations exposed versus unexposed to a new medical product.

One implements the CUSUM method by performing sequential hypothesis tests continuously in time as each new unit enters the production process. The hypothesis test compares the amount of evidence in support of a null hypothesis (H_0) that represents a state of 'control' versus a simple alternative hypothesis (H_A) that represents a specific state of process deterioration. Each test is performed by computing a score based on accumulated information from all previous observations. The score will remain close to zero for a long time if no change has occurred in the process but will become large and move away from zero if a process has deteriorated. Unlike SPRT which has a preset lower threshold (A), CUSUM has a 'holding barrier' at zero. This means that if the computed score value falls below zero, then the score is set and held to be zero until more observations accumulate and the statistic becomes positive again. Thus, in CUSUM monitoring one can never fall beneath a lower threshold and 'accept' the null hypothesis. It is therefore intended for situations in which one desires on going monitoring until an alarm is raised. Further, CUSUM procedures are designed to signal if sufficient evidence has accumulated that an industrial process has changed substantially are thus ideal for detecting small persistent process changes.[18] CUSUM thus allows rapid detection of deterioration (or improvement) in a process while not overreacting to the expected smaller fluctuations due to chance.

2.2.2 Statistical description

The CUSUM procedure sequentially tests the hypothesis H_0: $\theta = \theta_0$ versus H_A: $\theta = \theta_A$, where θ is a parameter that characterizes a particular feature of the process that is of interest for detecting potential process changes. The value of θ_0 is typically determined by current process performance, and θ_A corresponds to a specific level of inferior performance. The CUSUM procedure involves computing and plotting a cumulative score (S_i) versus the observation number ($i=1, 2, 3...$) and comparing S_i to a preset upper stopping threshold designed to ensure a pre-specified desired statistical performance level (analogous to pre-specifying a desired Type 1 error level). $S_i = max(0, S_{i-1} + W_i)$, where S_0 is defined to be 0 and W_i is the sample weight or score assigned to the ith observation on the basis of the observed outcome of interest. For example, W_i may be specified as the *log*(LRT) comparing baseline risk to a pre-specified increased risk. Moustakides showed

that the *log*(LRT) sample weight is the optimal choice for W_i for detecting shifts in the odds ratio (*OR*).[19]

At each test point, two decisions are possible:

1) If $S_i \geq B$: stop monitoring and signal
2) If $0 \leq S_i < B$: continue monitoring

where *B* is the upper signaling threshold for the plot.

Because the CUSUM procedure will eventually signal with probability one in most practical process control settings, the usual criteria for the evaluation of test procedures (false positive and negative errors) are not appropriate for assessing the performance of CUSUM. Instead, the performance of the CUSUM chart is determined by the average run length (ARL) defined to be the expected length of time (or, equivalently, the expected number of units observed) before a signal occurs. The in-control average run length (IC-ARL) is the ARL under H_0, when the process is in control, and should ideally be long since a signal would represent a false alarm. The out-of-control average run length (OC-ARL) is the ARL under H_A and should be short since we want to detect a change in the process as soon as possible when it occurs. IC-ARL is analogous to Type 1 error, and OC-ARL is analogous to Type 2 error. Determining the ARL for a particular threshold, *B*, is computationally intensive, but it can be approximated using Markov chain methods.[20] Ideally, one chooses a control limit, *A*, so that the IC-ARL is relatively long (i.e. thus yielding fewer false positive errors).

As an example, suppose we are interested in detecting a certain increase in the odds of surgery failure relative to the baseline odds using the *OR*. The relevant hypotheses are H_0: $OR = OR_0$ (usually 1) versus H_A: $OR = OR_A$. Let X_i denote surgical failure for the ith subject from a binomial distribution with the probability of surgical failure for patient *i* equal to p_i. The odds of failure for patient *i* under H_0 and H_A, respectively, are $\frac{OR_0 p_i}{(1-p_i)}$ and $\frac{OR_A p_i}{(1-p_i)}$, and the probabilities of failure are p_i and $\frac{OR_A p_i}{(1-p_i+OR_A p_i)}$. Thus, the log-likelihood ratio weight for patient *i* is:

$$W_i = \log\left(\frac{(1-p_i+OR_0 p_i)OR_A}{(1-p_i+OR_A p_i)OR_0}\right) \text{ if } X_i = 1 \text{ (if surgery fails), or}$$

$$W_i = \log\left(\frac{1-p_i+OR_0 p_i}{1-p_i+OR_A p_i}\right) \text{ if } X_i = 0 \text{ (if surgery succeeds)}$$

Thus, W_i is based on 1) the observed Bernoulli outcome, X_i, 2) the baseline probability of failure, p_i, and 3) a specified change in the failure probability that is of importance, *OR*. A CUSUM plot looks very much like the plot of SPRT decision rules shown in Figure 1.

The main difference is that, instead of having both an upper and lower thresholds, CUSUM has only an upper threshold at B and a lower barrier set at 0.

2.2.3 Applications

In addition to production quality, CUSUM has been used to monitor procedures in clinical chemistry[20] and to monitor rare congenital malformations.[21] A common application of CUSUM methods in medicine has been to monitor surgical performance,[22-24] and the concept of 'adjusting for case-mix' in the context of signal detection was first considered in this setting.[22] Specifically, to account for the fact that baseline mortality risk before surgery varies across patients, an adjustment for prior risk can be made to avoid incorrectly attributing an increase in mortality rate that is due to patient risk differences to the surgeon. Hypotheses based on the *OR*, for example, can be defined based on an *OR* that incorporates baseline risk differences between patients. A bivariate CUSUM methodology designed to simultaneously monitor two outcomes has also been developed[25] as well as other variations of risk-adjusted CUSUM including resetting SPRT (RSPRT) and fast initial response (FIR) CUSUM, which are discussed by Grigg et al.[26]

2.2.4 Advantages

Like SPRT, chart-based methods provide useful visual summaries of the monitoring process and are relatively simple to implement. They are slightly more complex computationally compared to SPRT since the thresholds do not have a simple analytic approximation and must be obtained via Markov chain methods. Other advantages include the following:

- Chart-based methods are more flexible than SPRT and can potentially be better tailored to specific applications:
  - The sample weight, W_i, is purposely general rather than being defined as a specific statistic, such as the *log*(LRT) for SPRT. Different choices for W_i can thus be made depending on the desired setting. Although specifying W_i as the *log*(LRT) has been shown to be optimal for settings in which detecting shifts in the *OR* is of interest, different values or statistics may be used for W_i in other settings.
  - The sample weight, W_i, may be defined using cumulative historical information from previous observations or just using recent prior history. For example, CUSUM depends on the sum of the entire process history while Shewhart chart evaluation depends on a recent subgroup.
- Chart-based methods need not be performed on a strictly continuous basis. For instance, in production monitoring, data are analyzed after groups of a pre-specified size are observed.

2.2.5 Limitations

The primary limitation of chart-based methods that derive from process control such as the CUSUM is that the type of signal detection for which they are designed is not the same type of signal that is generally of interest in a setting like the Sentinel Initiative. In particular, CUSUM is designed to detect a change in a process over time (i.e. to detect a temporal change in risk within the same population). The process is assumed to begin in a state of 'control' and may later shift to an 'out-of-control' state at a later time point. This is in contrast to the Sentinel Initiative's primary purpose, which is to detect a difference (that is not expected to change over time) in adverse event risk between a group exposed versus unexposed to a new medical product. One exception in the context of post-licensure safety when a method like CUSUM may be of interest, is if one wants to detect an elevated safety risk that may be due to a problem with a specific drug or vaccine lot. In this instance, one would expect to see an elevated risk only at certain points in time (i.e. during uptake of the 'bad lot'), and the CUSUM would be suited to detect this type of increase.

In addition, chart-based methods share many of the same limitations as SPRT:

- Potential lack of accuracy (due to testing of a simple alternative hypothesis) and an unpredictable length of surveillance (since B is not guaranteed to be exceeded)
- They are designed for continuous testing which may not be feasible or desirable in all settings.
- The signaling threshold is restrictive in that it is constant (flat) over time.
- There is a lack of interpretability and transparency in determining the stopping threshold.
- There is relatively limited methodology to handle confounder adjustment.

Other disadvantages include the following:

- Chart-based methods are designed for monitoring a process *until a signal is detected.* There is no stopping rule to allow acceptance of the null hypothesis (of no signal).
- Different chart methods are better for detecting different types of potential signals, but no single method is best for all signal types. For example, CUSUM does not detect large abrupt shifts as fast as a Shewhart chart. Shewart charts are sensitive to large process shifts but do not detect small shifts well.
- Comparison of performance with other sequential methods, such as SPRT, is more complicated because the statistical performance criteria (i.e., the ARL), although analogous to standard hypothesis testing criteria (Type 1 and 2 errors), are not the same.
- Calculation of ARL and the signaling threshold is complex but possible.
- Separate chart analyses are needed if a 2-sided hypothesis test is desired.

2.3. Maximized sequential probability ratio test

2.3.1 Overview

The maximized sequential probability ratio test (maxSPRT)[6] a continuous sequential testing method that extends the SPRT to test a composite rather than a simple alternative hypothesis. For example, one may test H_0: $RR = R_0$ versus H_A: $RR > R_O$. In addition, maxSPRT allows comparisons of risk to be made between an exposed group and concurrently matched controls (by assuming binomial distributions in the two groups and comparing the probabilities of events) as well as between an exposed group and historical controls (by assuming a Poisson distribution in the exposed group and comparing that rate with the rate estimated from historical controls). Like SPRT, maxSPRT involves computing a LRT statistic continuously in time after each new subject (or matched pair) enters the study and has an upper signaling threshold designed to ensure that the Type 1 error rate is maintained at a pre-specified level. In addition, maxSPRT has an 'end of study' threshold, a preset point in time after which (if no signal has been reached) the study terminates and fails to reject H_0.

2.3.2 Statistical description

Kulldorff proposed the maxSPRT of a null hypothesis $H_0 : f = f_0$ versus a composite alternative hypothesis $H_A : f \neq f_0$ based on independent observations $X_1, X_2, \ldots$ having a common density function f.[6] Usually, the observations are known to be from a specific family f with an unknown parameter θ, and so the null and alternative hypotheses become $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$. The alternative can also be one-sided, $H_A : \theta > \theta_0$ or $H_A : \theta < \theta_0$. As the ith observation arrives, the log-likelihood ratio statistic Λ_i is calculated:

$$\Lambda_i = \log\left(\frac{\sup_{\theta\in\Theta_A\cup\Theta_0} L(X_1, X_2, \ldots, X_i \mid \theta)}{\sup_{\theta\in\Theta_0} L(X_1, X_2, \ldots, X_i \mid \theta)}\right) = \log\left(\frac{\prod_{j=1}^{i} f(X_i \mid \hat{\theta})}{\prod_{j=1}^{i} f(X_i \mid \hat{\theta}_0)}\right),$$

where Θ_0, Θ_A are parameter spaces under H_0, H_A, and $\hat{\theta}_0, \hat{\theta}$ are the maximum likelihood estimates (MLEs) of θ in $\Theta_0, \Theta_0 \cup \Theta_A$. The stopping rule is as follows:

- $\Lambda_i \geq B$: stop monitoring and accept H_A
- $\Lambda_i < B$, $i < N$: continue monitoring
- $\Lambda_i < B$, $i = N$: stop monitoring and accept H_0

where $0 < B < \infty$, $N < \infty$. N is a pre-specified upper limit of monitoring. If N subjects accumulate and no signal is detected, the study terminates and H_0 is accepted. B is chosen to maintain a pre-specified Type 1 error and is computed through simulation.

For example, let $X_1, X_2, \ldots$ from Bernoulli(p) denote the adverse event status (yes or no) for vaccinated subjects. In unvaccinated population, the probability of an adverse event is p_0. We want to test whether vaccinated subjects have a higher probability of getting adverse events than unvaccinated subjects ($H_0 : p = p_0$ versus $H_A : p > p_0$). The desired

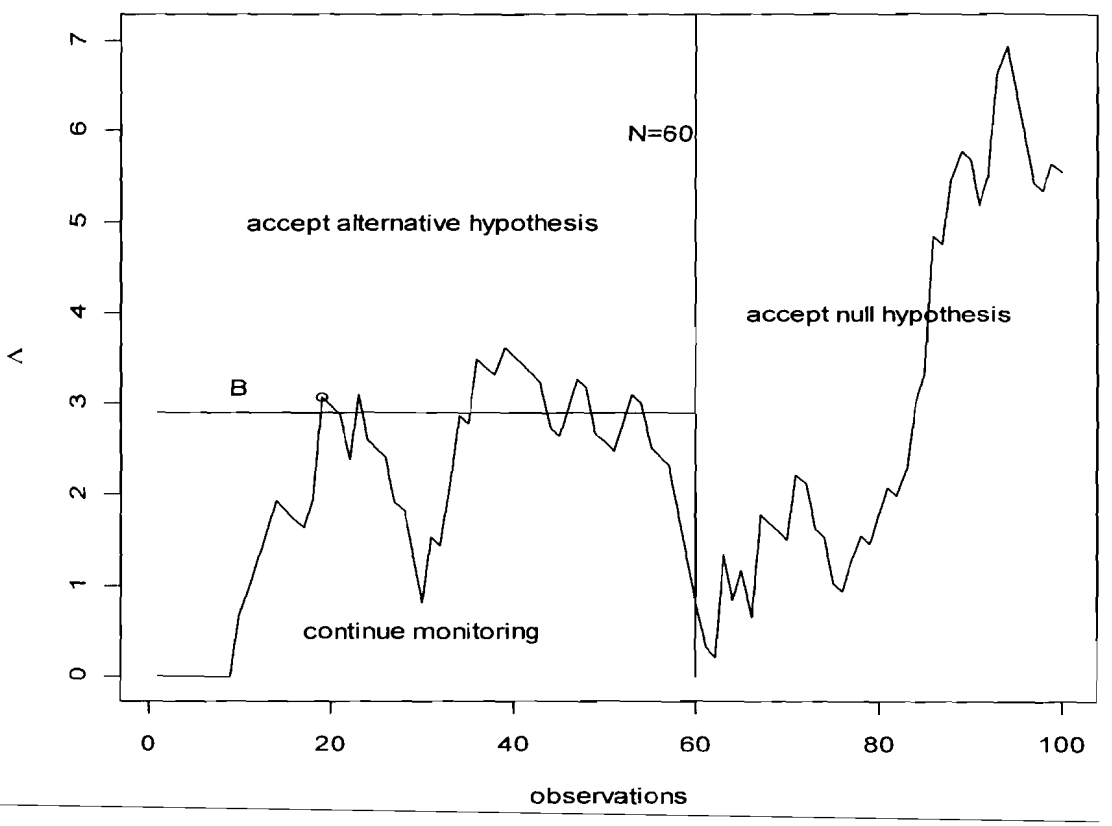
Type 1 error and power are 0.05 and 0.8, respectively. The test is equivalent to $H_0 : RR = 1$ versus $H_A : RR > 1$ with known p_0, and the log-likelihood ratio Λ_i is

$$\Lambda_i = \log\left(\frac{\sup_{p \geq p_0} \prod_{j=1}^{i} p^{X_j}(1-p)^{1-X_j}}{\prod_{j=1}^{i} p_0^{X_j}(1-p_0)^{1-X_j}}\right) = \log\left[\frac{\sup_{p \geq p_0} p^{\sum_{j=1}^{i} X_j}(1-p)^{i-\sum_{j=1}^{i} X_j}}{p_0^{\sum_{j=1}^{i} X_j}(1-p_0)^{i-\sum_{j=1}^{i} X_j}}\right]$$

$$= \begin{cases} (\sum_{j=1}^{i} X_j)\log(\frac{\sum_{j=1}^{i} X_j}{ip_0}) + (i - \sum_{j=1}^{i} X_j)\log(\frac{i - \sum_{j=1}^{i} X_j}{i(1-p_0)}), & \sum_{j=1}^{i} X_j \geq ip_0 \\ 0, & \sum_{j=1}^{i} X_j < ip_0 \end{cases}$$

since the MLE of p under $p \geq p_0$ is $\max(\frac{\sum_{j=1}^{i} X_j}{i}, p_0)$. We accept H_A if $\Lambda_i \geq B$ and accept H_0 if we have not exceeded B when we reach the Nth subject. A hypothetical example of the maxSPRT stopping threshold is shown in Figure 2. The likelihood ratio statistic exceeds the upper threshold at the 19th observation, so in this example we accept H_A.

Figure 2: MaxSPRT thresholds and observed likelihood ratio statistic values (Λ) from a hypothetical example.



2.3.3 Applications

MaxSPRT was first proposed for vaccine safety surveillance. It was the second approach used within the VSD[10] and has also been considered for application in longitudinal drug safety surveillance in the HMO Research Network's Center for Education and Research on Therapeutics (CERT). [27-28] Summaries of both the methodological and practical challenges that have arisen when applying maxSPRT for medical product safety surveillance based on input from statistician, epidemiologist, and clinician investigators from the VSD and CERT can be found in Appendices A1-3.

2.3.4 Advantages

The main advantage of maxSPRT compared to other continuous methods is that it tests for a composite versus a simple alternative hypothesis, which prevents the signaling accuracy problems that are associated with other methods. The maxSPRT is also computationally simply and easy to implement as Kulldorff has provided a comprehensive table of signaling threshold (i.e., critical) values for a range of potential RRs of interest and a range of expected number of adverse events likely to be observed.[6] Like SPRT, maxSPRT can be charted to provide useful visual summaries of the monitoring process. The maxSPRT also offers increased flexibility compared to other continuous methods by allowing for comparisons of risk to be made between an exposed group and concurrently matched controls as well as between an exposed group and historical controls.

2.3.5 Limitations

Although more accurate and flexible than SPRT, maxSPRT shares some of the same limitations as SPRT. First, maxSPRT is designed for continuous testing, and thus it performs optimally only in that setting. For instance, the VSD implements maxSPRT by performing a LRT test on a weekly (rather than a strictly continuous) basis. For weekly testing, the maxSPRT stopping thresholds (which control for a fixed Type 1 error rate under the assumption that continuous testing is being performed) are too conservative (i.e. too high) As an illustration, assume a 1-year study with μ expected events under the null hypothesis of no increased risk among recipients of a new medical product. Table 1 shows the stopping thresholds for maxSPRT (i.e. a LRT with continuous testing) and the stopping thresholds for a sequential LRT conducted on a weekly basis. Both thresholds are designed to hold the same Type 1 error rate of 0.05. In order for maxSPRT to hold the same Type 1 error rate as a weekly LRT test, maxSPRT requires a higher threshold since it allows for more frequent testing (i.e for maxSPRT, there are more opportunities for false positive errors). The data in Table 1 show that if weekly testing is performed, the thresholds can be set considerably lower than the maxSPRT boundaries particularly as the expected number of events, μ, increases. In other words, use of continuously-designed maxSPRT thresholds when testing is actually performed on a weekly basis means that higher-than-necessary thresholds are used. This yields a Type 1 error rate that is lower than desired and suboptimal statistical power.

Table 1. Stopping thresholds for a sequential LRT: continuous (maxSPRT) vs. weekly testing

Expected number of events	maxSPRT	Weekly LRT
μ =1	2.86121	2.70944
2	3.05217	2.55459
5	3.30220	2.88088
10	3.47399	2.88088
20	3.63490	3.03593
50	3.82760	2.95867
100	3.95952	3.05170
200	4.07969	3.12476
500	4.22903	3.13165
1000	4.33164	3.14484

A second potential restriction of maxSPRT is that it involves a flat signaling threshold over time. Compared to other potential threshold shape choices, a flat threshold may result in more false positives early on in a study (when there is less data and inherently more variability) and less power at later time points in the study. Third, the maxSPRT stopping thresholds are computed on the scale of the LRT, which can be difficult to interpret and yields a somewhat obtuse stopping threshold selection process. Specifically, when designing a safety surveillance study using maxSPRT, one must pre-specify the number of adverse events that are expected to occur during the study period and this then determines the LRT signaling threshold. However, it may be difficult for an investigator to estimate the expected number of events in advance. Further, the relationship of this quantity to the resulting stopping threshold is mathematically complex, and the resulting value of the stopping threshold on the scale of the LRT is not clinical meaningful. Therefore, as currently implemented, it is difficult to assess the operating characteristics of the maxSPRT threshold in advance. Alternatively, threshold selection could be based directly on the size of the *RR* that represents a concerning risk difference for which an investigator would want to signal a safety problem.

A fourth limitation is that, as with other continuous monitoring methods, accommodation of confounders when comparing an exposed group to a concurrently matched control group is achieved by the binomial-based maxSPRT through stratification. This may provide limited adjustment compared to a more continuous regression-based adjustment method if stratification is not fine enough. Conversely, when very fine stratification/matching categories are used for the binomial-based maxSPRT, there is potential for a loss of information due to an inability to identify exposed-unexposed matches within such narrowly defined strata. Several investigators within the VSD, including members of our group at Group Health, are currently developing methods to generalize maxSPRT to more comprehensively adjust for confounders (i.e. using a regression-based covariate-adjustment approach) and to avoid the information loss

problem that may occur if matching is too restrictive (i.e. broaden from a matched sampling design to an unmatched cohort design).

Last, the Poisson-based maxSPRT which involves comparing the adverse event rate in an exposed group with a rate estimated using historical controls, is fairly sensitivity to misspecification of historical rate. This sensitivity occurs because maxSPRT assumes that the historical comparison rate is known and does not acknowledge the uncertainty in this estimated value. This becomes particularly problematic when the amount of available historical data upon which to estimate the historical rates is relatively small, as can often be the case with rare adverse events within pre-specified strata of interest. Kulldorff and his Harvard colleague Lingling Li are currently developing a new extension to maxSPRT that addresses this issue.

3 Group Sequential Monitoring Methods

3.1 Overview

In contrast to continuous monitoring methods, group sequential monitoring involves analyzing data after a group of subjects accrues in the study, rather than analyzing the data after each individual subject accrues. Group sequential designs can be considered to be a generalization of the continuous monitoring approach. Specifically, continuous monitoring is a special case of a group sequential design where each individual is its own 'group.' In group sequential monitoring, the size of the groups analyzed or equivalently the frequency with which monitoring is performed can be varied depending on the desired monitoring specifications. In clinical trials, for example, many group sequential stopping designs have been developed and are widely used as described in the recent reviews by Emerson et. al.[7-8] As with continuous testing methods, signaling thresholds in group sequential designs are computed to account for the multiple testing that maintains an overall pre-specified Type 1 error rate.

3.2 Statistical description

The key feature of a group sequential test is that, instead of testing after each new subject accrues, as is done for continuous monitoring methods such as SPRT and maxSPRT, analyses are conducted after a group of subjects accumulates. In addition, the stopping thresholds can be specified to vary over time, if desired. Two commonly used thresholds are the Pocock[29] threshold, which is constant over time, and the O'Brien-Fleming[30] threshold, which is wider at earlier time points and narrows at later time points. Group sequential tests can be applied to a wide variety of hypotheses testing H_0 versus H_A, and several standard test statistics of these hypotheses can be used. The desired Type 1 and 2 errors and the total number of planned tests, T, are specified in advance. If testing is planned after m subjects accumulate, as the ith group of m observations arrive, the test statistic, Z_i, is calculated based on these $i*m$ subjects, and the stopping rule is as follows:

- $|Z_i| \geq B_i(T,\alpha)$: stop monitoring and accept H_A

- $|Z_i| < B_i(T,\alpha)$, $i < T$: continue monitoring
- $|Z_i| < B_i(T,\alpha)$, $i = T$: stop monitoring and accept H_0

where $0 < B_i < \infty$, $T < \infty$. B_i is a pre-specified value, representing the threshold at the ith test, that depends on the number of tests planned, T, and the desired Type 1 error, α. Less frequent testing (smaller T) will result in a lower B_i. The Pocock threshold is constant over time: $B_i(T,\alpha) = B_P(T,\alpha)$, $i = 1,2,...,T$. The O'Brien-Fleming threshold assumes that $B_i(T,\alpha) = B_{OF}(T,\alpha)\sqrt{T/i}$, $i = 1,2,...,T$, which is a decreasing function over time. Wang and Tsiatis[31] proposed a family of thresholds, which encompass Pocock and O'Brien-Fleming, indexed by a parameter Δ. Changing this parameter allows different thresholds shapes: $B_i(T,\alpha) = B_{WT}(T,\alpha,\Delta)(i/T)^{\Delta-1/2}$, $i = 1,2,...,T$, $\Delta \in [0,1/2]$. The Pocock and the O'Brien-Fleming thresholds are special cases with $\Delta = 1/2$ and $\Delta = 0$, respectively.

For example, suppose we are interested in testing whether vaccinated subjects have a higher probability of having a certain adverse event than unvaccinated subjects, as described previously for the maxSPRT example. At the ith test, the likelihood ratio statistic is the same as calculated before. However, now instead of i subjects, we have $i*m$ subjects:

$$\Lambda_i = \log\left(\frac{\sup_{p\ge p_0}\prod_{j=1}^{im} p^{X_j}(1-p)^{1-X_j}}{\prod_{j=1}^{im} p_0^{X_j}(1-p_0)^{1-X_j}}\right) = \log\left[\frac{\sup_{p\ge p_0} p^{\sum_{j=1}^{im}X_j}(1-p)^{im-\sum_{j=1}^{im}X_j}}{p_0^{\sum_{j=1}^{im}X_j}(1-p_0)^{im-\sum_{j=1}^{im}X_j}}\right]$$

$$= \begin{cases} (\sum_{j=1}^{im}X_j)\log(\frac{\sum_{j=1}^{im}X_j}{imp_0}) + (im-\sum_{j=1}^{im}X_j)\log(\frac{im-\sum_{j=1}^{im}X_j}{im(1-p_0)}), & \sum_{j=1}^{im}X_j \ge imp_0 \\ 0, & \sum_{j=1}^{im}X_j < imp_0 \end{cases}.$$

Since $2\Lambda_i \sim \chi_1^2$, rejecting H_0 if $|Z_i| \ge B_i(T,\alpha)$ is equivalent to $2\Lambda_i \ge B_i(T,\alpha)^2$. Critical values $B_i(T,\alpha)$ can be calculated through simulation. We accept H_A if $2\Lambda_i \ge B_i(T,\alpha)^2$, and accept H_0 if we don't detect any signal until the Tth test. A hypothetical example using both a Pocock and an O'Brien-Fleming threshold and T=5 is shown in Figure 3.

3.3 Applications

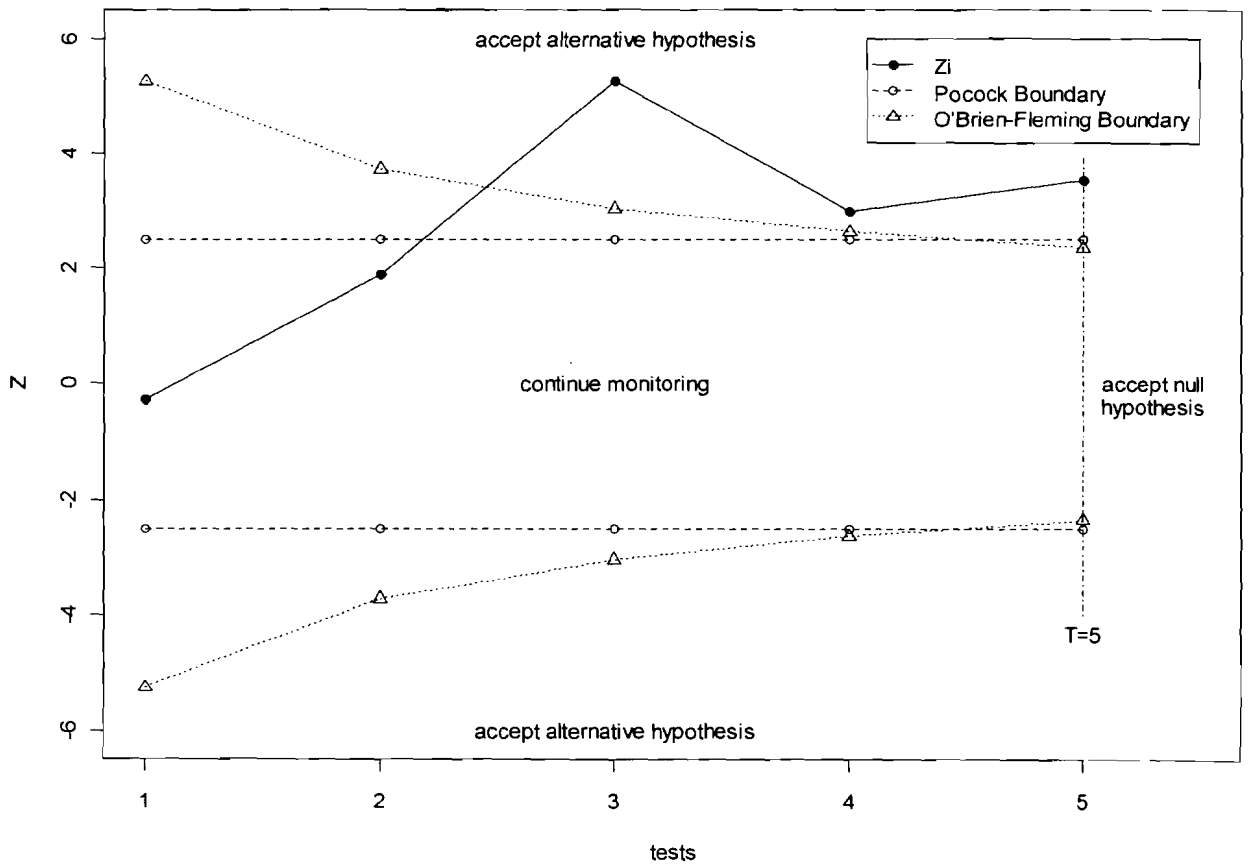
Group sequential testing methods are widely used in randomized clinical trials to monitor the efficacy and safety of new medical products. Clinical trial sequential testing procedures were originally proposed by Armitage[32] and considerable methodological developments have been made since that time (see texts by Jennison and Turnbull[33] and Whitehead[34]). For instance, flexible and wide-ranging stopping rules have been proposed based on both scientific and statistical criteria and using different interim monitoring

schedules and adaptive designs. Emphasis has been given in these contexts to fully investigating the operating characteristics of a particular stopping rule to ensure that the design choice is satisfactory from a clinical, statistical, and ethical perspective.[7-8]

3.4 Advantages

Group sequential testing is advantageous because it can result in lower expected sample sizes and a shorter average study length than a conventional fixed sample study (where a single analysis is performed at the end of study follow-up). Compared with a continuous sequential testing approach, the group sequential testing framework offers flexibility in the size of the groups analyzed or equivalently the frequency with which monitoring is performed. And, in general, less frequent testing leads to increased overall study power. In addition to flexibility in the frequency with which data are monitored, group sequential monitoring methods offer flexibility in the shape of the signaling threshold over time as well as the type hypothesis and statistic/criteria used for hypothesis testing.[7-8]

Figure 3: Pocock and O'Brien-Fleming thresholds with T=5 tests and observed likelihood ratio statistic values from a hypothetical example.



3.5 Limitations

Although careful study of a variety of group sequential designs has been performed in randomized clinical trial settings, their use in active observational safety surveillance, in which more frequent (e.g. monthly, weekly) safety monitoring may be desired and where confounding is a potential concern, has not been considered. In fact, there are many unanswered questions about how to best apply a group sequential approach in an active observational safety surveillance setting:

- How frequently should sequential testing be performed? Should testing frequency change over time?
- What signaling threshold shape is most desirable?
- What test statistic is best for quantifying risk differences between exposed and unexposed groups?
- Should these design choices vary by the adverse event of interest? If so, how? (e.g., based on known risk-benefit information, prevalence of the outcome, severity of the outcome)

To guide such design decisions, statistical information about the operating characteristics of a variety of group sequential designs of potential interest in an active surveillance setting is needed. In particular, conducting simulation studies designed to compare different group sequential designs (i.e. designs with different testing frequencies, signaling threshold shapes, and test statistics for quantifying risk differences) based on relevant statistical criteria including power (i.e., the probability of correctly concluding that an exposure is unsafe) and timeliness (i.e., the time to correct detection of safety problem) would be useful. For example, although it is apparent that testing less frequently (i.e., in a group sequential fashion) will yield higher overall statistical power and longer time-to-signal detection than testing continuously, specific power-timeliness trade-offs associated with particular degrees of less frequent testing have not been studied. Members of our team at Group Health are currently conducting this type of comprehensive simulation study evaluation in order to develop a much needed formal statistical framework for optimally designing group sequentially monitored observational safety surveillance studies.

A second limitation of group sequential testing methods in the context of active observational safety surveillance is that confounder adjustment methods are limited. In clinical trials, randomization generally equalizes confounder differences between comparison groups. In an active observational surveillance setting, however, group sequential analysis methods need a robust mechanism to adjust for potential confounders since comparison groups are not randomized by design. VSD investigators, including members of our team at Group Health are currently developing such methodological advances.

A final limitation of a group sequential framework compared to maxSPRT is that it is inherently more complex. No single table of existing signal thresholds can be used as the threshold level will vary by testing frequency and possibly over time. However, signaling

thresholds for a given group sequential design can be obtained in a fairly straightforward fashion via simulation. This is often done in sequentially monitored clinical trial planning. Further, although there is added complexity in designing group sequential monitoring plan that varies depending upon the specific adverse event outcome and medical product being monitored, a tailored group sequential design will likely better achieve the desired scientific goals compared to a one-size-fits-all design approach.

4 Data Mining Methods

4.1 Overview

Data mining is a technique of extracting hidden associations or patterns of association in large datasets when manual inspection is not feasible. Data mining methods are generally aimed for hypothesis generation and signal strengthening rather than hypothesis testing. The data mining methods that are most commonly used in pharmacovigilance include the proportional reporting ratio (PRR)[35], a Bayesian confidence propagation neural network (BCPNN)[36,37], and the multi-item gamma Poisson shrinker (MGPS).[1,2] All of these methods are based on a quantitative measure of disproportionality which compares the observed with the expected reports of a certain outcome-exposure combination. The observed number of reports of a particular outcome-exposure combination is compared to an estimate of the expected number based on all other reports in the dataset (i.e. reports for other adverse event outcomes and exposures). Therefore, all other exposures and outcomes in the dataset are used as the comparison group. The dataset can be restricted to a subset of exposures and adverse event outcomes in order to achieve the desired comparison group. The PRR, BCPNN, and MGPS each derive a different statistical measure to quantify the ratio of observed-to-expected reports. Thus, they each have a different approach to statistically determine which outcome-exposure combinations generate a signal indicative of a safety problem.

4.2 Statistical description

4.2.1 Proportional reporting ratio (PRR)

The Proportional Reporting Ratio (PRR) is a very simple statistical tool that was developed by Evans et. al[35] to generate a signal of a possible unrecognized hazard from spontaneous adverse event reporting data (Table 2).

Table 2. Structure of data used by the PRR and other data mining methods

	Drug of interest	All other drugs in database	
Adverse event (s) of interest	a	b	a+b
All other adverse events	c	d	c+d
	a+c	b+d	N

where

a = number of a specific adverse event for the drug of interest
a+b = total number of that specific adverse event for all drugs in the dataset
a+c = total number of reports of the drug of interest in the dataset
N = total number of reports of any drug-adverse event in the dataset

For a given drug of interest, the proportion of a pre-specified adverse event (or group of adverse events) among all events related to that drug is compared to the proportion of that same adverse event(s) among all events related to all other drugs in the database. The ratio of these two proportions is called the PRR and is defined as follows:

$$\text{PRR} = \frac{a/(a+c)}{b/(b+d)}.$$

Measures of statistical association may be calculated using a chi-squared test on one degree of freedom with Yate's correction.

Three pieces of information are used to define the signaling threshold for a specific adverse event related to the drug of interest: 1) the value of the PRR, 2) the value of the chi-squared statistic based on the 2X2 table data, and 3) the total number of adverse event(s) reported. Specifically, if the PRR is at least 2, the chi-squared statistic is at least 4, and number of adverse events is at least 3, then a signal is generated.

4.2.2 Bayesian Confidence Propagation Neural Network (BCPNN)

Bate et al. described a Bayesian confidence propagation neural network (BCPNN) for adverse event signal detection.[36,37] The method uses a fully Bayesian framework, which enables the method to account for uncertainty when a small number of drug-adverse event combinations is observed. The strength of the dependency between a drug and an adverse event is measured by the information component (IC) defined as the logarithm of the ratio of the observed rate of a specific drug-adverse event to the expected rate of adverse event under the null hypothesis of no association between the drug and the event. Using the same notation as in Table 2, the crude estimate of IC, $\hat{I}C_0$, without using a Bayesian modeling framework is defined as follows:

$$\hat{I}C_0 = \log_2 \frac{a/(a+c)}{(a+b)/N},$$

Therefore, IC is defined as

$$\log_2 \frac{P(\text{event is y} \mid \text{drug is x})}{P(\text{event is y})}.$$

By incorporating a Bayesian framework, Bates et. al. use the property that under the null hypothesis of no association between drug x and reaction y,

$$P(\text{drug is x, event is y}) = P(\text{drug is x})P(\text{event is y})$$

Therefore,

$$\text{IC} = \log_2 \frac{P(\text{event is y} \mid \text{drug is x})}{P(\text{event is y})} = \log_2 \frac{P(\text{drug is x, event is y})}{P(\text{drug is x})P(\text{event is y})}$$

where

P(drug is x, reaction is y) = probability of a specific drug-adverse reaction combination (x and y) among all adverse reaction reports,

P(drug is x) = probability of drug, x, among all adverse reaction reports, and

P(reaction is y) = probability of adverse reaction, y, among all adverse reaction reports

Bates et. al. assume that P(drug is x, reaction is y), P(drug is x), and P(drug is y) each follows a beta-binomial distribution, and IC is estimated using the posterior mean and variance from a fully Bayesian model specification. A valuable feature of this type of estimate, compared to the crude estimate for IC, is that if a drug has very few reports of the adverse event of interest (i.e., if the sample size is small), then the posterior mean estimate of IC will be shrunk accordingly towards the null to reflect this uncertainty due to the limited amount of available data. The crude estimate of IC does not have this feature and thus can be more subject to false positive signals when sample sizes are small. When sample sizes are large, then the BCPNN posterior mean estimate of IC will be similar to the crude estimate of IC. A 95% credible interval for the posterior mean estimate of IC is calculated to provide a measure of robustness. If the lower 95% limit of that interval is greater than zero, a signal is generated.

4.2.3 Multi-Item Gamma-Poisson Shrinker (MGPS)

DuMouchel[1-2] developed a Bayesian data mining method called the Multi-item Gamma Poisson Shrinker (MGPS) that is designed for analysis of a large outcome-by-exposure frequency table that has many cells. His work was motivated by interest in analyzing the FDA's AERS database, which is essentially a frequency table containing millions of cell counts of spontaneous drug-adverse event reports. The goal of his method is to find the 'interestingly large' cell counts in a large frequency table for possible further evaluation. DuMouchel recommended a signaling criterion based on the results of fitting an empirical Bayes model to the cell counts. It is similar to the BCPNN data mining method except that MGPS uses a more robust empirical Bayesian approach that can account for covariate adjustment through stratification.

The quantity of interest is the logarithm of the ratio of observed number of reactions of type y that are related to drug x (i.e. the quantity 'a') from Table 2 over the expected number of events (E_{xy}):

$$\log_2 \hat{RR} = \frac{a}{E_{xy}},$$

where the expected number of events is calculated as

$$E_{xy} = \sum_z \frac{N_{xz} N_{yz}}{N_z}$$

and

x = drug of interest
y = adverse event of interest
z = stratum such as age and/or gender
N_{xz} = total number of reports with the drug of interest (x) in stratum z
N_{yz} = total number of reports with the adverse event of interest (y) in stratum z
N_z = total number of reports in stratum z

Using the same notation as in Table 2,

$$E_{xy} = \sum_z \frac{(a_z + b_z)(a_z + c_z)}{N_z}.$$

Note that without stratification and without incorporating the empirical Bayesian modeling,

$$\log_2 \hat{RR} = \log_2 \frac{a}{(a+b)(a+c)/N} = \log_2 \frac{a/(a+c)}{(a+b)/N} = \hat{IC}_0.$$

The method further assumes that the observed count for each drug-event combination, N_{xy}, is drawn from a Poisson distribution with an unknown mean, μ_{xy}. Additionally, the ratio, μ_{xy}/E_{xy}, is assumed to be drawn from a prior distribution which is a mixture of two gamma distributions. An empirical Bayes estimate of the *RR*, the empirical Bayes geometric mean (EBGM), is obtained from the model. It is used to rank all cell counts to determine which cells have unusually large observed counts compared to the expected counts. Similar to the BCPNN approach the EBGM can be seen as a 'shrinkage' estimate of the true relative risk for a particular drug-event combination. If the observed or expected number of events of drug x (a and E_{xy}, respectively) is large, then EBGM will be close to $\hat{RR}$, the crude RR, but otherwise EBGM will be shrunk towards the null value of 1. If the lower 95% credible interval of the posterior distribution is greater than 2, then a signal is generated. Compared to BCPNN, MGPS is based on a more flexible model with fewer assumptions that may fit data better. Although EBGM and BCPNN

approaches have not been directly compared, the major advantage for EBGM compared to the BCPNN is the ability to adjust for covariates through stratification.

4.3 Applications

The BCPNN method has been routinely used for signal detection in the World Health Organization (WHO) adverse drug reactions (ADR) database since 1998,[38,39] and a pilot investigation applying BCPNN to the Intercontinental Medical Statistics (IMS) Disease Analyzer-Mediplus UK clinical records database of about 2.4 million patients was also done.[40] The MGPS has been used in evaluating drug adverse reaction signals in the FDA's AERS database.[3,41] The MGPS method was first applied to the Vaccine Adverse Event Reporting System (VAERS) data to assess feasibility,[42] and later a simulation study was performed to assess the sensitivity and specificity of the MGPS data mining method within the VAERS database.[43] The MGPS method was also used to detect opioid analgesic abuse using data from a network of poison control centers.[44] The feasibility of adapting MGPS to analyze longitudinal administrative claims data has been evaluated, and the validity of this Bayesian method on claims data seems promising.[45] It is important to note, though, that this feasibility assessment involved applying MGPS at a single point in time to a longitudinal claims database that was 'collapsed' into a cross-sectional dataset. This is in contrast to repeatedly and sequentially applying MGPS at multiple points in time in prospective fashion.

4.4 Advantages

There are several advantages of the PRR, BCPNN, and MGPS data mining methods. First, since all three methods use a proportionate approach, and thus they avoid potential biases related to variable reporting (e.g., if the overall level of reporting is high for a new drug) and underreporting. Furthermore, these methods can easily handle concomitant medication use by conducting subgroup analysis (i.e., by varying the group of drugs of interest).

For the PRR, calculation and interpretation is straightforward, and it is easy to implement the method in standard software packages. Also, the PRR represents a direct measure of the strength of the signal. For the BCPNN and the MGPS, use of the Bayesian framework provides for estimate 'shrinkage' when the number of drug-adverse event reports is small, which can reduce the occurrence of false positive associations based on small sample variability. The empirical Bayes methodology used by MGPS is good at minimizing the effect of sampling variance on the interpretation of the relative reporting rate. A unique advantage of PRR and MGPS is that they can handle covariate adjustment through stratification. Stratification should be applied if confounding for the outcome-exposure association is known or suspected. In a study that applied PRR and MGPS data mining methods to the VAERS database, stratification applied to both methods reduced confounding and unmasked some vaccine-adverse event associations that the methods without stratification did not detect.[46]

4.5 Limitations

General Limitations of all three methods:

A major limitation of all three data mining methods is that the statistical properties (e. g., Type 1 error, sensitivity, specificity) of the standard signaling thresholds may vary when the methods are applied to different datasets. The standard thresholds typically used by regulatory agencies (e.g., the WHO) derive from the original papers in which each method was developed and are as follows:

- For PRR: PRR ≥ 2, chi-square test statistic ≥ 4, and number of adverse event reports ≥ 3
- For BCPNN: (IC value – 2 standard deviations) > 0
- For MGPS: the lower 95% credible interval of the posterior distribution > 2

However, the statistical performance of these thresholds when the methods are applied to other datasets is not known and may vary greatly. For example, Matsushita et. al. investigated the performance of standard thresholds for signal detection in relatively small (compared to AERS, for example) spontaneous adverse event reporting databases maintained by pharmaceutical companies.[47] They found that application of the standard thresholds for all three methods in these smaller databases yielded very low sensitivity (<50%). In a simulation study, they showed that lower thresholds are needed for these smaller spontaneous reporting datasets to achieve desired sensitivity levels. They also showed that different thresholds may be needed for more common versus rare outcomes, and this is not currently done in practice.

Another limitation of these data mining approaches is that a large association between an adverse reaction and a drug that is *not* of specific interest (i.e. for large values of 'b' in Table 2) can reduce the likelihood of detecting a true signal between that adverse event and the drug that *is* of interest. This situation could occur when a drug that is an indication for the adverse event of interest is included in the dataset (and thus included in the 'all other drugs' comparison group). A solution to this problem is to remove that drug from the dataset before calculating the test statistic. However, this requires knowing and identifying all such drugs in advance, which may not be realistic in practice since drug indications may not always be recorded.

Last, since these data mining methods were developed specifically for pharmacovigilance using large spontaneous adverse reporting databases where the denominator includes people who take <u>any</u> drug and experience <u>any</u> adverse event, these methods may not be the most appropriate data mining methods for safety monitoring in a prospective cohort like the Sentinel Initiative. It would be useful to consider how to specifically extend these methods to a prospective cohort data rather than to simply apply these methods in their existing form to cohort data. One such extension is currently in development by DuMouchel for the MGPS data mining method. In addition, Kulldorff is developing a tree-based data mining method for use in prospective longitudinal surveillance that explicitly controls for the overall false positive error rate (versus controls in *ad hoc* way as done by current approaches) and that searches for optimal outcome groupings of ICD-9 codes (rather than requires a pre-specified list of outcomes as current approaches do). It

may also be beneficial to review the broader statistical literature on data mining beyond these three methods as they may possess properties better-suited for prospective cohorts.

Specific Limitations of the PRR method:

There are several limitations to be noted that are specific to the PRR method. One major limitation is that the PRR cannot be calculated if there are no adverse events of interest reported for the comparison drug(s) (i.e., if b=0 in Table 2). Further, since the PRR is a ratio of two proportions, its value is unstable when sample sizes are small. Specifically, when the adverse event of interest is rare, the PRR can easily signal when there is only one adverse event following the drug of interest. Data in the following example showcase this limitation.

Table 3. Example dataset A for a rare adverse event of interest

	Drug of interest	All other drugs in database	
Reaction(s) of interest	1	4	5
All other reactions	9	9,996	10,005
	10	10,000	10,010

In this case with only 10 individuals reporting use of the drug of interest and only 1 adverse event of interest related to that drug, the PRR is $\frac{1/10}{4/10{,}000}$=250, which indicates a very strong signal, well-above the standard threshold of 2. Even when the number of observed drug reports increases from 10 to 100, the PRR=$\frac{1/100}{4/9{,}910}$=24.8, is still very large and indicates a strong signal (Table 4).

Table 4. Example dataset B for a rare adverse event of interest

	Drug of interest	All other drugs in database	
Reaction(s) of interest	1	4	5
All other reactions	99	9,906	10,005
	100	9,910	10,010

Thus, the PRR method is more likely to yield a false signal than the Bayesian data mining methods.[48-51] Simple corrections to the PRR were initially attempted to solve this problem, but the BCPNN and MGPS data mining methods, which were developed in part due to this problem with PRR, are preferred.

5 Bayesian updating

5.1 Overview

Bayesian updating is a method that has been explored recently as a possible safety monitoring methodology for interventional cardiovascular procedures.[52,53] Typically,

very limited pre-licensure safety information for new interventional cardiology devices is available due to small study sample sizes and the low occurrence of specific adverse reactions in clinical trials. In addition, pre-licensure clinical trial patient populations can be very select, often including healthier patients than those who may actually receive the new device in clinical practice. Thus, methods for continual post-licensure monitoring of new devices have been explored, including Bayesian updating.

Bayesian updating is a method that incorporates prior knowledge of adverse event rates (e.g., expert opinions, data from previous studies, data observed at earlier time points in the current study) to routinely monitor safety. Specifically, one recursively uses all available prior knowledge plus new data available at the current study time point to continually refine an estimate of adverse event risk in a population of interest. In statistical terms, a prior adverse event probability estimate is combined with current study data to obtain a posterior adverse event probability estimate for the current time point. This updating process is repeated to combine the new additional observed data at each pre-specified time point. Summary statistics (e.g., means, medians and 95% credible intervals from the posterior distribution) for the adverse event probability estimated at the final time point are compared to those estimated using the initial/prior data to determine if there is a higher than expected safety risk in the current study compared to initial estimates based on previous data. Adjustment for potential confounders is accomplished through stratification.

5.2 Statistical description

Resnic et. al. described the Bayesian updating method they applied to monitor the safety of 309 patients who underwent rotational atherectomy (RA) in a hospital setting.[52] There are four main steps involved in the process:

(1) Risk stratification
In order to account for patient characteristics or confounding factors, patients were stratified into different risk groups and separate prior and posterior estimates were calculated for each stratum. Risk-scores were computed for each study subject using previously developed models for predicting the risks of two outcomes of interest, death and the major adverse cardiac events (MACE). Based on these models, patients were classified into low, moderate, and high risk groups.

(2) Development of initial prior probability estimates
Initial prior probability estimates were developed separately for the three risk groups. Information on adverse events among patients undergoing RA was obtained from 5 previous studies of low risk patients. Prior risk estimates for moderate and high risk groups were extrapolated from the estimates for the low risk group. Since the risk of outcomes was modeled as a binomial process, a beta distribution (which is the conjugate prior distribution of the binomial distribution) was used. The probability density function for a binary outcome x under a beta distribution is

$$f(x \mid q, r) = \frac{1}{B(q,r)} x^{q-1}(1-x)^{r-1},\ 0 \le x \le 1,\ q > 0,\ r > 0$$

where B(q, r) is a beta function with $B(q,r) = \int_0^1 x^{q-1}(1-x)^{r-1}dx$

Using the mean and variance of the estimated adverse event probability from the 5 previously published studies, the two parameters in the beta prior distribution, q and r, can be easily calculated for the low risk group. To extrapolate estimates for the moderate and high risk groups, the relationship of the event risks in the low, moderate and high risk patients who underwent non-RA interventional procedures was used.

(3) Updating method to obtain posterior estimates
Bayesian updating was performed every 3 months to continually refine the adverse event risk estimate in each stratum. With a beta prior distribution and a binomial distribution for the outcomes, the posterior distribution is also a beta distribution under the Bayesian framework. At each pre-specified time point during the study, new observed data (e.g., the number of patients undergoing RA, the number of events) were combined with the prior (beta distribution) information by updating the two beta distribution parameters (q and r) from the previous time point and generating a new posterior beta distribution based on these updated parameters as follows:

$$q_{i+1}=q_i+k_i \quad \text{and} \quad r_{i+1}=r_i+n_i-k_i$$

where i=current time period; i+1=subsequent time period; n=number of new patients observed at time i, and k=number of new adverse events at time i. The updated (i.e., posterior) beta distribution parameters are then used to compute an updated estimate of the mean and variance of the adverse event risk for each risk group at each time point.

(4) Interpretation of results
Summary statistics for the adverse event risk (e.g., means, medians and 95% credible intervals from the posterior distribution) estimated at the final time point are compared to those estimated using the initial/prior data to determine if there is a higher than expected safety risk in the current study compared to initial estimates. Specifically, if the 95% credible intervals for the initial and final posterior distributions overlap, then no evidence for an excess risk of adverse events in the current study data compared to previous studies has been found. However, no formal statistical test is performed to compare the initial prior versus final posterior estimates.

5.3 Applications

Bayesian updating has been applied in the field of nuclear reactor engineering to monitor safety.[54] The method has also been recently applied to monitor the safety of interventional cardiovascular procedures using high quality clinical data and relatively smaller sample sizes, compared to sample sizes for spontaneous reporting data systems.[52]

5.4 Advantages

The main advantage of Bayesian updating is the incorporation of prior knowledge of the event risk with the current observed data. In addition, calculations are quite simple and straightforward, particularly when using a conjugate prior distribution (e.g. a beta prior

distribution and a binomial distribution for outcomes, which yields a beta posterior distribution). Other distributions, including discrete and continuous probability distributions, may also be used. Bayesian updating is best-suited to find differences between an initial/prior and a final/posterior estimate under two conditions: 1) when sufficient data accrue over time, and 2) when accurate prior estimates are obtained.[55]

5.5 Limitations

The main limitation of the Bayesian updating method is that the type of signal detection for which it is designed is not the same type of signal that is generally of interest in a setting like the Sentinel Initiative. In particular, similar to the CUSUM method, Bayesian updating is designed to detect a change over time in an estimated adverse event risk estimate as more study data accrue in a given population of interest. This is distinct from the primary purpose of the Sentinel Initiative, which is to detect a difference in the adverse event risk between a group exposed versus unexposed to a new medical product.

Another major challenge for the Bayesian updating method in monitoring the safety of a new medical product is accurate and precise estimation of the initial prior probability. For certain new products, there may not be adequate existing knowledge to inform prior probability estimation. Or, the population in which prior knowledge exists may not be representative of the population that will actually receive the new product in clinical practice. In these instances, it may be necessary to use a very rough prior estimate with wide confidence intervals. However, since the Bayesian updating signal detection criterion involves comparing the final posterior estimate to the initial prior estimate, the accuracy, precision, and assumptions used to estimate the prior have a great influence on the signal detection results. For instance, if there is considerable uncertainty in the initial prior estimate, a false negative result will be likely since it will be inherently difficult to detect any difference between a prior risk estimated with such high variability and a truly elevated final posterior estimate.

However, a false negative result can occur for the opposite reason as well. Specifically, it has been illustrated in a randomized clinical trial setting that Bayesian updating can also have a high false negative rate when 'too much' prior information is available.[56] In this case, the vast amount of information used to inform the prior can outweigh the relatively smaller amount of information observed in the current study. Thus, the posterior estimate doesn't incrementally change much from the initial prior estimate over time, and differences between the prior and posterior are therefore not readily detected.

Other limitations of the Bayesian updating method include the following:

- Lack of a formal test to detect a signal (beyond comparing means, medians and 95% credible intervals from the posterior and the prior distribution, which is heavily dependent on the initial estimates from existing data as previously noted).
- Adjustment for potential confounding factors is accomplished through stratification. More thorough continuous risk or covariate adjustment would be preferable.

6 Other methods

In our Screening Evaluation, we identified several other general methodological areas related to signal detection or to large-scale multiple testing problems that may be of potential relevance to the Sentinel Initiative. However, we did not conduct a Full Evaluation of these methods since their potential for application in the Sentinel Initiative was less clear than for the methods detailed in Sections 2-5 of this report. However, we list these method areas here as a guidepost for future potential investigations:

- Methods to control false positive discovery in genome-wide and microarray studies
- Syndromic and bioterrorism surveillance methods for early aberration and disease outbreak detection
- Commercial data mining, data aggregation, and data abstraction methods/software (see Appendix C)

7 Recommendations

The following recommendations are based on our comprehensive evaluation of published signal detection methods and unpublished methods currently in development by leading statistical and medical product safety experts.

7.1 Hypothesis identification

At the time of product approval, comprehensively review risk-benefit information from existing data sources (e.g. pre-licensure industry studies, post-licensure data outside the U.S., medical officer product reviews and disagreements) to aid in planning post-licensure studies. Use this information to determine whether or not active observational surveillance of a large cohort is the preferred surveillance method for a given outcome-exposure pair compared to other potential approaches such as a more traditional Phase IV epidemiological cohort study or a randomized clinical trial.

Rationale: There is a considerable amount of valuable pre-licensure safety information available at the time of product approval collected at all phases of product development (including animal toxicity, human studies of the product and of similar products), and the Sentinel Initiative offers an opportunity to make increased use of such data to inform post-licensure surveillance. Industry has an asymmetric interest in efficacy and safety, which has led to fewer resources being spent previously to identify, verify, and publish safety problems or to develop statistical methods for analyzing pre-licensure safety data. However, high quality synthesis and conflict-free independent review of these data can be used to identify post-licensure hypotheses worth pursuing, inform the prioritization of outcomes and products for post-licensure surveillance, and to more accurately estimate expected event rates to inform post-licensure study power and sample size calculations.

Implications for signal detection method: This recommendation impacts the methods used to analyze pre-licensure safety data in order to improve the planning and design of post-licensure surveillance studies. It involves better assembly, analysis, and synthesis of pre-licensure information to improve the processes by which outcomes and products are selected for surveillance and to improve the quality of the outcome incidence rate estimates used for post-licensure study planning (e.g., sample size calculations).

Gaps in existing methods and future work needed:

a) Retrospectively, for previously-approved products, study pre-licensure risk-benefit information to determine what factors are markers of products that are later shown to have problems that require major warnings or withdrawals.
b) Prospectively, once informative markers are identified, conduct analyses of pre-licensure data on newly licensed products to obtain and use those markers, along with estimated outcome rates from pre-licensure data, to aid in planning post-licensure studies.
c) Explore Bayesian and empirical Bayes approaches for incorporating prior pre-licensure adverse event rate data (when such data are considered accurate and applicable to the post-licensure population) into methods used for post-licensure studies. For example, existing sequential testing methods could be modified to explicitly make use of outcome rate estimates from prior/pre-licensure studies and then more accurately and efficiently detect signals.

A comprehensive review of pre-licensure risk-benefit information at the time of medical product approval should also be used to determine whether or not active observational surveillance of a large cohort is the preferred surveillance method for a given outcome-exposure pair compared to other potential approaches such as a more traditional Phase IV epidemiological cohort study or a large Phase IV randomized clinical trial. Active surveillance using large, linked databases is best suited for monitoring rare events and for ruling out relatively large relative risk effects (e.g. RR>10 or possibly RR>5). If, based on known risk-benefit evidence from pre-licensure studies, it is clinically important to rule out a more modest relative risk (e.g. RR≤10), then a Phase IV randomized clinical trial with non-surrogate endpoints should be considered since unmeasured confounding in database studies for many outcome-exposure pairs can easily mask increased safety risks.

7.2 Primary signal detection methods

7.2.1 Active surveillance

If the pre-licensure risk-benefit evidence supports the use of an active surveillance framework, then, as a primary signal detection approach in this setting, we recommend targeting specific medical products (exposures) and outcomes for routine post-licensure surveillance and testing these target hypotheses by conducting sequential analyses with automated health care data, provided that high quality health care data are available.

Rationale: Targeted longitudinal surveillance will be more powerful than all-by-all data mining of medical products and adverse events due to the magnitude of the multiple testing problem (i.e., multiple products, outcomes, subgroups, testing time points). Targeted surveillance will also allow focus on the most important (from a risk-benefit standpoint) and the most reliably-measured products and outcomes.

Implications for signal detection method: Sequential testing methods offer a useful and proven framework for conducting pre-specified hypotheses at multiple points in time throughout a study as data accrue while controlling the overall false positive statistical error rate across all tests performed. The maximized sequential probability ratio test (maxSPRT) has been successfully used on a weekly basis for prospective vaccine safety monitoring within the Vaccine Safety Datalink (VSD). Group sequential methods have been widely used to sequentially monitor medical product efficacy and safety in clinical trials on a less frequent basis (e.g., quarterly or semi-annually throughout the trial). Similarly, sequential testing could be applied in the Sentinel Initiative to efficiently monitor safety. In addition, routine testing results should be accompanied by routine descriptive analyses of the safety data to fully characterize outcome rates by site, demographics, dose, time interval after the product is administered and other factors.

Gaps in existing methods and future work: The use of and statistical performance of sequential testing methods in observational safety surveillance settings, in which relatively frequent monitoring plans (compared to clinical trials, for e.g.) may be desired and in which confounding is a potential concern, has not been considered. Several methodological developments are needed to evaluate and adapt existing sequential methods for potential use in the Sentinel Initiative:

a) Develop a formal statistical framework to optimally design a sequentially monitored observational safety surveillance study for pre-specified outcomes by extending group sequential clinical trial methods. Such a framework will provide statistical information to guide design decisions such as: How frequently should testing occur? What boundary shape is best? What test statistic to compare risk should be used? Next steps:
    - Conduct simulation studies to compare the statistical operating characteristics (e.g. power, time-to-signal detection) of the standard continuous sequential monitoring method (maxSPRT, used in the VSD project) to a variety of group sequential designs (which vary the monitoring frequency, boundary shape, and test statistic to compare risks)
    - Apply a variety of sequential designs using existing data sources (e.g., past and ongoing VSD safety studies) and compare their performance to maxSPRT.
b) In settings with concurrent controls, improve the way confounders are accounted for in sequential testing by developing a 'covariate-adjusted analysis' approach to construct monitoring boundaries. Most existing methods, including maxSPRT, involve matching (which yields information loss if matches cannot be found) and stratification (which may not be fine enough for adequate confounder control).
c) Surveillance using selected sequential testing methods has been conducted to date in relatively simple outcome-exposure settings (e.g., for vaccines paired with outcomes

that occur in relatively narrowly-defined post-vaccination intervals). The use and implementation of sequential methods for monitoring more complex exposures (e.g., drugs, which may be cumulatively-defined or time-varying) and outcomes (e.g., those with delayed onset) requires further evaluation.

7.2.2 Conventional surveillance

If high quality outcome and exposure data are *not* available from automated health care data sources or if confounding is too strong by factors not reliably measured in such data sources, conduct a conventional Phase IV observational safety study of pre-specified hypotheses using more reliable data sources (e.g., information abstracted from medical charts or data from existing prospectively collected large cohort studies) or conduct a large simple Phase IV randomized clinical trial. Choice of data source and study design may also depend on expected effect sizes and sample size.

Rationale: Given the extensive resources that are involved in confirming or refuting signals detected using information from automated health care data sources, if such signals are falsely generated from data sources that are inaccurate or unreliable, much effort will be wasted. In such circumstances, it will be better to simply invest at the outset in a rigorously-designed Phase IV confirmatory study.

Implications for signal detection method: Methods to test pre-specified safety-related hypotheses at a single point in time (e.g., at the end of an observational Phase IV safety study) are well developed (e.g. logistic regression).

Gaps in existing methods and future work needed: A systematic framework is needed to determine whether the quality of outcomes and exposures defined using automated health care data sources is high enough to support their use in sequential surveillance Data must be of high enough quality to guarantee that the resulting potential bias on surveillance results due to data inaccuracies is reasonably small. Important next steps:

a) Determine what data quality factors are essential to consider when selecting outcomes and products for surveillance (e.g., database outcome and exposure accuracy, complexity of database exposure definitions, variability in coding practices across health plans, level of confounding in the population of interest by variables not available in automated sources).
b) Develop a systematic approach to use existing studies and/or to conduct pilot studies using automated healthcare data sources to obtain information on these data quality parameters for outcomes and products at the time of product approval (i.e., in advance of post-licensure study) to aid in post-licensure study planning. For example:
    - Review existing literature on automated health care database quality for candidate outcomes and exposures and on confounders for candidate outcome-exposure pairs.

- Conduct pilot data quality studies at a local site or with a small group of sites to estimate candidate outcome and exposure accuracy and to identify and estimate the influence of confounders on candidate outcome-exposure pairs.

c) Conduct sensitivity analyses to gauge the impact (i.e. bias) of data quality deficiencies of varying type and magnitude (as measured by data quality parameters) on surveillance results.

7.3 Primary signal confirmation methods

As a primary approach to confirm signals generated by active post-licensure surveillance of automated healthcare data sources, use a two-phase sampling study design[57] to select a subset of participants on which to collect additional information (e.g. medical chart data).

Rationale: Signals detected using automated and often aggregated (i.e. event counts by age, exposure, and site subgroups) health care data sources will likely require follow-up, including the collection of additional individual-level information (e.g., from medical charts). Because both false and true signals can occur (and both may require follow-up) and because the resources required to obtain additional information (e.g., to abstract data from charts) are large, minimizing the amount of additional information needed to accurately confirm or refute a signal is essential.

Implications for signal detection method: 2-phase sampling methods are potentially useful tools for efficient collection of data on a targeted and informative subset.

Gaps in existing methods and future work needed: 2-phase sampling methods, and other techniques, warrant further development and evaluation in the medical product safety surveillance setting:

a) Develop an optimal 2-phase sampling design strategy to determine which (based on outcome, exposure, and/or confounder status) and how many medical charts (i.e., Phase 2 data) minimally need to be reviewed to reliably confirm or refute a signal detected using automated (i.e., Phase 1) data.
b) Compare the performance of candidate 2-phase analysis strategies (e.g. generalized raking[58] and other 2-phase sampling strategies from the survey literature)
c) Develop a framework to determine what Phase 2 confounders should be collected.
   - Review literature on the nature and magnitude of confounding that exists.
   - Propose further research, if needed, to more fully understand the potential impact of confounding.
d) Develop a framework to determine how to best incorporate Phase 2 confounders into analytic regression models:
   - Describe approaches to minimize confounding (e.g., stratify, restrict, adjust).
   - Assess the ability of different data sources to measure comorbidity (ICD-9 codes, pharmacy data, home health care data, problem lists, telephone encounter data, medical charts, etc.)

- Assess different approaches to adjust for comorbidity (e.g. summary scores versus individual variables; more nuanced measures such as severity; use of cognitive and functional status)

Note that methods proposed in c and d above to determine what confounders (beyond those available in automated health care data sources) should be collected and how to best incorporate them into analytic regression models are similarly needed to properly design and analyze a conventional Phase IV study.

7.4 Secondary and supplementary methods

In addition to the primary signal detection and confirmation methods recommended above for testing pre-specified hypotheses, continue to collect and analyze other supplementary post-licensure safety data (e.g., spontaneous reports), and investigate further development of complimentary signal detection approaches (e.g., data mining for unanticipated adverse events in prospective cohorts) as secondary methods to ensure comprehensive surveillance. Specifically:

a) Other data sources
   i. Spontaneous reporting databases should continue to be analyzed.
   ii. NIH prospectively collected cohort studies such as the Multi-Ethnic Study of Atherosclerosis[59] and the Cardiovascular Health Study[60] could also potentially serve as datasets where safety questions might be addressed quickly.

b) Other signal detection methods
   i. To detect unanticipated adverse outcomes, thus complimenting sequential testing for pre-specified adverse events, further develop cross-sectional data mining approaches and possibly adapt data mining methods for use in prospective longitudinal surveillance. In particular, data mining methods that explicitly (versus in *ad hoc* way) control for the overall false positive error rate and that search for optimal outcome groupings of ICD-9 codes (rather than require a pre-specified list of outcomes) are most promising.
   ii. Due to likely heterogeneity in data quality and coding practices across sites, develop meta-analysis methods to combine safety results across sites. In general, conducting meta-analyses of multiple post-licensure confirmatory studies to rule out regression to the mean effects (i.e. to avoid over-estimation of safety risks that may occur by chance in some samples) is also important.
   iii. Further explore existing literature on statistical process control and aberration and outbreak detection methods for potential use in identifying elevated safety risks due to bad drug or vaccine lots.

7.5 Methods not recommended at this time

Methods we reviewed and would not recommend pursuing further at this time include SPRT, CUSUM, PRR data mining, and Bayesian data updating.

Acknowledgements

We would like to thank the FDA Sentinel Initiative Safety Signal Identification contract team including Melissa Robb, Robert Ball, David Graham, Thomas Gross, Abraham Hartzema, Karen Lee, and Judy Staffa for their ongoing feedback on our methods investigations which critically influenced the evolution and content of this report.

We would also like to thank Lisa Jackson (Group Health, University of Washington), Bruce Psaty (University of Washington, Group Health), Martin Kulldorff (Harvard Medical School, Harvard Pilgrim Health Care), and Tom Fleming (University of Washington) for providing technical insight on several statistical, epidemiological, and clinical issues that importantly informed this report. We would also like to thank the following external experts for sharing their experiences developing and/or implementing methods for use in safety surveillance:

Jeffrey Brown -- Harvard Medical School, Harvard Pilgrim Health Care, Boston MA
Katherine Yih -- Harvard Medical School, Harvard Pilgrim Health Care, Boston MA
Eric Weintraub -- Centers for Disease Control and Prevention, Atlanta GA
James Baggs -- Centers for Disease Control and Prevention, Atlanta GA
Matthew Daley -- Kaiser Permanente Institute for Health Research, Denver CO
James Donahue -- Marshfield Clinic Research Foundation, Marshfield WI
Nicola Klein -- Northern California Kaiser Permanente, Oakland CA
Allison Naleway -- Kaiser Permanente Northwest, Portland OR
Steven Jacobsen – Southern California Kaiser Permanente, Pasadena CA
Frederick Resnic--Brigham and Women's Hospital, Harvard Medical School, Boston MA
Willam DuMouchel -- Lincoln Technologies Division, Phase Forward, Waltham MA

References

1. DuMouchel W, Pregibon D. Empirical bayes screening for multi-item associations. Proceedings of the conference on knowledge discovery and data. 2001 Aug 26-29; San Diego, CA. ACM Press: 67-76.
2. DuMouchel W. Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system. The American Statistician. 1999;53:177-90.
3. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's Spontaneous Reports Database. Drug Safety 2002;25(6):281-392.
4. Levine JG, Szarfman A. Standardized data structures and visualization tools: a way to accelerate the regulatory review of the integrated summary of safety of new drug applications. Biopharmaceutical Report. 1996;4(3)12-7.
5. Wald A. Sequential tests of statistical hypotheses. Annals of Mathematical Statistics. 1945;16:117-186.
6. Kulldorff M, Davis RL, Kolczak M, et al. A maximized sequential probability ratio test for drug and vaccine safety surveillance. Department of Ambulatory

Care and Prevention. Working Paper. 2007; available at www.dacp.org/faculty_kulldorff.html.
7. Emerson SS, Kittelson JM, Gillen DL. Tutorial in biostatistics: Frequentist evaluation of group sequential clinical trial designs. Stat in Med. 2007:26:5047-5080.
8. Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential clinical trial designs. Stat in Med. 2007:26:1431-1449.
9. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. Epidemiology 2005;16:336–341.
10. Lieu TA, Kulldorff M, Davis RL, et al. Real-time vaccine safety surveillance for the early detection of adverse events. Med Car 2007 Oct;45(10 Supl 2):S89-95.
11. Armitage P. Sequential tests in prophylactic and therapeutic trials. Q J Med. 1954; 23:255-74.
12. Bartholomay AF. The sequential probability ratio test applied to the design of clinical experiments. N Engl J Med 1957; 256:498-505.
13. Spiegelhalter D, Grigg O, Kinsman R, et al. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. International Journal for Quality in Health Care. 2003; 15(1): 7-13.
14. Leandro G, Rolando N, Gallus G, et al. Monitoring surgical and medical outcomes: the Bernoulli cumulative SUM chart: a novel application to assess clinical interventions. Postgraduate Medical Journal. 2005; 81: 647-652.
15. Reynolds MR, Stoumbos ZG. Monitoring a proportion using CUSUM or SPRT control charts. Frontiers in Statistical Quality Control 2001; 6:155-176
16. Page ES. Continuous inspection schemes. Biometrika. 1954; 41:100-114.
17. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. J R Statist Soc A 2003; 166(1):5-21.
18. Montgomery DC. Introduction to statistical quality control, 2nd ed. 1991. New York: John Wiley and Sons.
19. Moustakides GV. Optimal stopping times for detecting changes in distributions. Annals of Statistics 1986; 14:1379-1387.
20. Nix A, Rowlands RJ, Kemp KW. Internal quality control in clinical chemistry. Statistics in Medicine 1986; 6:425-440.
21. Gallus G, Mandelli C, Marchi M, Radaelli G. On surveillance methods for congenital malformations. Statistics in Medicine 1986; 5:565-571.
22. Steiner SH, Cook RJ. Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics. 2000;1(4):441-452.
23. DeLeval MR, Francois K, Bull C, Brawn WB, Spiegelhalter DJ. Analysis of a cluster of surgical failures. Journal of Thoracic and Cardiovascular Surgery 1994; 104:914-924.
24. Steiner SH, Cook RJ, Farewell VT. Risk-adjusted monitoring of binary surgical outcomes. Med Decis Making 2001; 21:163-169.
25. Steiner SH, Cook RJ, Farewell VT. Monitoring paired binary surgical outcomes using cumulative sum charts. Statistics in Medicine. 1999; 18:69-86.
26. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Statistical Methods in Medical Research. 2003; 12: 147-170.

27. Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiology and Drug Safety. 2007; 16: 1275-1284.
28. 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. Pharmacoepidemiology and Drug Safety. 2009; 18(3):226-34.
29. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64(2): 191-199.
30. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979; 35(3): 549-556.
31. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics. 1987; 43(1): 193-199.
32. Armitage P. Sequential methods in clinical trials. Am J Pub Health. 1958; 48:1395-1402.
33. Jennison C and Turnbull BW. Group sequential methods with applications to clinical trials. CRC Press: Boca Raton FL, 2000.
34. Whitehead J. The design and analysis of sequential clinical trials. Wiley: New York, 1997.
35. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety 2001; 10:483-486.
36. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54:315-321.
37. Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. Drug Safety 2002; 25(6):393-397.
38. Bate A. Bayesian confidence propagation neural network. Drug Safety 2007; 30:623-625.
39. Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundamental & Clinical Pharmacology 2008; 22:127-140.
40. Bate A, Edwards IR, Edwards J, Swahn E, Noren N, Lindquist M. Knowledge finding in IMS disease analyzer Mediplus UK database – effective data mining in longitudinal patient safety data. Drug Safety 2004; 27:917-918.
41. Gibbons RD, Segawa E, Karabatsos G, et al. Mixed-effects Poisson regression analysis of adverse event reports: the relationship between antidepressants and suicide. Statistics in Medicine. 2008; 27: 1814-1833.
42. Niu MT, Erwin DE, Braun MM. Data mining in the US Vaccine Adverse Event Reporting System (VAERS): early detection of intussusception and other events after rotavirus vaccination. Vaccine 2001; 19:4627-4634.
43. Rolka H, Bracy D, Russell C, et al. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. Statistics in Medicine. 2005; 24: 551-562.
44. Smith MY, Irish W, Wang J, et al. Detecting signals of opioid analgesic abuse: application of a spatial mixed effect Poisson regression model using data from a

network of poison control centers. Pharmacoepidemiology and Drug Safety. 2008 [epub ahead of print].
45. Curtis JR, Cheng H, Delzell E, et al. Adaptation of Bayesian data mining algorithms to longitudinal claims data. Medical Care. 2008; 46(9): 969-975.
46. Woo EJ, Ball R, Burwen DR, Braun MM. Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System (VAERS). Drug Safety 2008; 31(8):667-674.
47. Matsushita Y, Kuroda Y, Niwa S, et al. Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer. Drug Safety. 2007; 30(8): 715-726.
48. Hauben M, Reich L. Safety related drug-labeling changes: findings from two data mining algorithms. Drug Safety. 2004; 27(10): 735-744.
49. Hauben M, Reich L, Chung S. Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. European Journal of Clinical Pharmacology. 2004; 60: 747-750.
50. Almenoff JS, LaCroix KK, Yuen NA, et al. Comparative performance of two quantitative safety signally methods: implications for use in a pharmacovigilance department. Drug Safety. 2006; 29(10): 875-887.
51. Banks D, Woo EJ, Burwen DR, Perucci P, Braun MM, Ball R. Comparing data mining methods on the VAERS database. Pharmacoepi and Drug Safety 2005; 14;601-609.
52. Resnic FS, Zou KH, Do DV, Apostolakis G, Ohno-Machado L. Exploration of a Bayesian updating methodology to monitor the safety of interventional cardiovascular procedures. Medical Decision Making 2004; 24:399-407.
53. Matheny ME, Ohno-Machado L, Resnic FS. Monitoring device safety in interventional cardiology. J Am Med Inform Assoc 2006;13:180-187.
54. Siu N, Apostolakis G. Modeling the detection rates of fires in nuclear plants: development and application of a methodology for treating imprecise evidence. Risk Anal. 1986; 6:43-59.
55. Resnic FS. Personal communication. March 4, 2009.
56. Matheny ME, Morrow DA, Ohno-Machado L, Cannon CP, Sabatine MS, Resnic FS. Validation of an Automated Safety Surveillance System with Prospective, Randomized Trial Data. Med Decis Making 2008 Dec 2. [Epub ahead of print]
57. Breslow NE, Chatterjee. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. Applied Statistics 1999;48:457-468.
58. Zhang LC. Generalized raking procedures in survey sampling. The American Statistician 2000;54:178-184.
59. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacobs DR, Kronmal R, Liu K, Clark Nelson J, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. American Journal of Epidemiology 2002; 156: 871-881.
60. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1991; 1:263-76.