

A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance

Martin Kulldorff^{1,*} Robert L. Davis²
Margarette Kolczak^{3,†} Edwin Lewis⁴ Tracy Lieu¹
Richard Platt¹

¹ Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston, MA 02215, USA.

² Department of Research, Kaiser Permanente Georgia, Atlanta, GA 30305, USA.

³ Immunization Safety Branch, National Immunization Program, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA.

⁴ Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612, USA.

* martin_kulldorff@hms.harvard.edu

† deceased

Abstract

Because of rare but serious adverse events, pharmaceutical drugs and vaccines are sometimes withdrawn from the market, either by a government agency such as the Food and Drug Administration (FDA) in the United States or by the manufacturing pharmaceutical company. In other cases, a drug may be generally safe but increase the risk for serious adverse events for certain sub populations such as pregnant women or people with heart problems. Due to limited sample size and selected study populations, rare adverse events are often impossible to detect during phase 3 trials conducted before the drug is approved for general use. It is then important to conduct post-approval drug safety surveillance, using for example health insurance claims data. In such surveillance, the goal should be to detect serious adverse events as early as possible without too many false alarms, and it is then natural to use a sequential test procedure. In this paper, we first show that

Wald's classical sequential probability ratio test (SPRT) is very sensitive to the choice of relative risk required in the specification of the alternative hypothesis, making it difficult to use for drug and vaccine safety surveillance. We then propose a maximized sequential probability ratio test (MaxSPRT) based on a composite alternative hypothesis, which works well across a range of relative risks. We illustrate the use of this method on vaccine safety surveillance and compare it with the classical SPRT. A table of critical values for the MaxSPRT is provided, as well as estimates of statistical power, the time until the null hypothesis is rejected and the average length of surveillance.

Keyword: Drug safety; Rapid cycle analysis; Surveillance; Sequential analysis.

1 Introduction

The early detection of unexpected adverse events is very important in both drug and vaccine safety surveillance. While common adverse events are often detected during phase 2 and 3 clinical trials, rare but serious adverse events may go undetected due to limited sample size. Other adverse events may go undetected if they only affect a subpopulation that was excluded from the clinical trial. To catch these types of adverse events it is important to conduct post-marketing drug and vaccine safety surveillance [1, 2, 3, 4]. This can be done by monitoring adverse events among patients receiving these drugs/vaccines as part of their regular medical care, using for example health insurance claims data. Even when no adverse events are found, it is important to do this type of surveillance to ensure the public that new drugs and vaccines are not only effective but also safe, so that patients do not avoid taking important and life saving drugs/vaccines due to safety concerns.

In order to detect a problem with an adverse event as early as possible, the ideal is to do weekly or monthly monitoring of patients as they receive the drug or vaccine under study, generating an adverse event signal if and when the number of adverse events are so great that they are unlikely to be due to chance alone. For such sequential analyzes, Wald proposed a sequential probability ratio test (SPRT) [5, 6], where a signal is generated if the likelihood ratio exceeds a certain pre-determined value, while the observation ends if the likelihood falls below another predetermined lower bound. The key aspect of this method is that the p-values are adjusted for the continual

looking at the data (i.e. multiple testing). Sequential probability ratio tests have been extended and refined in various ways, for the comparison of two binomial proportions [7, 8, 9], for normally distributed data with composite alternatives [10, 11], for variance components [12], for stepwise sequential probability ratio tests [13], for multihypothesis testing [14], for sequential tests with cost functions [15, 16], and so on. Both the theoretical and practical aspects of the field has been summarized in excellent books by Gosh [17], Armitage [18], Siegmund [19], Whitehead [20], Ghosh, Mukhopadhyay and Sen [21], Jennison and Turnbull [22], Mukhopadhyay [23] and Govindarajulu [24], among others.

One problem with Wald’s classical sequential probability ratio test is that the result is highly dependent on the relative risk used to specify the alternative hypothesis [25]. In this paper we illustrate this problem in the context of vaccine safety surveillance, showing that an unfortunate choice of the relative risk for the alternative hypothesis may either delay the detection of an important signal or completely miss it. We propose instead the use of a maximized sequential probability ratio test (MaxSPRT), where the alternative hypothesis is composite rather than singular, with the relative risk defined as being greater than one rather than a specific value. For different surveillance schemes, we provide tables with critical values for both Poisson and binomial type data. The method is illustrated using historical data on fever and neurological symptoms after PediarixTM vaccination.

2 Vaccine Data

To illustrate the use of both the classical SPRT and the new MaxSPRT for vaccine safety surveillance, we have applied them using a historical time series of health insurance claims data from the Centers for Disease Control and Prevention (CDC) Vaccine Safety Datalink (VSD) project. With this data, we will mimic a prospective weekly surveillance system for evaluating whether there is any increased risk of either fever or neurological symptoms within 28 days after Pediarix vaccination. Manufactured by GlaxoSmithKline, PediarixTM is a combination vaccine that with a single injection protects children from five different diseases: diphtheria, tetanus, whooping cough, hepatitis B, and polio. The VSD project and the data it uses have been described in detail elsewhere [1, 26]. Here we only give a brief overview.

Started in 1991, the Vaccine Safety Datalink is a collaborative project

between CDC and eight different health plans: Group Health Cooperative of Puget Sound (Seattle, WA); Harvard Pilgrim Health Care / Harvard Vanguard Medical Associates (Boston, MA); Health Partners (Minneapolis, MN), Kaiser Permanente Colorado (Denver, CO), Marshfield Clinic (Marshfield, WI), Northern California Kaiser Permanente (Oakland, CA), Northwest Kaiser Permanente (Portland, OR), and Southern California Kaiser Permanente (Torrance, CA). Together, these plans cover approximately 650,000 children in the United States under the age of six, 3.5 percent of the total United States population in that age group. As part of the project, immunizations of these children are automatically tracked. Moreover, information about disease diagnoses made during routine medical care at hospitals, emergency departments and outpatient clinics are available. The date is recorded for all events, so that one can tally the number of adverse event seen within a risk window of fixed number of days after vaccination.

3 Wald's Sequential Probability Ratio Test

3.1 Mathematical Definition

Sequential analysis was first developed by Wald in the 1940's [5, 6, 27], at which time he introduced the sequential probability ratio test (SPRT). The likelihood based SPRT proposed by Wald is very general in that it can be used for many different probability distributions. In our setting, it is defined as follows.

Let C_t be the random variable representing the number of adverse events within D days following a vaccination (or drug prescription) that was given during the time period $[0, t]$, and let c_t be the corresponding observed number of adverse events. Note that time is defined in terms of the time of the vaccination rather than the time of the adverse event, and that hence, we actually do not know the value of c_t until time $t + D$.

Under the null hypothesis (H_0), C_t follows a Poisson distribution with mean μ_t , where μ_t is a known function reflecting the population at risk. In our setting, μ_t reflects the number of people who received the drug/vaccine during the time interval $[0, t]$ and a baseline risk for those individuals, adjusting for age and gender. Under the alternative hypothesis (H_A), the mean is instead $RR\mu_t$, where RR is the increased relative risk due to the drug/vaccine. Note that $C_0 = c_0 = \mu_0 = 0$.

With the classical SPRT, tests are performed continuously at every time point $t > 0$ as additional data are collected. The test statistic is the likelihood ratio, which for the Poisson distribution is defined as

$$LR_t = \frac{P(C_t = c_t | H_A)}{P(C_t = c_t | H_0)} = \frac{e^{-RR\mu_t} (RR\mu_t)^{c_t} / c_t!}{e^{-\mu_t} \mu_t^{c_t} / c_t!} = e^{(1-RR)\mu_t} (RR)^{c_t}$$

or equivalently, as the test statistic is often defined using the log likelihood ratio

$$LLR_t = \ln(LR_t) = (1 - RR)\mu_t + c_t \ln(RR)$$

This test statistic is sequentially monitored for all values of $t > 0$, until either $LLR_t \geq \ln((1 - \beta)/\alpha)$, in which case the null hypothesis is rejected, or until $LLR_t \leq \ln(\beta/(1 - \alpha))$, in which case the alternative hypothesis is rejected. With this stopping rule, the null hypothesis will be falsely rejected with probability α when it is true (type 1 error) while the alternative hypothesis will be falsely rejected with probability β when it is true (type 2 error), although it should be noted that these are approximate results [5, 6]. Note that $LLR_0 = 0$.

As an example, for $\alpha = 0.05$ and $\beta = 0.20$, the upper and lower rejection levels are 2.77 and -1.56 respectively. We will use these two values for the SPRT throughout the paper. The SPRT is designed for continuous monitoring, but in practice, it is often evaluated at frequent but discrete time intervals, resulting in a slightly conservative test procedure. In this paper we use it for weekly data.

3.2 Pediarix Vaccination Safety Surveillance

The first question we will ask using the historical vaccine data is if there is an increased risk of fever during the four weeks following Pediarix vaccination. The top left of Figure 1 shows the result of the classical SPRT. With an alternative hypothesis of $H_A : RR = 2.0$, there is enough evidence to reject the alternative hypothesis after 7 weeks, with the conclusion that there is no evidence that Pediarix increases the risk of fever. With $H_A : RR = 1.2$, we get the opposite result, with a rejection of the null hypothesis after 13 weeks, with the conclusion that Pediarix increases the risk of fever.

[Figure 1 about here.]

Why do we get these seemingly contradictory results? Suppose that the true $RR = 1.2$. If the alternative hypothesis is $H_A : RR = 2$, then there is more evidence for the null hypothesis than for the alternative hypothesis, and hence, the alternative hypothesis will be rejected. If the alternative is $H_A : RR = 1.2$, then there is more evidence for the alternative hypothesis than for the null hypothesis, and the null hypothesis will be rejected. Hence, which hypothesis is rejected depends on the alternative hypothesis chosen. This makes perfect mathematical sense, but the practical implications are disturbing, as we do not know beforehand what excess relative risk we should look for.

One option would be to take a ‘conservative’ approach by always choosing a very low relative risk for the alternative hypothesis, so that any true relative risk below that threshold is clinically unimportant and uninteresting to detect. That can also lead to problems though. In the top right part of Figure 1, the results are shown when using the classical SPRT to evaluate an increased risk of neurological symptoms during the four weeks following Pediarix vaccination. With $H_A : RR = 1.2$, there is some evidence of an excess risk, and after 65 weeks there is enough evidence to reject the null hypothesis. If instead, we use an alternative model with $H_A : RR = 2.0$, then the null hypothesis is rejected after 32 weeks.

What is going on now? Suppose the true $RR = 2$. If the alternative model is $H_A : RR = 1.2$, then it is almost as bad as the null model with $RR = 1$, so the log likelihoods are similar and the log likelihood ratio stays close to zero, and it will take a long time until we reach the upper boundary to reject the null hypothesis. If the alternative model is instead $H_A : RR = 2$, then there is much more evidence for the alternative than for the null hypothesis, resulting in a larger log likelihood ratio, and the null will be rejected much sooner. Again, this makes perfect mathematical sense while the practical consequences are worrisome, since the time until we detected a serious risk would be longer when using a low conservative relative risk for the alternative hypothesis than if we had used a higher relative risk. Note that, if we are only concerned about power, but not in minimizing the time to signal, then this is not a problem and we can safely use the classical SPRT with a single alternative chosen as the lowest relative risk of interest [6] (p73).

Another way to look at this latter problem is in terms of statistical power and sample size. If we want to detect a true relative risk of 1.2 with 80% power ($\beta = 0.20$), then we need a larger sample size than if we want to detect a true relative risk of 2.0 with the same power. Hence, with $H_A :$

$RR = 1.2$, we would expect to have to wait longer until the null is rejected. One option around this problem would be to modify the SPRT so that the likelihood calculations are based on the lowest relative risk of interest to detect (e.g. $H_A : RR = 1.2$) while the threshold is calculated to guarantee the desired power for a higher relative risk (e.g. 80% power for $H_A : RR = 2$). While it would be fairly easy to use computer simulations to calculate the correct critical values for any combination of power and pairs of relative risks, we think that a more natural approach is to use a maximized sequential probability ratio test (MaxSPRT) with a composite alternative hypothesis, as described below.

4 A Maximized Sequential Probability Ratio Test: Poisson Data

4.1 Log Likelihood Ratio

A better approach for drug and vaccine safety surveillance is, we think, to use a Maximized Sequential Probability Ratio Test (MaxSPRT), with a composite alternative hypothesis $H_A : RR > 1$. We then divide the maximum likelihood under the composite alternative hypothesis with the likelihood under the single null hypothesis [28]. The likelihood ratio based test statistic is then

$$LR_t = \max_{H_A} \frac{P(C_t = c_t | H_A)}{P(C_t = c_t | H_0)} = \max_{RR > 1} \frac{e^{-RR\mu_t} (RR\mu_t)^{c_t} / c_t!}{e^{-\mu_t} \mu_t^{c_t} / c_t!} = \max_{RR > 1} e^{(1-RR)\mu_t} (RR)^{c_t}$$

The maximum likelihood estimate of RR is c_t / μ_t , so

$$LR_t = e^{\mu_t - c_t} (c_t / \mu_t)^{c_t}$$

Equivalently, when defined using the log likelihood ratio

$$LLR_t = \ln(LR_t) = \max_{RR > 1} ((1 - RR)\mu_t + c_t \ln(RR)) = (\mu_t - c_t) + c_t \ln(c_t / \mu_t)$$

4.2 Critical Values

When defining the critical values for the test statistic there are a few different options. One is to use the standard approach and reject the null when the

LLR reaches an upper bound and accept the null when the LLR reaches a lower bound. An alternative approach is the reject the null when the LLR reaches an upper bound and accept the null when the surveillance has been going on for a predetermined length of time, defined in terms of the expected number of events accrued under the null hypothesis. There are pros and cons with either approach, and the choice will depend on the application. For drug and vaccine safety surveillance we prefer the latter. Since we are doing observational surveillance using data that is collected regardless, there is no harm in continuing the surveillance if the drug/vaccine is safe except for minor data analytic costs. It also makes the length of the study more predictable. With the first approach we do not know for how long the surveillance may go on.

[Table 1 about here.]

We have not found any analytical way of calculating the critical values for the MaxSPRT. This is not a problem though, as they are easy to calculate using computer based simulations for either of the two scenarios described above, or for any other type of rejection region. In Table 1 we present the upper bounds used for the rejection of the null hypothesis under the second scenario, for different upper limits on the length of surveillance. These critical values are based on 9,999,999 simulated data sets. Note that, as expected, the longer we are willing to do the surveillance, the larger the LLR must be before we reject the null hypothesis. This is because there is more multiple testing that needs to be adjusted for. Hence, the less willing we are to stop early and accept the null hypothesis, the longer it takes to reach the critical value needed to stop and reject the null hypothesis.

Note that the simulations only have to be done once, so users of the MaxSPRT do not need to do their own computer simulations. As long as they use one of the upper limits presented in Table 1, they can simple use that table.

It is important to note that the null hypothesis should be rejected as soon as the LLR reaches the critical value, even if it subsequently drops below again. Allowance for this is taken into account when the critical value is determined. In fact, due to the randomness of the data, it is rather typical that the LLR falls below the critical value soon after it reaches it for the first time, but then climbs above again and stays above.

[Table 2 about here.]

4.3 Statistical Power

In addition to ensuring the correct alpha level, it is important to consider the statistical power to reject the null for different true relative risks. Table 2 presents such estimates, based on 100,000 simulations for each relative risk defining the alternative hypothesis. The power is obviously higher when the true relative risk is larger, so the main interest is to compare the power for different upper bounds on the length of surveillance. As expected, the power for the maximized SPRT is higher when T , the expected number of events defining the maximum length of surveillance, is longer. This is natural since the sample size is allowed to grow larger. In fact, the power obtained is one possible criterion to use when selecting this upper limit. The trade-off is that we must be willing to collect data for a longer time period if the null is not quickly rejected. Hence, the choice of the upper limit on surveillance length is the classical trade-off between sample size and power, although we often do not have to utilize the full sample size that we allow for.

[Table 3 about here.]

4.4 Signal Timeliness and Length of Surveillance

In sequential analyzes is not only the α -level and power that is important, but also the time it takes to reject the null hypothesis when the alternative is true. Conditioned on the null being actually rejected, Table 3 (top) shows the expected time until rejection for different MaxSPRT parameter values. The bottom of the same tables shows instead the expected length of surveillance, until either the null hypothesis is rejected or accepted. In some applications, it is also important to consider the time until the alternative hypothesis is rejected when the null hypothesis is true, but in drug and vaccine safety surveillance that is not the major concern.

4.5 Aggregated Data

The MaxSPRT is, just as the classical SPRT, formulated for data that is continuously collected and evaluated. In drug and vaccine surveillance, it is often more practical to collect data on a slightly aggregate bases such as weekly or monthly counts. If the log likelihood ratio is only evaluated at the end of each week, the MaxSPRT will be slightly conservative in that

the probability of rejecting the null when it is true is somewhat less than the nominal alpha level. It may also result in a slight delay in detecting a true signal. A slightly modified approach, which will maintain the correct alpha level, is to randomly allocate the observations within the expected counts accrued during that week by using a uniform distribution. In most practical settings, either approach will work fine without major differences in the results. The MaxSPRT should not be used when the sequential testing is done in a less frequent manner though, such as once every year, as it would then adjust for more multiple testing than necessary. It is then more appropriate to use group sequential methods [22].

5 A Maximized Sequential Probability Ratio Test: Binomial Data

Reliable estimates for the expected number of events are not always available before the start of drug and vaccine safety surveillance. An alternative design is then to collect information about potential adverse events from both exposed and unexposed times. For example, for the same individual, we may compare an exposed time period after vaccination with an unexposed time period before vaccination; or with an unexposed time period long after vaccination. Alternatively, we may compare individuals exposed to the drug/vaccine with matched unexposed individuals. Unless the unexposed time period is much longer than the exposed, we cannot then use the Poisson distribution. We should instead use a binomial probability model when calculating the log likelihood function and the critical values. The upper limit on the length of surveillance will also be different, and should now be defined in terms of the number of adverse events seen. That is, we would continue the surveillance until either there is a signal rejecting the null hypothesis or when we have observed a total of N adverse events in the exposed and unexposed time periods combined. In essence, we have a number of coin tosses (adverse events) which may either turn up as head or tail (exposed or unexposed), and under the null hypothesis the probability of a head is known to be p ($p = 0.5$ for a 1:1 matching ratio when the exposed and unexposed time periods are of the same length, $p = 0.25$ for a 1:3 matching ratio, etc).

Other than these differences, the principles behind the MaxSPRT are the same for Poisson and binomial type data.

5.1 Log Likelihood Ratio

Let n be the number of adverse events seen so far during the sequential data collection, and among those n events, let $c_n \leq n$ be the number of adverse events among the exposed individuals. Let z be the number of matched unexposed individuals per exposed individual. Conditional on the number of adverse events n , we can then write the likelihood ratio for the binomial model as:

$$LR_n = \max_{H_A} \frac{P(C_n = c_n | H_A)}{P(C_n = c_n | H_0)} = \max_{RR > 1} \frac{(RR/(z + RR))^{c_n} (z/(z + RR))^{n-c_n}}{(1/(z + 1))^{c_n} (z/(z + 1))^{n-c_n}}$$

The maximum likelihood estimate of RR is $zc_n/(n - c_n)$. So

$$LR_n = \frac{(c_n/n)^{c_n} ((n - c_n)/n)^{n-c_n}}{(1/(z + 1))^{c_n} (z/(z + 1))^{n-c_n}}$$

when $zc_n/(n - c_n) > 1$ and $LR_n = 1$ otherwise. Equivalently, when defined using the log likelihood ratio

$$LLR_n = \ln(LR_n) = c_n \ln\left(\frac{c_n}{n}\right) + (n - c_n) \ln\left(\frac{n - c_n}{n}\right) - c_n \ln\left(\frac{1}{z + 1}\right) - (n - c_n) \ln\left(\frac{z}{z + 1}\right)$$

when $zc_n/(n - c_n) > 1$ and 0 otherwise.

[Table 4 about here.]

5.2 Critical Values

The critical values for the MaxSPRT for binomial data are provided in Table 4. Note that the critical values are often identical for different values of the upper limit on the survival length N . This is because of the discrete nature of the data. For example, with $N = 10$ there are only $2^{10} = 1024$ possible outcomes of the surveillance, since each of the ten adverse events will either be for an exposed or an unexposed individual. This discreteness also means that the actual alpha level is usually somewhat less than the nominal 0.05, but never more.

These critical values were calculated analytically, using an iterative Markov chain approach. Because of the discrete nature of the data, there is only a finite number of values that the likelihood can take. For each of these likelihood values l , a separate Markov chain is constructed. The state space of

the Markov chain is (n, c_n) , where $n > 0$ and $0 \leq c_n \leq n$. For the value n , the probability for each state can easily be computed iteratively from the probabilities for the value $n - 1$, with the initial condition that $P[(0, 0)] = 1$. Those states for which $LLR_n(c_n) \geq l$ are absorbing states, that cannot be departed. By summing the probabilities of these states we get the alpha level when the likelihood value l is used as the critical value.

6 Example: Pediarix Vaccine Safety Surveillance

We applied the Poisson based MaxSPRT to the same Pediarix data that we analyzed using the classical SPRT in Section 3 above. As the upper limit on the length of surveillance we choose 800 and 15 expected events for fever and neurological symptoms respectively, corresponding approximately to two years of surveillance. The results are shown in bottom of Figure 1.

For fever, the MaxSPRT rejects the null hypothesis after 13 weeks at the alpha 0.05 level, due to 97 observed cases when 69.7 were expected under the null, with $RR = 1.39$ and $LLR = 4.78$. For neurological symptoms, the MaxSPRT rejects the null hypothesis after 42 weeks at the $\alpha = 0.05$ level, due to 15 observed cases when 5.5 were expected under the null, with $RR = 2.7$ and $LLR = 5.51$.

[Table 5 about here.]

In Table 5, we compare the results when using the MaxSPRT with different upper limits on the length of surveillance, and the classical SPRT for different relative risks used for the alternative hypothesis. Results are provided for alpha levels of 0.05 and 0.01. The power for the classical SPRT depends on the true relative risk, but is set to be 0.80 for the alternative chosen. The power for the MaxSPRT depends the upper limit on the length of surveillance as well as on the true relative risk, as shown in Table 2. Note that, under normal circumstances one would do at most one of these analysis using prespecified parameter values, and we only present the multiple results for methodological comparisons.

7 Discussion

In this paper we have demonstrated an inherent problem when utilizing Wald’s classic SPRT for surveillance of vaccine and drug adverse events. We have then presented a maximized SPRT that uses a composite rather than a single alternative hypothesis. The new method was developed for two different probability models using the Poisson and binomial distributions respectively, but it can also be extended to other distributions such as the hypergeometric, suitable for other types of data. The maximized SPRT has been shown to work well for vaccine safety surveillance, with good statistical power and timeliness until signals are generated.

While the focus of this paper is purely methodological, the clinically relevant findings of our analysis deserve a brief comment. First, as in any disease surveillance setting, it is important to realize that a signal may either be due to a true excess risk or to other issues, including systematic differences in coding or diagnostic practices. A signal is hence a call for a detailed epidemiological study rather than proof of a clinical problem. Mild fever is a known side effect of Pediarix vaccination [29], so it is not surprising that we see a 16 percent elevated risk in our data. For neurological symptoms, we found that the excess number of cases is at least partly explained by changes made in the medical health records encounter forms affecting two different neurological symptoms.

7.1 Abt’s SPRT

Ours is not the first sequential probability ratio test with a composite alternative. Abt [25] provided an important first step in that direction, using a different approach. For some values of a and b specified by the user, with $0 < a < 1$ and $0 < b < 1$, define $R(a, b, t)$ as

$$R(a, b, t) = R : \frac{\ln(\frac{1-b}{a}) + \mu_t(R-1)}{\ln(R)} = \min_{RR>1} \frac{\ln(\frac{1-b}{a}) + \mu_t(RR-1)}{\ln(RR)}$$

This means that, for each time t , $R(a, b, t)$ is the value of the relative risk that minimizes the number of cases that is needed to reject the null hypothesis of the classical SPRT with $\alpha = a$ and $\beta = b$. The test statistic is then defined as

$$A_t = \frac{P(C_t = c_t | H_A : RR = R(a, b, t))}{P(C_t = c_t | H_0)} = e^{(1-R(a,b,t))\mu_t} (R(a, b, t))^{c_t}$$

The upper and lower rejection limits are set to be $\ln((1-b)/a)$ and $\ln(b)/(1-a)$ respectively, as with the classical SPRT. As Abt [25] points out, because of the minimization done when calculating $R(a, b, t)$, a and b no longer represent the approximate type 1 and 2 errors. Rather, for any pair (a, b) , the true type 1 and 2 errors are calculated using simulations. For example, with $a = 0.07$ and $b = 0.08$, the type one error is $\alpha = 0.1$ and the type 2 error is $\beta = 0.05$ [25].

The difference between Abt's SPRT and the MaxSPRT is that the former finds the relative risk that minimizes the number of cases needed to reject the null hypothesis with the classical SPRT, while the latter defines the test statistic by maximizing the likelihood over different relative risk parameter values. This latter approach is the standard way to deal with composite alternative hypotheses, through the creation of a likelihood ratio test statistic [30].

7.2 Rejection and Acceptance Region

In this paper we have defined the critical bounds so that the null is rejected when the LLR reaches a certain fixed value, and the null is accepted when the pre-specified upper limit on the length of surveillance is reached. This is a natural choice for drug and vaccine safety applications but not the only option. The MaxSPRT can be used with any other type of critical bounds as well, including the traditional upper and lower bounds used by the classical SPRT as well as various rejection regions of triangular or other shapes. To calculate critical values, statistical power and timeliness for such versions of the MaxSPRT, is an important area for further research.

7.3 Tables of Critical Values

While the critical values are based on extensive computations, a nice feature of the MaxSPRT is that the users do not have to do any of these computations themselves, but rather, can simply use the tables provided in this paper in the same old fashioned way that we used to do for most statistical distribution functions. The only exception is if the user wants to use different parameter values in terms of the alpha level, the upper length of surveillance or the number of unexposed individuals per exposed, although approximate critical values can in some cases be obtained from the existing tables through linear interpolation.

7.4 Weekly Vaccine Safety Surveillance

The examples provided in this paper used historical data to mimic a real-time surveillance system. The CDC sponsored Vaccine Safety Datalink project is currently using MaxSPRT for weekly surveillance of the safety of Menactra, Tdap, RotaTeq, MMRV, and HPV vaccines. Menactra is a tetravalent meningococcal conjugate vaccine, produced by Sanofi-Pasteur, which was approved in 2005, and the MaxSPRT is used to monitor for four different potential adverse events: Guillain-Barre syndrome, facial paralysis, thrombocytopenia and seizures. There have been no signals to date. This surveillance project is described in detail elsewhere [31]. The method is also currently being evaluated for drug surveillance using historical data from the HMO Research Network [32].

For the Poisson based MaxSPRT it is necessary to choose a comparison group to calculate the expected counts. Likewise, for the binomial based MaxSPRT, it is necessary to choose a set of matched unexposed individuals for each exposed person. The appropriate way of doing this depends on both the application and the available data, in order to minimize bias and confounding. In the two medical companion papers, we discuss how one may do this for vaccine and drug safety surveillance when using HMO data [31, 32], but for other applications it would be done very differently.

Acknowledgment

This work was supported in part by the Centers for Disease Control and Prevention through the Vaccine Safety Datalink Project and in part by grant HS10391 to the HMO Research Network Center for Education and Research on Therapeutics (CERTs), from the Agency for Health Care Research and Quality. We thank Ruihua Yin for data support, Elizabeth Pfoh for creating the figure and Paul Gargiullo for valuable comments on an earlier draft.

References

- [1] RL Davis, M Kolczak, E Lewis, J Nordin, M Goodman, DK Shay, R Platt, S Black, H Shinefield, and RT Chen. Active surveillance of vaccine safety: A system to detect early signs of adverse events. *Epidemiology*, 16:336–341, 2005.

- [2] A Szarfman, SG Machado, and RT O’Neill. 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*, 25:381–392, 2002.
- [3] W DuMouchel. Bayesian data mining in large scale frequency tables, with an application to the fda spontaneous reporting system. *American Statistician*, 53:177–202, 1999.
- [4] RT O’Neill and A Szarfman. Some us food and drug administration perspectives on data mining for pediatric safety assessment. *Current Therapeutic Reserach*, 62:650–663, 2001.
- [5] A Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186, 1945.
- [6] A Wald. *Sequential Analysis*. Wiley, 1947.
- [7] DN Joanes. Sequential tests of composite hypotheses. *Biometrika*, 59:633–637, 1972.
- [8] DG Hoel, GH Weiss, and R Simon. Sequential tests for composite hypotheses with two binomial populations. *JRSSB*, 38:302–308, 1976.
- [9] WQ Meeker. A conditional sequential test for the equality of two binomial proportions. *Applied Statistics*, 30:109–115, 1981.
- [10] JM Lachin. Sequential clinical trials for normal variates using interval composite hypotheses. *Biometrics*, 37:87–101, 1981.
- [11] I van der Tweel, R Kaaks, and PAH van Noord. Comparison of one-sample two-sided sequential t-tests for application in epidemiological studies. *Statistics in Medicine*, 15:2781–2795, 1996.
- [12] BK Ghosh. Sequential range tests for components of variance. *Journal of the American Statistical Association*, 60:826–836, 1965.
- [13] W Huang. Stepwise likelihood ratio statistics in sequential studies. *JRSSB*, 66:401–409, 2004.

- [14] CW Baum and VV Veeravalli. A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory*, 40:1994–2007, 1994.
- [15] S Holm. On the optimality of differentiated spr tests of composite hypotheses. *Metrika*, 32:15–33, 1985.
- [16] M Schipper, J den Hartog, and E Meelis. Sequential analysis of environmental monitoring data: Optimal sprts. *Environmetrics*, 8:29–41, 1997.
- [17] BK Ghosh. *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, 1970.
- [18] P Armitage. *Sequential Medical Trials, 2nd edition*. Blackwell, 1975.
- [19] D Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, 1985.
- [20] J Whitehead. *The Design and Analysis of Sequential Clinical Trials, 2nd edition*. Ellis Horwood, 1992.
- [21] M Ghosh, N Mukhopadhyay, and PK Sen. *Sequential Estimation*. Wiley, 1997.
- [22] C Jennison and BW Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, 1999.
- [23] N Mukhopadhyay. *Sequential Methodologies and Applications*. Chapman and Hall, 2002.
- [24] Z Govindarajulu. *Sequential Statistics*. World Scientific Publishing Company, 2004.
- [25] K Abt. Poisson sequential sampling modified towards maximal safety in adverse event monitoring. *Biometrical Journal*, 40:21–41, 1998.
- [26] RT Chen, JE Glasser, PH Rhodes, RL Davis, WE Barlow, RS Thompson, JP Mullooly, SB Black, HR Shinefield, CM Vadheim, SM Marcy, JI Ward, RP Wise, SG Wassilak, SC Hadler, and the Vaccine Safety Datalink Team. Vaccine safety datalink project: A new tool for improving vaccine safety monitoring in the united states. *Pediatrics*, 99:765–773, 1997.

- [27] A Wald and J Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339, 1948.
- [28] G Lorden. Open-ended tests for koopman-darmon families. *Annals of Statistics*, 1:633–643, 1973.
- [29] S Partridge and SH Yeh. Clinical evaluation of a dtap-hepb-ipv combined vaccine. *American Journal of Managed Care*, 9:S13–S22, 2003.
- [30] EL Lehmann. *Testing Statistical Hypotheses, 2nd edition*. Springer-Verlag, 1986.
- [31] TA Lieu, M Kulldorff, RL Davis, EM Lewis, KW Yih, JS Brown, and R Platt. Real-time vaccine safety surveillance for the early detection of adverse events. *Medical Care*, 45:S89–S95, 2007.
- [32] JS Brown, M Kulldorff, KA Chan, RL Davis, D Graham, PT Pettus, SE Andrade, M Raebel, L Herrinton, D Roblin, D Boudreau, D Smith, JH Gurwitz, MJ Gunter, and R Platt. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety*, 16:1275–1284, 2007.

Figure 1: Analyses of the safety of Pediarix vaccination with respect to fever (left) and Neurological symptoms (right) during the 28 days following vaccination, using the classical SPRT (top) with different relative risks defining the alternative hypothesis ($RR = 1.2$ and $RR = 2.0$) and the MaxSPRT (bottom) with a composite alternative ($RR > 1$).

The dashed lines are the critical value bounds. The solid lines are values of the log likelihood ratio test statistics. The final point estimates for the true relative risk were 1.16 for fever and 2.75 for neurological symptoms.

	0.05	0.01	0.001
T=1	2.86121	4.67676	7.17783
1.5	2.97072	4.78607	7.29289
2	3.05217	4.86978	7.34532
2.5	3.11777	4.93266	7.41141
3	3.16829	4.97900	7.46244
4	3.24997	5.04482	7.53088
5	3.30220	5.09726	7.57910
6	3.34850	5.14328	7.62238
8	3.41997	5.21307	7.69026
10	3.47399	5.26847	7.73711
12	3.51755	5.31141	7.77738
15	3.56791	5.35847	7.83106
20	3.63490	5.42270	7.90162
25	3.68379	5.47566	7.93953
30	3.72352	5.51255	7.97629
40	3.78259	5.57427	8.02826
50	3.82760	5.61884	8.06905
60	3.86431	5.65224	8.10018
80	3.91818	5.70639	8.15230
100	3.95952	5.74879	8.19729
120	3.99301	5.77973	8.23150
150	4.03131	5.81844	8.27277
200	4.07969	5.87074	8.31647
250	4.11724	5.90856	8.35589
300	4.14836	5.93869	8.37958
400	4.19457	5.98641	8.42991
500	4.22903	6.01998	8.46520
600	4.25648	6.04649	8.49589
800	4.29878	6.09461	8.53977
1000	4.33164	6.12942	8.58290

Table 1: Critical values for the MaxSPRT based log likelihood ratios for Poisson data. T is the upper limit on the length of surveillance, expressed in terms of the expected number of events under the null.

	True Relative Risk				
	1.2	1.5	2	3	5
T=1	0.068	0.107	0.184	0.378	0.729
1.5	0.071	0.118	0.218	0.472	0.851
2	0.075	0.129	0.253	0.558	0.924
2.5	0.077	0.139	0.287	0.634	0.963
3	0.079	0.149	0.320	0.701	0.982
4	0.084	0.171	0.387	0.807	0.996
5	0.088	0.190	0.444	0.875	0.999
6	0.091	0.209	0.497	0.919	0.9998
8	0.099	0.247	0.597	0.970	1.000
10	0.106	0.283	0.684	0.989	1.000
12	0.112	0.318	0.755	0.996	1.000
15	0.122	0.368	0.835	0.999	1.000
20	0.136	0.451	0.918	0.99998	1.000
25	0.151	0.528	0.962	1.000	1.000
30	0.166	0.599	0.983	1.000	1.000
40	0.195	0.716	0.997	1.000	1.000
50	0.223	0.805	0.9996	1.000	1.000
60	0.252	0.869	0.99995	1.000	1.000
80	0.309	0.944	1.000	1.000	1.000
100	0.366	0.977	1.000	1.000	1.000
120	0.421	0.991	1.000	1.000	1.000
150	0.502	0.998	1.000	1.000	1.000
200	0.622	0.99998	1.000	1.000	1.000
250	0.720	1.000	1.000	1.000	1.000
300	0.799	1.000	1.000	1.000	1.000
400	0.902	1.000	1.000	1.000	1.000
500	0.956	1.000	1.000	1.000	1.000
600	0.982	1.000	1.000	1.000	1.000
800	0.997	1.000	1.000	1.000	1.000
1000	0.9996	1.000	1.000	1.000	1.000

Table 2: Estimated statistical power for the Poisson based MaxSPRT. The type 1 error is $\alpha = 0.05$. T is the upper limit on the length of surveillance, expressed in terms of the expected number of events under the null.

	True Relative Risk				
	1.2	1.5	2	3	5
<i>Time until H_0 is rejected</i>					
T=1	0.26	0.30	0.35	0.39	0.37
2	0.53	0.63	0.75	0.79	0.62
5	1.38	1.85	2.10	1.79	0.83
10	3.02	4.05	4.15	2.46	0.87
20	6.64	8.65	6.97	2.68	0.91
50	19.87	20.41	8.99	2.82	0.96
100	43.73	29.76	9.35	2.92	0.99
200	90.13	32.90	9.66	3.02	1.02
500	172.55	34.30	10.05	3.13	1.06
1000	197.26	35.23	10.33	3.21	1.08
<i>Time until surveillance ends</i>					
T=1	0.95	0.92	0.88	0.77	0.54
2	1.89	1.82	1.68	1.33	0.72
5	4.68	4.40	3.71	2.19	0.83
10	9.26	8.32	6.00	2.54	0.87
20	18.19	14.88	8.04	2.68	0.91
50	43.27	26.19	9.01	2.82	0.96
100	79.43	31.34	9.35	2.92	0.99
200	131.64	32.90	9.66	3.02	1.02
500	186.99	34.30	10.05	3.13	1.06
1000	197.54	35.23	10.33	3.21	1.08

Table 3: The estimated expected length of surveillance for the Poisson based MaxSPRT. The top part of the table is the time until a signal is generated rejecting the null hypothesis, given that the null was rejected. The lower part of the table is the time until the end of surveillance, either because of a signal or because of reaching the upper limit on the length of surveillance. The type 1 error is $\alpha = 0.05$. T is the upper limit on the length of surveillance. All times are expressed in terms of the expected number of events under the null.

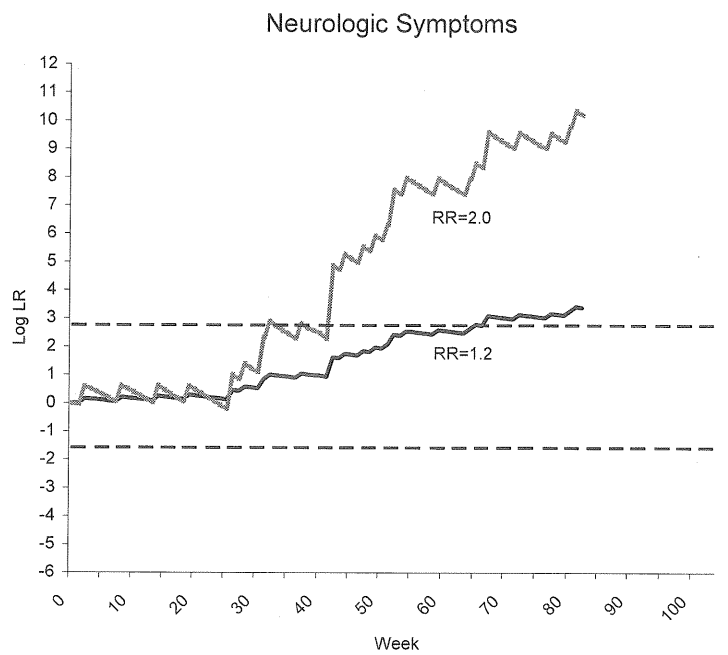
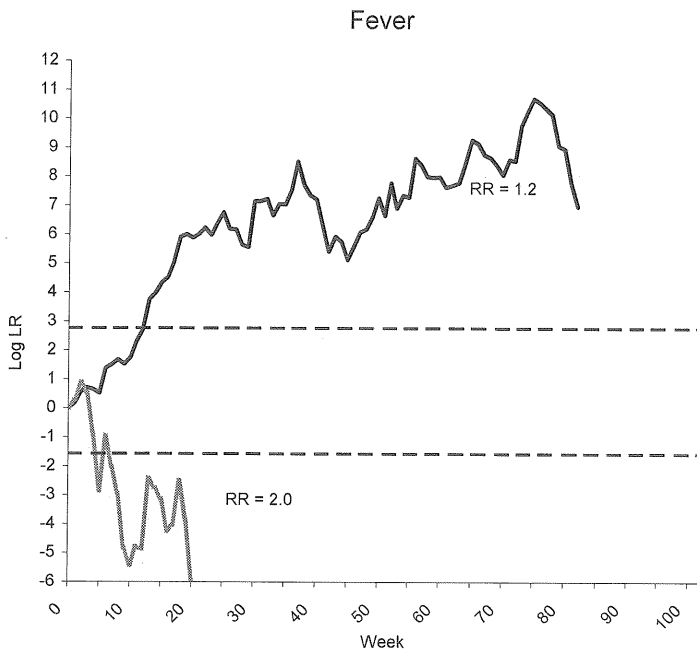
	Matching ratio 1:1			Matching ratio 1:2			Matching ratio 1:3		
	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
N=3	n/a	n/a	n/a	2.19723	n/a	n/a	2.77259	n/a	n/a
4	n/a	n/a	n/a	2.19723	n/a	n/a	2.77259	4.15889	n/a
5	2.77259	n/a	n/a	2.29791	4.39445	n/a	2.77259	4.15889	5.54518
6	2.77259	n/a	n/a	2.29791	4.39445	n/a	2.77259	4.15889	5.54518
8	2.77259	4.15889	n/a	2.90393	4.39445	6.59168	2.77259	4.15889	6.93148
10	2.77259	4.15889	6.23833	2.90393	4.39445	6.59168	2.77259	4.15889	6.93148
12	2.77259	4.27363	6.23833	3.19516	4.39445	6.59168	2.77259	4.39445	6.93148
15	2.89118	4.50710	6.72326	3.19516	4.39445	6.59168	2.77259	4.45847	6.93148
20	3.09884	4.85204	6.93148	3.29584	4.39445	6.89371	2.77259	4.51579	6.93148
25	3.13949	4.85204	6.93148	3.29584	4.59581	7.04215	2.87683	4.51579	6.93148
30	3.39139	4.85204	6.93148	3.29584	4.86283	7.04215	3.00043	4.60292	6.93148
40	3.46574	4.85989	7.34969	3.29584	4.93395	7.31447	3.23539	4.79047	6.97759
50	3.46574	4.88272	7.36129	3.29584	5.08104	7.41709	3.31895	4.92501	7.08667
60	3.46574	4.96051	7.44608	3.29584	5.08160	7.63273	3.33085	5.05772	7.18758
80	3.46574	5.11924	7.62462	3.30986	5.17987	7.69029	3.33085	5.23455	7.37322
100	3.46574	5.31315	7.62462	3.43691	5.30927	7.69029	3.45219	5.36765	7.56519
120	3.47863	5.37043	7.62462	3.48031	5.45081	7.69029	3.45660	5.47676	7.67217
150	3.60597	5.44610	7.64717	3.53437	5.49307	7.69029	3.52540	5.51207	7.70334
200	3.68065	5.51372	7.77200	3.59470	5.49307	7.69029	3.64449	5.54518	7.77539
250	3.75290	5.54518	7.87014	3.67774	5.49307	7.76739	3.72423	5.54518	7.83649
300	3.82197	5.54518	7.94491	3.73387	5.49307	7.78945	3.78278	5.54518	7.89501
400	3.89723	5.54518	7.98030	3.78859	5.49307	7.90802	3.86715	5.54518	7.97126
500	3.95630	5.55891	8.05717	3.87392	5.53040	7.99256	3.92779	5.54518	8.05510
600	3.97652	5.62593	8.12388	3.93147	5.55664	8.04060	3.98762	5.54518	8.11146
800	4.06641	5.68788	8.19592	4.02452	5.62280	8.07877	4.07561	5.60324	8.21196
1000	4.12966	5.76441	8.25052	4.09409	5.67413	8.14459	4.11634	5.64641	8.23855
1200	4.15888	5.79584	8.30479	4.12633	5.71949	8.20833	4.15889	5.69864	8.23855
1500	4.15888	5.83888	8.31777	4.14389	5.78536	8.25265	4.15889	5.73464	8.30161
2000	4.15888	5.91773	8.31777	4.18194	5.83545	8.31836	4.15889	5.74860	8.31777

Table 4: Critical values for the log likelihood ratios from the MaxSPRT for binomial data, for 1, 2 and 3 unexposed individuals per exposed respectively. N is the upper limit on the length of surveillance, defined in terms of the observed number of adverse events. For small N, small alpha levels and few unexposed per exposed, it may not be possible to reject the null even if all adverse events are among the exposed. Such combinations of parameter values make the MaxSPRT non-applicable (n/a).

	Fever		Neurological	
	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$
<hr/>				
MaxSPRT				
T~ 2 years	13	17	42	42
T~ 1 year	13	17	42	42
T~ 3 months	13	17	32	42
<hr/>				
Classical SPRT				
RR=1.05	36	73	>82	>82
1.1	16	30	>82	>82
1.2	13	16	65	>82
1.5	13	13	42	52
2.0	<i>7</i>	<i>7</i>	32	42
5.0	<i>1</i>	<i>1</i>	<i>13</i>	<i>13</i>
10.0	<i>1</i>	<i>1</i>	<i>6</i>	<i>6</i>

Table 5: The number of weeks until a signal is seen for the maxSPRT, with different upper limits on the length of surveillance, and for the classical SPRT, with different relative risks used for the alternative hypothesis. For values in bold, the null hypothesis was rejected, indicating that the vaccine causes fever. For values in italic, the null hypothesis was accepted, indicating that the vaccine does not increase the risk of fever / neurological symptoms. After 82 weeks of surveillance, the observed relative risk was 1.16 for fever and 2.7 for neurological symptoms.

Classical SPRT



MaxSPRT

