HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2019 April 25.

Published in final edited form as:

Stat Med. 2018 January 15; 37(1): 107–118. doi:10.1002/sim.7504.

Type I Error Probability Spending for Post-market Drug and Vaccine Safety Surveillance with Binomial Data

Ivair R. Silva*

Department of Statistics, Federal University of Ouro Preto

Abstract

Type I error probability spending functions are commonly used for designing sequential analysis of binomial data in clinical trials, but it is also quickly emerging for near-continuous sequential analysis of post-market drug and vaccine safety surveillance. It is well known that, for clinical trials, when the null hypothesis is not rejected, it is still important to minimize the sample size. Unlike, in post-market drug and vaccine safety surveillance, that is not important. In post-market safety surveillance, specially when the surveillance involves identification of potential signals, the meaningful statistical performance measure to be minimized is the expected sample size when the null hypothesis is rejected. The present paper shows that, instead of the convex Type I error spending shape conventionally used in clinical trials, a concave shape is more indicated for post-market drug and vaccine safety surveillance. This is shown for both, continuous and group sequential analysis.

Keywords

Sequential probability ratio test; Score statistic; Expected time to signal

1. Introduction

In sequential analysis hypothesis testing, multiple tests are applied to cumulative data in time. Such multiple testing approach favors to anticipate statistically accurate decisions even for small sample sizes. For this reason, sequential analysis is now an important tool for post-market drug and vaccine safety surveillance. When a new drug, or vaccine, is commercialized, fast identification of potential risks from serious adverse events is a central goal for the population's safety[1, 2, 3, 4].

The conventional approach for sequential analysis is based on monitoring a real-valued test statistic in comparison to a flat signaling threshold at each single test. The analysis is interrupted as soon as the test statistic becomes greater than such threshold. The sequential probability ratio test (SPRT) of Wald [5], the Pocock's test [6], and the O'Brien & Fleming's test [7], are some of the conventional sequential methods based on flat thresholds given in the scale of a test statistic. These methods are traditionally used for clinical trials. More

*Correspondence to: Department of Statistics, Federal University of Ouro Preto, Campus Morro do Cruzeiro, CEP 35400 000, Ouro Preto, MG, Brasil. ivairest@gmail.com.

recently, new test statistics has been developed specially for monitoring adverse events caused by recently approved drugs. This is the case of the maximized sequential probability ratio test, introduced by [8], and the conditional MaxSPRT (CMaxSPRT), proposed by [9].

Alternatively, in place of a flat threshold given in the scale of a test statistic, sequential hypothesis testing can be designed through a Type I error spending approach [10]. The Type I error spending is a non-decreasing function, say $S(n)$, taking values in the $[0, \alpha]$ interval, where α is the overall significance level up to the termination of the analysis, and n is the time index defined as a fraction of the total length of surveillance, N. $S(n)$ dictates, in advance, the rate at which the Type I error probability is to be spent in the course of the multiple sequential tests. A common choice is the power-type shape:

$$S(n) = \alpha \times \left(\frac{n}{N}\right)^{\rho},\ \rho > 0,\ n \in (1, N]\,. \quad (1)$$

Most of the discussions on the choice of appropriate $S(n)$ forms can be tracked back to statistical challenges appearing in clinical trials. The book of [10] offers a rich overview on the history and main contributions dealing with usage of Type I error spending in clinical trials. As in clinical trials it is often expensive to recruit many patients to participate of the pre-market study, an important question that has been intensively scrutinized in the last three decades is how to find $S(n)$ shapes in order to optimize the trade-off between statistical power and expected sample size, also called ‘expected time of surveillance’. The expected time of surveillance is a statistical performance measure defined as the average of the sample size at which the surveillance is stopped, which can occur in either of two situations, when the null hypothesis is rejected, or when the null is not rejected. In this direction, [11] performed an intensive study aimed to compare three different Type I error spending shapes. The study revealed that Type I error spending functions with convex shapes will, typically, benefit sample size reductions. In the same direction, [10] showed that, for the specific power-type shape, ρ values around 2, which also leads to a convex shape, can provide a near-optimal balancing between power and expected time of surveillance.

Unlike clinical trials, in post-market safety surveillance it is usually easy to increase the total sample size. This is so because, when a surveillance system is up and running over electronic health records, the costs of keeping the monitoring for some additional months are negligible [12]. Therefore, when the null hypothesis is not rejected, it is not a priority to minimize the sample size. By another hand, as emphatically stated by [12], the very important statistical performance measure in post-market safety surveillance is the expected number of events when the null hypothesis is rejected, called ‘expected time to signal’. The expected time to signal is the conditional expectation that measures the average sample size under the situations where the null hypothesis is rejected. Because many people are exposed to the drug/vaccine, and only a small number of patients may be part of the monitoring system, a delayed identification (signal) of elevated risks can lead to a large number of affected patients. This is the reason why, in post-market safety surveillance, the focusing on finding $S(n)$ shapes for minimizing expected time to signal is much more important than for minimizing expected time of surveillance.

Unfortunately, there is a lack of works dedicated to explore how to choose the Type I error spending in order to minimize expected time to signal. This lack is probably explained by two related reasons: (i) sequential analysis for post-market safety surveillance is just emerging, and the first proposals dates back in no more than 10 years, [4], and (ii) the expected time to signal is a very new concept which was formalized, possibly by the first time through an official publication, in the year of 2015, with the work of [12]. Therefore, finding appropriate $S(n)$ shapes in response to the post-market safety surveillance goals is also a very new challenge.

Aiming to face this challenge for sequential testing with binomial data, the present paper compares four of the classical and most important Type I error spending functions. Again, as this work is concentrated on the goals of the post-market drug/vaccine safety surveillance, the design criterion (performance measure) used in the present comparison study is the expected time to signal. Exact calculations show that the Type I error spending of the form in (1) is, in general, an appropriate choice. The value of ρ that minimizes the expected time to signal depends on the desired α level and on the actual Bernoulli success probability under the alternative hypothesis. But, in most of the cases, the optimal ρ solution is a number in the (0,1) interval, that is, associated to a concave shape, which contrasts to the convex shape usually adopted in clinical trials. Also, we found that ρ values in the range [0.5,1] are appropriate choices in most of the evaluated scenarios. All investigations in this paper are applied for both, continuous and group sequential approaches.

This paper is organized as follows: the next section establishes notation and meaningful definitions. Section 3 offers a brief discussion on how the expected time to signal behaves for the methods of Pocock, O'Brien & Fleming, and MaxSPRT. Section 4 presents the results of the comparison study of four different Type I error spending shapes in terms of expected time to signal. Section 5 closes the paper with the main conclusions.

2. Notation and Definitions

Binomial data take place when, for example, a population is monitored in both, exposed and unexposed periods. This is the case, for example, of self-control studies, where the goal is to compare the period when an individual is exposed to an administered drug against an unexposed period, before administrating the drug, in the same individual. Another situation where a binomial model is applicable is the case where patients exposed to a drug are compared with matched unexposed subjects.

Let C_n denote a non-negative discrete stochastic process of discrete time. For fixed n, assume that C_n is the random variable: number of events from the exposed population when a total of n adverse events is observed.

For the purposes of this paper, the random variable

$$Y_n = C_n - C_{n-1}$$

follows a Bernoulli distribution with success probability $p_{n,R}$, for $n = 1, 2, \ldots$, and $C_0 = 0$. Assume that $Y_1, Y_2, \ldots$ are independent, i.e.:

$$Pr[Y_{n+1} = 1 | Y_1 = y_1, \ldots, Y_n = y_n] = p_{n,R},$$

for any sequence of observed 0's and 1's, $y_1, \ldots, y_n$. In the post-market vaccine safety surveillance context, the success probability is given by:

$$p_{n,R} = 1/\left(1 + z_n/R\right),$$

where z_n is the matching ratio associated to the nth adverse event. For example, if there are 2 controls matched to each case at the very first test, then $z_1 = 2$. In a self-control analysis, z_n is the ratio between the control time window by the risk time window associated to the nth test. For example, if the risk window is 3 days long and the control window is 5 days long when the first test is performed, then $z_1 = 5/3$. The parameter of interest, R, is interpreted as the relative risk implied by the vaccination. If the vaccine is safe, then R is smaller than or equal to 1. But, if the vaccine offers real threats to population's health, then R is greater than 1.

Sequential analysis can be performed by a continuous or a group sequential fashion and, for binomial data, are constructed according to the following definitions.

Definition 1. (Group Sequential Analysis)

For a set of constants $a_1 < a_2 < \cdots < a_G$, and a sequence $\{n_i\}_{i-1}^{G}$ of times taken from the set $\{1,\ldots, N\}$, which is based on the number of events rather than calendar time, a group sequential analysis design for binomial data is any procedure that rejects the null hypothesis if $C_{n_i} \geq a_i$ for some $i \in \{1,\ldots, G\}$, and does not reject the null otherwise.

Definition 2. (Continuous Sequential Analysis)

For a non-decreasing sequence of integers b_n, with $n \in \{1,\ldots, N\}$ representing the number of events rather than calendar time, a continuous sequential analysis design for binomial data is any procedure that rejects the null hypothesis if $C_n \geq b_n$ for some $1 \leq n \leq N$.

The tuning parameter N is an arbitrary positive integer representing the maximum length of surveillance (maximum sample size) for interrupting the surveillance without rejecting the null hypothesis. Continuous sequential designs are directed for applications where the events arrive one-by one, and hence even a single observation can signalize rejection of the null hypothesis. With group sequential designs, the null hypothesis can be rejected only after observing a predefined number of events for each test. Naturally, the choice between continuous and group designs is made according to peculiarities of each application. For example, it can be defined according to logistical and/or financial aspects of data collection. In either approach, the hypotheses considered in this paper are of the form:

$$H_0: R \leq 1 \text{ against } H_1: R > 1. \quad (2)$$

[12] show the important fact, valid not only for binomial data but for any positive stochastic process, that any group sequential design can be rewritten in terms of a continuous design with same or better performance in terms of maximum sample size, Type I error probability, statistical power, and expected time to signal. This is so due to the obvious but still remarkable fact that, for a given group sequential design with signaling threshold a_{n_i} with $i = 1,\ldots, G$, a continuous design with same power and alpha level can be constructed in the following way:

$$b_n = a_{n_i}. \text{ for } n \in (n_{i-1}, n_i], \quad (3)$$

where $N = n_G$. This result is very useful for many reasons, but it is specially convenient in this paper because it ensures that we can express all formulas for statistical performance measures (e.g., power, expected time to signal) using only the notation of continuous sequential analysis without loss of generality.

Let n_1 denote the minimum number of events required before allowing for rejection of H_0, i.e., n_1 is the minimum n such that:

$$Pr[C_n \geq b_n | R = 1] > 0.$$

For matching ratio z_n, signaling threshold b_n, and maximum sample size N, the overall probability of rejecting the null hypothesis is given by:

$$\begin{aligned} \beta(R) &= Pr[C_{n_1} \geq b_{n_1} \cup \cdots \cup C_N \geq b_N | R] \quad (4) \\ &= Pr[C_{n_1} \geq b_{n_1} | R] + Pr[C_{n_1} < b_{n_1} \cap C_{n_1+1} \geq b_{n_1+1} | R] \\ &+ \cdots + + Pr[\cap_{n=n_1}^{N-1} \left\{C_n < b_n\right\} \cap C_N \geq b_N | R] \\ &= \pi_{n_1}(R) + \cdots + \pi_N(R), \end{aligned}$$

where, for $n = n_1$:

$$\pi_{n_1}(R) = Pr[C_{n_1} \geq b_{n_1} | R] = \sum_{c=b_1}^{n_1} \binom{n_1}{c} \left[p_{n,R}\right]^c \left[1 - p_{n,R}\right]^{n_1 - c}. \quad (5)$$

For $n = 2,\ldots,N$:

$$\pi_n(R) = Pr[\cap_{n=n_1}^{N-1}\{C_n < b_n\} \cap C_N \geq b_N | R]$$

$$= \sum_{c_{n_1}=0}^{b_{n_1}-1} \sum_{c_{n_1+1}=c_{n_1}}^{\min\left\{b_{n_1+1}-1, c_{n_1}+1\right\}} \cdots +$$

$$+ \sum_{c_{n-1}=c_{n-2}}^{\min\left\{b_{n-1}-1, c_{n-2}+1\right\}} \sum_{c_n=b_n}^{n} \prod_{j=0}^{N-n_1+1} \left[\frac{p_{n,R}}{1-p_{n,R}}\right]^{c_{n_1+j}-c_{n_j}} [1-p_{n,R}] I(0 \leq c_n - c_{n-1} \leq 1).$$

(6)

Note that expression (4) is non-decreasing with R. This is true because, for each fixed $n \in \{n_1, \cdots, N\}$, the probability of the event $\{C_n \geq b_n\}$ is increasing with R for any choice $b_n \leq n$, and it equals to zero otherwise. The last assertion holds because $p_{n,R}$ is increasing with R Therefore, the overall significance level is calculated by evaluating expression (4) using $R = 1$. Likewise, the punctual Type I error probability, spent at the nth event, is given by $\pi_n(R = 1)$. Therefore, the actual cumulative Type I error probability, spent up to the nth test, is given by:

$$\alpha_n = \sum_{t=1}^{n} \pi_t(R=1). \quad (7)$$

There is an implicit dependence of α_n on the sequence of thresholds $b_{n_1}, \cdots, b_n$, and this can be made explicit through the notation $\alpha_n(b_{n_1}, \cdots, b_n)$. But, for finding critical values during the planing phase, the calculation has to be done in the opposite direction, i.e., one has to elicit critical values that match with the target Type I error spending. Thus, if $n > n_1$, and for fixed constants $b_{n_1}, \cdots, b_{n-1}$, the critical value, b_n, of the nth test, is given by:

$$b_n = \min\left\{j \in [1, 2, \cdots] : \alpha_n(b_{n_1}, \cdots, b_{n-1}, j) \leq S(n)\right\}. \quad (8)$$

If $n = n_1$, then:

$$b_{n_1} = \min\left\{ j \in [1,\ 2, \cdots \rceil : \alpha_{n_1}(j) \leq S(n_1) \right\}, \quad (9)$$

Under this notation, as C_n has a discrete support, there will exist testing times where there is no positive punctual Type I error probability spending due to the impossibility of attending the target S_n, and such cases will lead to $b_n = n + 1$, that is, $\alpha_n - \alpha_{n-1} = 0$.

Usually, the signaling threshold is not established directly in the scale of C_n through b_n, but constructed in terms of a flat critical value, cv, which is usually a number in the scale of a real-valued test statistic, $W(C_n)$. That is, the sequential analysis is interrupted, and H_0 is rejected, for the first n such that:

$$W(C_n) \geq cv\,. \quad (10)$$

For arbitrary α, if $W(C_n)$ is non-increasing with C_n, the exact flat thresholds given in the scale of $W(.)$ can be obtained through a numerical bisection algorithm, [8]. This is the case, for instance, of the most important sequential testing methods, such as Pocock's method, [6], O'Brien & Fleming's test, [7], maximized sequential probability ratio test (MaxSPRT), [8], and the modified MaxSPRT, [13]. Table 1 presents the test statistics related to each of these methods.

3. Expected Time to Signal

Now, we formally define the most important statistical performance measure of this paper, the so called 'expected time to signal'. Let T denote the number of events when the surveillance is interrupted. The expected time to signal is a conditional expectation, denoted by $\mathbb{E}[T|H_0 \text{ rejected},\ R]$, given by:

$$\mathbb{E}[T|H_0 \text{ rejected},\ R] = 1 \times Pr[T = 1|H_0 \text{ rejected},\ R] + 2 \times Pr\left[T = 2|H_0 \text{ rejected},\ R\right] + \cdots\cdots + N \times Pr[T = N|H_0$$

$$\text{rejected},\ R] = 1 \times \frac{Pr[T = 1|R]}{Pr\left[H_0 \text{ rejected}|R\right]} + 2 \times \frac{Pr[T = 2|R]}{Pr[H_0 \text{ rejected}|R]} + \cdots\cdots + N \times \frac{Pr[T = N|R]}{Pr[H_0 \text{ rejected}|R]}$$

$$[\text{from (4), (5), and (6)}] = \frac{\sum_{n=1}^{N} n \times \pi_n(R)}{\beta(R)}\,. \quad (11)$$

Note that the term $Pr[T = n/R]$ is also denoted by $\pi_n(R)$, which can be calculated according to (6).

Statistical performance measures, such as expected time to signal, are strongly dependent upon the Type I error spending shape. Also, it is important to emphasize: the expected time

to signal will usually differ from the expected time of surveillance for a fixed Type I error spending shape. The expected time of surveillance, denoted by $\mathbb{E}[T|R]$, is given by:

$$\mathbb{E}[T|R] = 1 \times Pr[T = 1|R] + 2 \times Pr[T = 2|R] + \cdots + N \times Pr[T = N|R] = 1 \times \pi_1(R) + 2 \times \pi_2(R) + \cdots + N \times \pi_N(R) + N \times Pr[H_0 \text{ not rejected}|R]$$

$$[\text{from } (4)] = \sum_{n=1}^{N} n \times \pi_n(R) + N \times [1 - \beta(R)]\,. \quad (12)$$

Although these measures are related to each other, the Type I error spending shape that minimizes $\mathbb{E}[T]H_0$ rejected, $R]$ will usually differ from the shape that minimizes $\mathbb{E}[T|R]$ if compared under the same fixed power. For example, consider again the power-type spending shape, i.e., $S(n) = \alpha\,(n/N)^{\rho}$, and a constant matching ratio $z_n = z = 1$. For a significance level $\alpha = 0.05$, we can elicit the minimum sample size that attains a power of 0.8 for a given target relative risk, say r, and fixed ρ. Hence, it is possible to evaluate the behavior of each performance measure for arbitrary scenarios of true R values under a range of target r and ρ choices.

It is important to stress the distinction between the notations R and r. While R represents the unknown parameter of interest, the term r works as a tuning parameter for sequential analysis designing. For example, one can base the choice of the maximum length of surveillance under the requirement of detecting relative risks of at least $r = 2$ with probability (power) of at least 0.8. But, the actual performance of such a sequential design depends on the actual R, that is, the actual power will be smaller than 0.8 if $R < 2$, and it will be greater than 0.8 if $R > 2$. We could assume, by simplicity, that target and actual relative risks coincides. For instance, take the scenarios $R = r = 1.2, 1.3, 1.5, 2$, having target power of 0.8 for $\alpha = 0.05$, and $z = 1$. In this case, if we calculate expected time to signal and expected time of surveillance for each ρ value in [0.1, 0.2, …, 3], we will find that the ρ values that minimize expected time to signal are 1.2, 1, 0.8, and 0.9, for $R = 1.2, 1.3, 1.5, 2$, respectively. Differently, for the same R values, the ρ values that minimize expected time of surveillance are 1.5, 1.7, 2, and 1.2. These results are shown in Table 3 and illustrated with Figure 1 of the Supplementary Material Part I.

We observe that solutions minimizing expected time to signal are related to ρ values smaller than or around 1, and those for minimizing expected time of surveillance are greater than 1. But, situations where $R = r$, as discussed in the paragraph above, are unlikely in practice. Actually, it is more likely that the tuning parameter r and the actual R will differ in practice.

For a more realistic evaluation, Figure 1 presents expected time to signal and expected time of surveillance as a function of ρ, but also considering scenarios where r and R are different. This is done for $R = 1.2, 1.3, 1.5, 2$, with $r = 2$ in all scenarios. Each line represents a fixed R value, where $R = 1.2$ is represented by the thinnest dashed line, $R = 1.3$ has the thinnest solid line, $R = 1.5$ has the the thickest dashed line, and $R = 2$ is represented by the thickest solid

line. Little circles are used to point the optimal ρ values of each scenario. Figure 1(A) has expected time to signal $\left(\mathbb{E}[T]H_0 \text{ rejected, } R]\right)$. The minimum values occur for ρ values of 0.1,0.2,0.9, and 0.9 for R equal to 1.2,1.3,1.5, and 2, respectively. These solutions lead to concave shapes. Figure 1(B) brings the curves for expected time of surveillance $(\mathbb{E}[T|R])$. There the optimal ρ values are 2.2, 2.2, 2.2, and 1.2 for R = 1.2, 1.3, 1.5, 2, respectively. These solutions lead to convex shapes. Therefore, ρ solutions minimizing $\mathbb{E}[T]H_0 \text{ rejected, } R]$ greatly differ from the solutions minimizing $\mathbb{E}[T|R]$.

While it is well-known in the literature of sequential analysis that convex shapes are indicated to optimize expected time of surveillance for fixed powers, a concave shape seems to be a better option if expected time to signal is the important measure. This preliminary evidence shall be further explored in Section 4.

3.1. Expected Time to Signal and Conventional Methods

As already stated in Section 2, sequential analysis designs that are defined in terms of test statistics are just indirect ways of selecting $S(n)$ shapes. Hence, certain test statistics might perform very well for clinical trials, where $\mathbb{E}[T|R]$ is the meaningful measure, but not so well for post-market safety surveillance, where $\mathbb{E}[T]H_0 \text{ rejected, } R]$ is a very important measure.

Therefore, as sequential analysis methods are rapidly embracing post-market safety surveillance problems, it is important to explore the relation between time to signal and some of the well-known test statistics.

Table 2 offers critical value, power, and expected time to signal for each of the test statistics presented in Table 1. Figure 2 enables to evaluate how the performance measures relate to the the Type I error spending shape implied by each statistic. The maximum sample sizes used in Table 2, denoted by N_0, were fixed at the minimum values satisfying solutions under a target statistical power of 0.99, fixed relative risk of R = 2, and α = 0.01, 0.05. For simplicity, all calculations in this table are based on non-variable matching ratios, for which two scenarios were considered, z_n = 1 and z_n = 4.

Important: the exact performances in Table 2 differ from the target ones. The actual size, denoted by α_N, and power, denoted by $\beta(R)$, are not exactly equal to α and 0.99, respectively, due to the discrete nature of C_n.

We see from Table 2 that the smallest expected time to signal is promoted by Pocock's approach, and the largest one occurs with O'Brien & Fleming's method (O'BF). This can sound controversial at a first sight. O'Brien & Fleming's statistic is just a linearly weighted version of the score statistic, just as it is Pocock's statistic. Then, why do they perform so differently? A similar phenomenon is observed between the second best approach, the MaxSPRT test, and the second worse approach, modified MaxSPRT (M. MaxSPRT), since they are obtained by scalar transformations of the same statistic, the likelihood ratio test statistic.

Actually, what the best approaches have in common is the concave shape of their implied Type I error spending. For instance, take z = 1 in Figures 2(A) and 2(B). The Type I error

spending associated to the best, Pocock's statistic (black solid line), has a concave shape, exactly as the second best curve (black dotted line) from MaxSPRT. Now, take $z = 4$. In this case, Figures 2(C) and 2(D), Pococok's statistic and MaxSPRT switch places, i.e., MaxSPRT presents the smallest $\mathbb{E}[T]H_0$ rejected, R] and Pocock's statistic presents the second smallest. Among the concave curves, what determines the best is the way at which the speed of increasing is managed at the beginning and at the end of the surveillance. For $z = 1$, the curve for MaxSPRT is placed above Pocock's curve for small n, but then they cross each other and keep position until the end of the surveillance. For $z = 4$, we observe the reverse, i.e., Pocock's curve is placed above MaxSPRT's curve at the beginning of the surveillance, but they cross each other in a long term surveillance. It means that, as expected, spending a high amount of Type I error probability at the beginning of the surveillance leads to a fast detection of increased risks, but it has to be done in a balanced way in order to save sufficient Type I error probability for intermediate times of analysis. Although concave shapes and high alpha spending at the beginning of the surveillance characterize the designs of small expected time to signal, this is not a sufficient condition for achieving the best performance. There is a balance between power spending and time to signal, and, as shall be shown in Section 4, such a balance depends on z and R. Hence, the point is not if the best function is more, or less, concave, but that the proper concavity depends on the values of N, R, z and target power.

It is worth noting that the judgment of the best or worse is relative. It depends on the target performance measure. For instance, if expected sample size is the main performance measure under concern, hence convex shapes, like the O'Brien & Fleming test and the modified MaxSPRT, would present the best performances, [13].

Naturally, the shapes identified above (between convex or concave), for the implicit Type I error spending of each test statistic, are obtained when flat signaling thresholds are used. But, any of these methods could promote either concave or convex shapes by strategical usage of time-varying thresholds. For example, in the context of post-market safety surveillance, [14] generalizes MaxSPRT for Poisson data by considering the usage of non-flat critical values according to target Type I error probability spending. By comparing four different critical value functions in terms of statistical power and expected time to signal, they argue that, if an adverse event is rare but can lead to severe harmful consequences, then early rejection of H_0 should be permitted, that is, decreasing signaling thresholds should be preferred. But, if adverse events are not rare and of less severeness, then increasing thresholds should be adopted.

4. Comparison of Type I Error Spending Shapes in Terms of Expected Time to Signal

This study compares the following four shapes of Type I error spending:

$$S_1(n) = \alpha \times \left(\frac{n}{N}\right)^{\rho}, \rho > 0,$$

$$S_2(n) = 2 - 2 \times \Phi\left(x_{\alpha} \times \sqrt{N/n}\right), \text{ where } x_{\alpha} \text{ is such that } \Phi^{-1}(1 - \alpha/2) = x_{\alpha},$$

$$S_3(n) = \alpha \times \log\left\{1 + [\exp(1) - 1] \times \frac{n}{N}\right\},$$

$$S_4(n) = \alpha \times [1 - \exp\{-\frac{n\gamma}{N}\}]/[1 - \exp\{-\gamma\}].$$

According to [11], [15] and [16], the power-type function, $S_1(n)$, produces good approximations for the Pocock's and O'Brien & Fleming's tests for specific values of ρ. But, this function is also appropriate to mimic the MaxSPRT and the modified MaxSPRT designs. MaxSPRT and modified MaxSPRT match reasonably well to $S_1(n)$ for $\rho = 0.28$ and $\rho = 2.5$, respectively. Recall that $S_1(n)$ produces a line for $\rho = 1$, a convex curve for $\rho > 1$, and a concave curve for $0 < \rho < 1$. [10] offers a comprehensive overview of the literature concerned with good choices for ρ in the sense of producing a favorable balancing between expected time of surveillance and statistical power. Such studies suggest that ρ values around 2 (i.e., convex shape) provide small expected time of surveillance. There are not similar studies concerned with expected time to signal. Regarding $S_2(n)$, as verified by [17], that function is appropriate for approximating the O'Brien & Fleming's error spending. [17] also introduced $S_3(n)$ in order to approximate the error spending of Pocock's test, but the function $S_4(n)$, introduced by [18], fits far better to Pocock's Type I error spending curve than $S_3(n)$. This is illustrated with Figure 2 of the Supplementary Material Part I. The function S_4 is concave for $\gamma > 0$ and convex for $\gamma < 0$, and it is flat for $\gamma = 0$.

4.1. Tuning Parameters Settings

This comparison study is based on the following tuning parameters choices. The scenarios for the significance level are $\alpha = 0.01, 0.025, 0.05, 0.1$. The target powers are $\beta(r) = 0.8, 0.9, 0.99$. For each signaling threshold associated to a fixed R in the set {1.2, 1.3, 1.5, 2, 3, 4}, expected time to signal was calculated for target relative risks of $r = 1.2, 1.3, 1.5, 2, 3, 4$. For the matching ratio, it were considered fixed values $z_n = z$, given by $z = 0.5, 1, 2, 4$. Due to space restrictions for showing the numerical results of so many scenarios, only a fraction of the results are explicitly in the body of this material. But, the available numbers favor a resume of the general conclusions achieved with this intensive study that is also supported by many other tables and additional figures available in the Supplementary Material Part I. Calculations for any other scenario not shown here can be easily performed by a simple 'copy and paste' of the code, written in **R** language, [19], and available in the Supplementary Material Part II.

4.2. Comparison Under a Continuous Sequential Fashion

The forms $S_1(n)$ and $S_4(n)$ define entire families of functions. Some choices of ρ and γ are best suited than others. For $\alpha = 0.05$ and a target power of $\beta(2) = 0.99$, Figure 3 illustrates the behavior of the expected time to signal under a relative risk of $R = 2$.

It is clear that the expected time to signal depends on the value of ρ for $S_1(n)$, and on γ for $S_4(n)$. Thus, the comparison of $S_1(n)$ and $S_4(n)$ against each other, and against $S_2(n)$ and $S_3(n)$, demands a previous step: finding the optimal ρ and γ, in the sense of minimizing expected time to signal, for each scenario of r, power, α, and z. Such a preliminary investigation was based on the values of $\rho = 0.01, .02, \cdots, 2$ and $\gamma = -10, -0.99, \cdots, 0, 0.01, .02, \cdots, 10$. For target powers of 0.9 and 0.99, and significance levels of 0.01 and 0.05, Table

1 and Table 2 of the Supplementary Material Part I present the solutions, with a precision of two decimal places, for the choices of ρ and γ that minimize $\mathbb{E}[T]H_0$ rejected, R] rejected, R].

For $S_1(n)$, in most of the cases, the optimal solution for ρ is a value smaller than 1, leading to concave Type I error spending shapes. But, the exact optimal solution depends on z and α. As a whole, and as illustrated with Figure 3(A), the optimal solution is around 0.5 in most of the scenarios. For $S_4(n)$, as one can follow with Figure 3(B), the optimal γ is uniformly greater than zero, hence, again, concave Type I error spending shapes are the best options.

For a significance level of 0.01, Table 3 presents expected time to signal calculated for each Type I error spending function. Similarly, Table 5 and 6 of the Supplementary Material Part I show expected time to signal for $\alpha = 0.05$. The numbers for $S_1(n)$ and $S_4(n)$ are based on the optimal solutions for ρ and γ of Tables 1 to 4 of the Supplementary Material Part I.

As a rule, the power-type shape, $S_1(n)$, is the best option in scenarios where $R > r$. If $R \leq r$, then $S_4(n)$ performs better. In the other side of the bridge, function $S_2(n)$ presents invariably the largest expected time to signal.

4.3. Comparison Under a Group Sequential Fashion

Near-continuous sequential analysis, where the total number of events (cases+controls) at each test can be greater than 1, is more realistic than assuming strict continuous analysis. But, although irregular and sometimes even unpredictable, the number of events at each test is not expected to vary much during the surveillance. For example, a team of analysts may have a good idea about the sample size to be observed from electronic health records, like e.g. 10 events per record, but they also may expect that the actual numbers will show up a little smaller value, like 6 events, or a little larger, like 15, but unlikely much different, like 50. Thus, regular group sizes, where the number of events in each test is constant, is treated here as a good approximation for the reality.

Denoting the number of tests by G, there were considered three different scenarios for the group sizes, $G = 2, 5, 20$. Tables 7, 8 and 9 of the Supplementary Material Part II show the expected time to signal for each of these scenarios. Again, the superiority of $S_1(n)$ is evident for most of the scenarios where $R > r$, hence $S_4(n)$ presents smaller expected time to signal for $R \leq r$. There are some exceptions, but, as a whole, the results are similar to what we observed for the continuous sequential approach. These results are not affected by the values of the parameters G, α or z.

5. Concluding Remarks

In general, functions $S_1(n)$ and $S_4(n)$ promote better performances in terms of expected time to signal. But, the winner between these two functions depends on the relative magnitude between the actual relative risk (R) and the relative risk (r) that solves a given target power of interest. $S_4(n)$ tends to promote smaller expected time to signal than $S_1(n)$ in applications where $R \leq r$. But, if $R > r$, then $S_1(n)$ tends to show better performances. These are general results, but exceptions were observed since the behavior of power and expected time to signal is not uniform with the tuning parameters ρ and γ, and this is so due to the discrete nature of the binomial distribution. Another important point to stress is that, in all scenarios,

the optimal solution was found among concave functions. This fact is not in contradiction with the fact that, in few scenarios, ρ solutions were slightly greater than 1, cases that lead to convex Type I error spending shapes for $S_1(n)$. Such cases occurred for $R \leq r$, which are the scenarios where $S_4(n)$ was found the best option under positive γ values, that is, concave Type I error spending shapes.

Although concavity is a remarkable property of Type I error spending functions that minimize expected time to signal, this is not a sufficient condition for ensuring the best performance. There is also a trade-off between power spending and time to signal, which depends on *N, r, z, R*, and target power. Therefore, tuning parameters that define the shape to use, such as ρ and γ, should be elicited during the sequential analysis designing as it takes in account the tuning parameters and target performance measures of each application.

The assumption of a constant marching ratio is near-realistic. In practice, the Bernoulli success probability can present a certain variation while the surveillance advances, and an infinite number of possibilities for the trajectories of z_n could be drawn in each application. But, as the results indicate that the superiority of $S_1(n)$ and $S_4(n)$ is observed for all the evaluated z values, it is safe to conclude that the usage of a constant z in the present investigations leads to an accurate understanding of the correspondence between Type I error spending and expected time to signal.

Besides the concave Type I error spending shape, another remarkable characteristic that favors reduction of the expected time to signal is the requirement of a slow spending rate in initial surveillance times. From Figure 2, we see that the best shape among those options allocates moderate to small Type I error spending at the beginning of the surveillance. This result explains the findings of [20], which evidenced that requiring a minimum number of events before allowing the rejection of the null hypothesis, same as putting very small Type I error spending for first events, can lead to relevant gains in terms of expected time to signal.

As emphasized in early sections, in post-market surveillance it is cheap and easy to collect large data sets once a monitoring system is set up and running. Therefore, minimizing the overall sample size at termination of the surveillance is not imperative. In theory, if the null hypothesis is rejected, then the surveillance should stop. But, in practice, one could keep collecting/analyzing data in order to access more evidence rather than stopping the monitoring to make an early decision even when the null is rejected. This is so because sequential analysis can be performed for different goals, such as point estimate, interval estimation, hypothesis testing, or more than one of these goals simultaneously. Thus, when a signalization for rejection of H_0 occurs, irrespectively of having a small, moderate or large sample size at the signalization moment, the surveillance can still continue in order to collect more data for ensuring a more accurate relative risk estimation. Again, this is applicable in situations where the costs of keeping collecting data for a longer time is negligible. While an early signal is valuable for preliminary actions and further investigations, investigators should take advantage of an alive surveillance system for improving the precision of point and interval estimation, and for ratifying evidences obtained from early analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by the National Institute of General Medical Sciences, USA, grant #RO1GM108999. Additional support was provided by Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Minas Gerais, Brazil (FAPEMIG), andbyPROPP-UFOP.

References

1. 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. [PubMed: 15824549]
2. Lieu TA, Kulldorff M, Davis RL, et al. Real-Time Vaccine Safety Surveillance for the Early Detection of Adverse Events. Medical Care. 2007; 45(S):89–95.
3. Yih WK, Kulldorff M, Fireman BH, et al. Active Surveillance for Adverse Events: The Experience of the Vaccine Safety Datalink Project. Pediatrics. 2011
4. Leite A, Andrews NJ, Thomas SL. Near real-time vaccine safety surveillance using electronic health recordsa systematic review of the application of statistical methods. pharmacoepidemiology and drug safety. 2016; 25:225–237. [PubMed: 26817940]
5. Wald A. Sequential Tests of Statistical Hypotheses. Annals of Mathematical Statistics. 1945; 16:117–186.
6. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64:191–199.
7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979; 35:549–556. [PubMed: 497341]
8. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. Sequential Analysis. 2011; 30:58–78.
9. Li L, Kulldorff M. A Conditional maximized sequential probability ratio test for pharmacovigilance. Statistics in Medicine. 2010; 29:284–295. [PubMed: 19941282]
10. Jennison, V, Turnbull, BW. Chapman and Hall/CRC London. 2000.
11. Kim K, Demets DL. Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika. 1987; 74(1):149–154.
12. Silva IR, Kulldorff M. Continuous versus Group Sequential Analysis for Post-Market Drug and Vaccine Safety Surveillance. Biometrics. 2015; 71(3):851–858. [PubMed: 26011024]
13. Gombay E, Li F. Sequential Analysis: Design Methods and Applications. Sequential Analysis. 2015; 34:57–76.
14. Li R, Weintraub E, McNeil MM. Continuous Sequential Boundaries for Drug and Vaccine Safety Surveillance. Statistics in Medicine. 2014; 33(19):3387–3397. [PubMed: 24691986]
15. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach(with disussion). Journal of the Royal Statistical Society B. 1989; 51:305–361.
16. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: A review and commentary. Statistical Science. 1990; 5:299–317.
17. Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical Trials. Biometrika. 1983; 70(3):659–663.
18. Hwang IK, Shih WJ, DeCani JS. Group sequential designs using a family of type I error probability spending functions. Statistics in Medicine. 1990; 9:1439–1445. [PubMed: 2281231]
19. Team R Core R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna; Austria: 2015. http://www.R-project.org
20. Kulldorff M, Silva IR. Continuous Post-market Sequential Safety Surveillance with Minimum Events to Signal. 2015

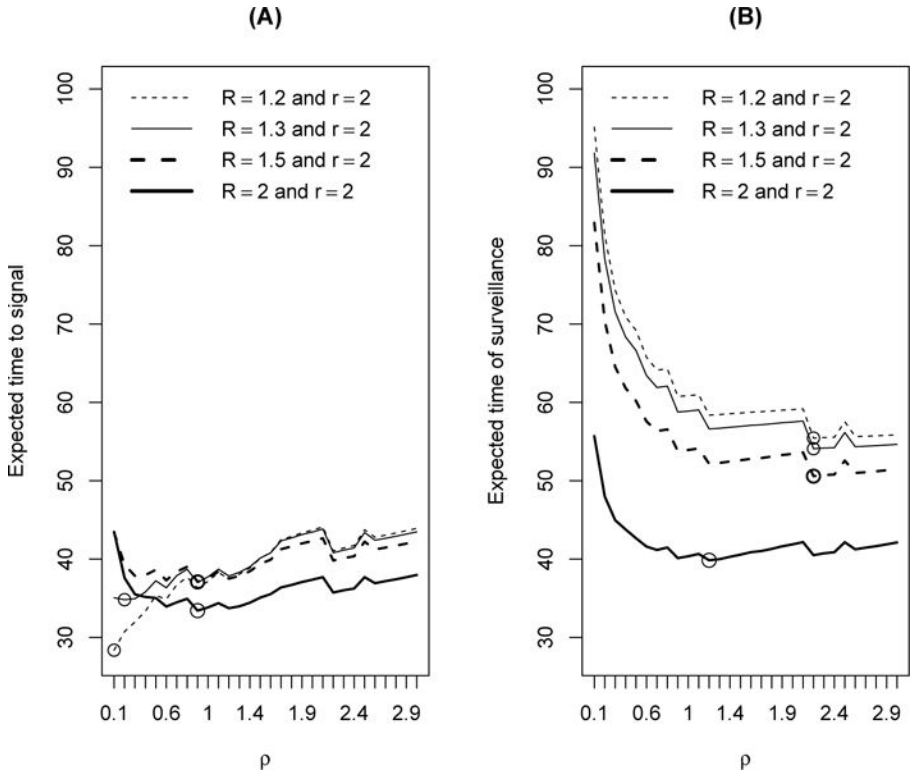

Figure 1.

Expected time to signal and expected time of surveillance for different Type I error spending shapes under $z = 1$, $\alpha = 0.05$, power of 0.8, ρ values in [0.1, 0.2, ⋯, 3], actual relative risks of $R = 1.2, 1.3, 1.5, 2$, and target relative risk of $r = 2$. Small circles point the ρ values that minimize the performance measures at each scenario.

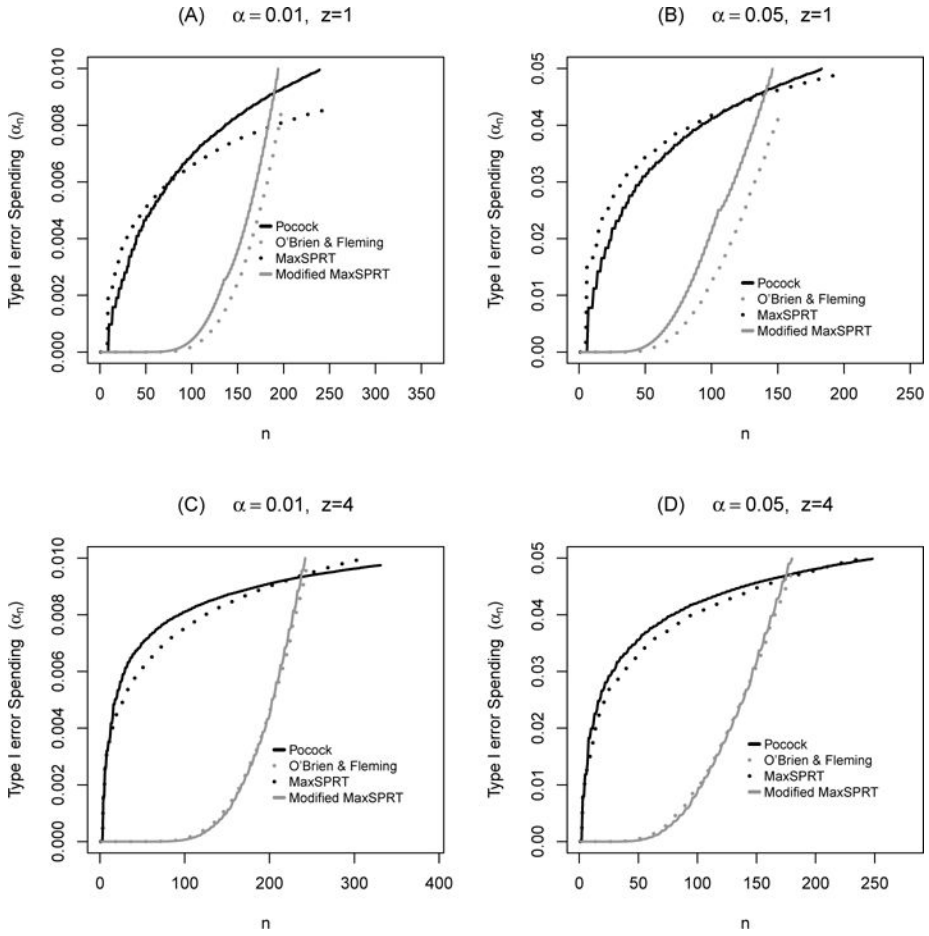

Figure 2.

Actual Type I error spending for continuous sequential testing, based on binomial data, with target power of 0.99 under $R = r = 2$, $\alpha = 0.01, 0.05$, and constant matching ratio $z_n = z$ Of 1 and 4.

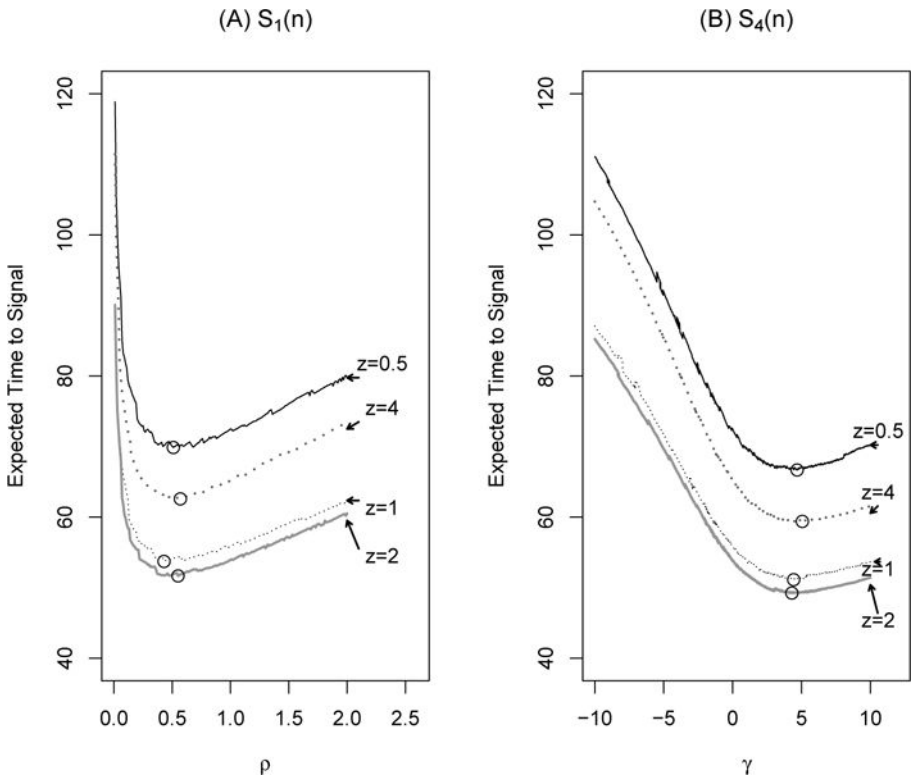

Figure 3.

Expected time to signal calculated for continuous sequential designs based on S_1 (n) and $S_4(n)$ considering a grid of values of $\rho = 0.01, .02, \cdots, 2$ and $\gamma = -10, -0.99, \cdots, 0, 0.01, .02, \cdots, 10$. The expected time to signal is evaluated for a true relative risk of $R = 2$, and a target power of 0.99 given target $r = 2$ under an overall significance level $\alpha = 0.05$. Minimum values are marked with little circles at each curve.

Table 1

Test Statistics for Sequential Analysis with Binomial Data. The parameter setting under H_0 is $p \leq p_0$, where $p_0 = (1 + z_n)^{-1}$

Pocock's Score Statistic (U_n)	O'Brien & Fleming
$W(C_n) = U_n = \begin{cases} \frac{c_n - np_0}{\sqrt{np_0(1-p_0)}}, & if \frac{c_n}{n} > p_0, \\ 0, \text{ otherwise}. \end{cases}$	$W(C_n) = \begin{cases} U_n\left(\frac{n}{N}\right)^{1/2}, & if \frac{c_n}{n} > p_0, \\ 0, \text{ otherwise}. \end{cases}$
MaxSPRT statistic (Λ_n)	
$W(C_n) = \Lambda_n \begin{cases} c_n\left(\log\frac{c_n}{n} - \log p_0\right) + (n - c_n)[\log\frac{n - c_n}{n} - \log(1 - p_0)], & if \frac{z_n c_n}{(n - c_n)} > 1, \\ 0, \text{ otherwise}. \end{cases}$	
Modified MaxSPRT	
$W(C_n) = \left(\frac{2n}{N}\Lambda_n\right)^{1/2}$	

Table 2

Expected time to signal for continuous sequential testing with target power of 0.99 ($r = 2$), $\alpha = 0.01, 0.05$, and constant matching ratios, $z_n = z$, of 1 and 4, for $n = 1,\ldots, N$.

	Test Statistic		cυ	N_0	α_N	$\beta(2)$	$\mathbb{E}[T]H_0$ rejected, $R = 2]$
z=1		$\alpha = 0.01$	3.160111	239	0.009958	0.9900	**82.23**
	Pocock	$\alpha = 0.05$	2.581989	183	0.049961	0.9900	**52.88**
		$\alpha = 0.01$	2.62286	199	0.008723	0.9906	109.97
	O'BF	$\alpha = 0.05$	1.953092	151	0.042229	0.9910	74.14
		$\alpha = 0.01$	5.545369	255	0.008643	0.9906	88.10
	MaxSPRT	$\alpha = 0.05$	3.680642	194	0.048905	0.9901	55.49
		$\alpha = 0.01$	2.52728	194	0.009990	0.9901	104.27
	M. MaxSPRT	$\alpha = 0.05$	1.910947	146	0.049936	0.9907	68.29
			cv	N_0	α_N	$\beta(2)$	$\mathbb{E}[T]H_0$ rejected, $R = 2]$
z = 4		$\alpha = 0.01$	3.618136	331	0.009746	0.9902	104.68
	Pocock	$\alpha = 0.05$	2.882307	248	0.049870	0.9902	64.26
		$\alpha = 0.01$	2.624756	244	0.009623	0.9902	124.79
	O'BF	$\alpha = 0.05$	1.969729	181	0.049028	0.9902	81.45
		$\alpha = 0.01$	5.418628	309	0.009987	0.9900	**101.00**
	MaxSPRT	$\alpha = 0.05$	3.551963	238	0.049900	0.9902	**63.65**
		$\alpha = 0.01$	2.523166	242	0.009993	0.9904	127.12
	M. MaxSPRT	$\alpha = 0.05$	1.899117	180	0.049902	0.9900	83.19

Table 3

Comparing type I error spending functions in terms of expected time to signal, $\mathbb{E}[T]H_0$ rejected, $R]$, for an overall significance level of α = **0.01** in a continuous sequential fashion. Target power of 0.9 and 0.99 for target relative risk of r = 2. It were considered fixed matching ratios of z = 0.5, 1, 2, 4

		$\beta(2) \approx 0.9$				$\beta(2) \approx 0.99$			
		$S_1(n)$	$S_2(n)$	$S_3(n)$	$S_4(n)$	$S_1(n)$	$S_2(n)$	$S_3(n)$	$S_4(n)$
R = 1.2	z = 0.5	8.72	12.42	8.64	8.10	19.38	28.67	19.59	18.02
	z = 1	6.79	9.13	6.51	6.32	14.60	21.18	14.76	13.24
	z = 2	6.75	8.49	5.99	5.72	13.59	20.07	13.69	12.39
	z = 4	7.59	9.73	6.82	6.50	15.86	23.98	16.24	14.32
R = 1.5	z = 0.5	46.61	57.52	48.45	46.39	103.83	122.99	103.20	101.73
	z = 1	35.25	42.81	37.08	35.53	80.20	94.42	79.85	77.82
	z = 2	32.72	40.57	34.50	33.08	76.29	91.79	76.25	74.39
	z = 4	37.93	47.73	39.68	38.46	91.43	111.94	92.35	88.59
R = 2	z = 0.5	88.62	101.85	89.43	88.46	107.71	136.88	106.57	105.62
	z = 1	67.93	78.12	68.98	67.89	83.19	106.99	82.26	81.21
	z = 2	65.63	75.83	65.62	64.82	79.67	104.47	79.10	78.00
	z = 4	78.28	92.38	78.23	77.55	96.46	128.20	95.55	94.06
R = 4	z = 0.5	41.18	69.33	43.18	42.95	38.43	80.38	41.81	40.79
	z = 1	28.81	51.40	30.45	30.92	26.69	60.07	29.47	28.50
	z = 2	23.93	46.33	25.70	25.89	22.37	54.22	25.26	24.43
	z = 4	25.32	52.73	27.45	27.32	23.41	61.60	26.71	25.72