Published in final edited form as:

Stat Med. 2020 February 10; 39(3): 340–351. doi:10.1002/sim.8405.

Exact sequential analysis for multiple weighted binomial end points

Ivair R. Silva¹, Joshua J. Gagne², Mehdi Najafzadeh², Martin Kulldorff²

¹Department of Statistics, Federal University of Ouro Preto, Brazil

²Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts

Abstract

Sequential analysis is used in clinical trials and postmarket drug safety surveillance to prospectively monitor efficacy and safety to quickly detect benefits and problems, while taking the multiple testing of repeated analyses into account. When there are multiple outcomes, each one may be given a weight corresponding to its severity. This paper introduces an exact sequential analysis procedure for multiple weighted binomial end points; the analysis incorporates a drug's combined benefit and safety profile. It works with a variety of alpha spending functions for continuous, group, or mixed group-continuous sequential analysis. The binomial probabilities may vary over time and do not need to be known *a priori*. The new method was implemented in the free **R** *Sequential* package for both one- and two-tailed sequential analysis. An example is given examining myocardial infarction and major bleeding events in patients who initiated non-steroidal antiinflammatory drugs.

Keywords

sequential testing; Type I error spending function; variable matching ratio

1 | INTRODUCTION

The standard strategy for statistical hypothesis testing is based on a decision rule at a single analysis. With such single-sample methods, the analyst has to wait until collecting the total number, say, N , of observations before performing the analysis. In contrast, with statistical sequential analysis, the analyst does not have to wait until observing the whole N -dimensional data but, instead, performs repeated analyses using cumulative data accruing over time, which can potentially lead to earlier but still accurate decisions.^{1–3}

In sequential data analysis, one of the most important probability models is the binomial distribution. For instance, in postmarket drug and vaccine safety surveillance, when a new

Correspondence: Ivair R. Silva, Department of Statistics, Federal University of Ouro Preto, Campus Morro do Cruzeiro, CEP, Ouro Preto-MG 35400-000, Brazil. ivaiarest@gmail.com.

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings of this study are available within the article.

product is commercialized after a phase III clinical trial, it can still cause unexpected adverse reactions in the exposed population, such as seizures. To test whether the product is associated with seizures, we can compare the number of events in a postexposure period following the vaccination, with a subsequent comparison period.⁴ In such a context, an adverse event is a Bernoulli random variable where the success probability is given by the ratio of the number of events during the comparison period and the total (postexposure + comparison) period. Bernoulli data are also commonly used in clinical trials. For example, suppose that a drug is administered to a selected group of patients and that an equal number of subjects were not exposed to the drug. In addition, assume that the studied end point is whether stroke occurs or not (0-1 response). Thus, we can compare the number of exposed subjects presenting with strokes against the number of unexposed subjects presenting with strokes. The ratio between the number of events in the unexposed by the number of events in the exposed subjects determines the Bernoulli probability of occurrence in the exposed group.

The two examples above describe single outcome sequential analysis, where just one outcome is entered in the sequential testing decision rule. Instead of monitoring for a specific outcome, such as seizures or stroke, we might want to monitor for a group of multiple outcomes that are of different levels of seriousness. For this, we assign relative weights to reflect seriousness of those events. For example, a single stroke could be considered equivalent to w episodes of seizure; stroke will then have a weight of w , whereas seizure will have a weight of 1. That is, stroke would be considered w times more serious than seizure. In such a case, the goal would be to monitor the vaccine or drug and generate a signal when the weighted sum of the adverse events in the exposed group is statistically different (larger or smaller) than the weighted sum of the adverse events in the unexposed group. The same elaboration arises when not only two but multiple (three or more) outcomes are studied, each having a different weight. Obviously, this rationale about weighted outcomes can be applied to problems other than postmarket surveillance or clinical trials.

The weighted sum of binomial end points is not a common test statistic for multiple outcomes sequential analysis. Most methods are based on adaptations of Pocock's and of O'Brien-Fleming's statistics.⁵⁻⁸ According to the work of Jennison and Turnbull,³ multiple-end point methods can be broadly divided in two categories. The first category consists of methods that run univariate sequential testing on each outcome separately but keeping the overall Type I error probability under the nominal level.⁸⁻¹² A simple but still important method is the group sequential Bonferroni procedure, which simply sets the significance level of each particular outcome equal to α/G , where α is the desired overall significance level, and G is a predefined number of group sequential tests.^{9,10} The Bonferroni approach is quite comprehensive in the sense of allowing for simultaneously analyzing heterogeneous outcomes and for any type of test statistic. However, the Bonferroni approach can lead to extremely conservative rejection criteria and, hence, relevant power losses.³

The second category of multiple-end point methods is the one of interest in this paper, formed by methods that summarize the cumulative outcomes in a single global test statistic. This is the case of the group sequential Hotelling test,¹³ the weighted combined index used

by Freedman et al,¹⁴ the proposal of Jennison and Turnbull,¹⁵ and the method for multivariate failure time data proposed by Cook and Farewell.¹⁶

Usually, multiple-end point sequential methods are used under the normal distribution assumption. That is, each outcome is assumed to follow a normal distribution, or even when outcomes are not assumed to be normally distributed, critical value calculations are still based on normal distribution approximations following classical asymptotic arguments, or related asymptotic results such as F-Snedecor or Chi-squared distributions. This is the approach of well-known methods proposed in other work.^{5–10,17–19}

Methods based on exact calculations, instead of asymptotic approximations, for multiple binomial end points have received little attention. Freedman et al¹⁴ discuss the usage of weights for monitoring long-term disease prevention trials with multiple binomial end points. In their work, the focus was on practical and heuristic aspects rather than formal statistical treatment concerning the overall significance level. In the work of Conaway and Petroni,²⁰ the joint distribution of two outcomes of interest was derived for group sequential phase II trials, but without use of weights for the outcomes. Hence, a formal derivation for exact sequential testing considering multiple weighted end points is lacking.

This paper derives the analytical framework for exact calculation of critical values for sequential analysis with multiple weighted binomial end points. All derivations are valid for the general case of one or more end points, and for different Bernoulli probabilities associated with each end point. Moreover, through usage of Type I error probability spending functions, one can apply the method for unpredictable sequential analysis. That is, the analyst does not need to know, in advance, the total number of observations at each look at the data. All these features are available through the openaccess **R** *Sequential* package, Version 3.0.²¹ The *Sequential* package can also be used to reproduce all numerical results shown in this paper, to perform calculations under arbitrary tuning parameters, and to analyze real data, as well. *Sequential* can also be used through its web interface (<http://www.sequentialanalysis.org>), where one can use *Sequential* functions without having to install or write code in R language.

This paper is organized in the following way: the next section describes the data structure considered in this work. Section 3 describes the proposed method. Section 4 derives critical value calculations. Section 5 offers an overview on the main Type I error probability spending functions. Section 6 gives an example of application for our case study. Section 7 discusses the choice of weights and its computational aspects. Section 8 closes this paper with some main conclusions.

2 | SEQUENTIAL ANALYSIS FREQUENCY AND DATA STRUCTURE

Sequential analysis methods can be broadly divided in three categories continuous, fixed group sizes, or unpredictable group sizes. Continuous methods are used when observations arrive one by one, and a test is performed for each single new observation. With fixed group sizes, the sample size at each test is specified in advance. In contrast, unpredictable group-size sequential methods are used when sample sizes cannot be established in advance.

Methods based on this last approach are comprehensive because they accommodate the first two categories and allow for mixed continuous-group sequential analysis. The method introduced in this paper is of this third, comprehensive, category.

Regarding the data structure, the proposed method is valid when the binomial distribution can be used to model each of K outcomes of interest. For example, it can be used for matched cohort designs, or placebo controlled two-arm trials as well, where two populations, say, exposure A and exposure B, are compared to each other. In this case, the binomial random variable is the number of individuals from group A presenting the j th end point among a total of $n_{i,j}$ occurrences of end point j at the i th test, with $j = 1, \dots, K$. In this case, the success probability, denoted by $p_{i,j}$, is the probability of having an individual from exposure A presenting outcome j at the i th test. Such an application motivates our terminology hereinafter.

For practical interpretation, we rewrite

$$p_{i,j} = 1 / (1 + z_i / R_j), \quad (1)$$

where z_i is the matching ratio between exposure A and exposure B for the i th test. For example, if for the i th test there are c exposures, A matched to each of the exposures B, then $z_i = c$. Usually, the matching ratio is constant for all tests. However, the comprehensive matching ratio setting, where z_i is allowed to vary with the test index i , can be useful for certain applications. For instance, in some cohort designs, because new patients can enter in the study during the analysis time, the ratio between the number of individuals in each group can change during the sequential surveillance.

For a second example, consider the postmarket drug/vaccine safety surveillance problem where, for the i th test, the number of adverse events from an exposure period window ($t_1 > 0$) is compared with the number of adverse events from a comparison period window ($t_2 > 0$). In this case, each adverse event of type j can be seen as a Bernoulli experiment with success probability equal to $p_{i,j} = R_j \times t_1 / (t_1 + t_2)$; hence, from (1), $z_i = t_2 / t_1$ for each $i = 1, \dots$. For instance, if the risk window is $t_1 = 2$ days long and the control window is $t_2 = 4$ days long when the first test ($i = 1$) is performed, and if $R_j = 1$ for each $j = 1, \dots, K$, then $p_{1,j} = 2 / (2 + 4) = 1/3$, and $z_1 = 2$.

The parameter R_j is the relative risk. Therefore, in essence, the relative risk is the parameter to be inferred. If the probability of having the j th end point from exposure A, instead of from exposure B, only depends on the matching ratio z_i , then $R_j = 1$. However, if there are effects on $p_{i,j}$ other than z_i , then $R_j \neq 1$.

Finally, the method is applicable in self-control studies, where, for the i th sequential test, the binomial random variable is the number of events associated to the outcome of type j in the risk window from a total of $n_{i,j}$ events in either the risk or control window.

3 | METHOD

Let $\vec{X}_1, \vec{X}_2, \dots$ denote a sequence of K -dimensional random vectors, where $\vec{X}_i = (X_{i,1}, \dots, X_{i,K})$ with $i = 1, 2, \dots$, and $X_{i,j} \sim \text{binomial}(n_{i,j}, p_{i,j})$, for $j = 1, \dots, K$. In addition, assume that $X_{i,j}$ is independent of $X_{i',j'}$ for each $i \neq i'$ or $j \neq j'$. The independence of $X_{i,j}$ in (i, j) holds when it is fair to assume that one would censor the patients at the time of any event such that, in the analysis, individuals would only ever experience a single event.

For the i th test, consider a test statistic, denoted by $S_{A,i}$, constructed as a weighted sum of cumulative outcomes, that is,

$$S_{A,i} = \sum_{g=1}^i (w_1 X_{g,1} + \dots + w_K X_{g,K}), \quad (2)$$

where w_j is the weight associated with the j th outcome. For simpler notation, we use S_i instead of $S_{A,i}$ when convenient and unambiguous.

In addition to simplicity, this metric, S_i , favors a friendly and intuitive interpretation in practice. For example, in a benefit-risk analysis, there may be some events that may be caused by the exposure of interest while there are other events that we hope are prevented by the exposure of interest. This classification does not matter for the analysis, and we will simply tally the number of events for each type of outcome, with their assigned weights. In theory, there can be both good or bad outcomes. Because we are primarily looking for bad outcomes, they will have a positive weight, while the good outcomes will have a negative weight.

3.1 | One-tailed testing

For testing if the relative risk is greater than 1 for at least one of the monitored outcomes, the statistical hypotheses are of the one-tailed, right-sided, form:

$$\begin{aligned} H_0: R_j &= 1 \text{ for each } j = 1, \dots, K, \\ H_1: R_j &> 1 \text{ for some } j \in \{1, \dots, K\}. \end{aligned} \quad (3)$$

With the hypotheses above, the signaling threshold for the i th test is an upper-limit critical value, denoted by cv_i . The critical values, $cv_1, cv_2, \dots$, are to be elicited so that the overall significance level, $\alpha \in (0, 1)$, is guaranteed. This is possible by considering a limit on the sample size. For this, let N denote the maximum length of surveillance given in the scale of the cumulative sample size. Therefore, the sequential testing continues without a decision while the cumulative sample size is smaller than N and S_i is smaller than cv_i . More precisely, at the i th test, if $\sum_{g=1}^i \sum_{j=1}^K n_{g,j} \geq N$ and $S_i > cv_i$, then stop the sequential monitoring and do not reject H_0 , but reject H_0 if $S_i > cv_i$. Otherwise, if $\sum_{g=1}^i \sum_{j=1}^K n_{g,j} < N$, then reject H_0 if

$$S_i > cv_i; \quad (4)$$

otherwise, proceed to a new test as soon as a new chunk of data is available.

If the hypotheses are of the left-sided form

$$\begin{aligned} H_0: R_j &= 1 \text{ for each } j = 1, \dots, K, \\ H_1: R_j &< 1 \text{ for some } j \in \{1, \dots, K\}, \end{aligned} \quad (5)$$

then the procedure is the same except for assigning a lower-limit critical value in place of an upper-limit one. That is, the null is rejected for the first i such that

$$S_i < cv_i. \quad (6)$$

3.2 | Two-tailed testing

For two-tailed testing, that is, with hypotheses of the form

$$\begin{aligned} H_0: R_j &= 1 \text{ for each } j = 1, \dots, K, \\ H_1: R_j &\neq 1 \text{ for some } j \in \{1, \dots, K\}, \end{aligned} \quad (7)$$

we consider both lower and upper signaling thresholds for each test i , which are denoted by cv_i^l and cv_i^u , respectively. In this case, the sequential testing is based on the following

procedure. At the i th test, if $\sum_{g=1}^i \sum_{j=1}^K n_{g|j} \geq N$, and $cv_i^l \leq S_i \leq cv_i^u$, then stop the sequential monitoring and do not reject H_0 , but reject H_0 if $S_i > cv_i^u$ or if $S_i < cv_i^l$. Otherwise, if $\sum_{g=1}^i \sum_{j=1}^K n_{g|j} < N$, then reject H_0 if

$$S_i > cv_i^u, \text{ or if } S_i < cv_i^l; \quad (8)$$

otherwise, proceed to a new test as soon as a new chunk of data is available.

Again, the critical values $cv_1^l, cv_2^l, \dots$, and $cv_1^u, cv_2^u, \dots$, are obtained in a way consistent with the overall significance level.

The notation for one-tailed and two-tailed tests can be unified to favor straightforward expressions. This is done by adopting lower and upper signaling thresholds notations for one-tailed testing, too. With right-sided testing, this is equivalent to putting lower-limit critical values that give zero probability to the event $\{S_i < cv_i^l\}$ for each i . Likewise, with left-sided testing, the upper-limit critical values, for each i , are simply established in a way that the event $\{S_i > cv_i^u\}$ has zero probability, too.

4 | CALCULATION OF CRITICAL VALUES

Consider the collection of end points, with their corresponding weights, to be fixed and that we condition the analysis on that collection. The random element is simply whether an outcome happened to occur with exposure A or exposure B.

The signaling thresholds are constructed in terms of a Type I error spending function. The Type I error spending function is a nondecreasing function, say, $F(t)$, taking values in the $(0, \alpha]$ interval, where $t \in \{1/N, 2/N, \dots, 1\}$ is a fraction of the maximum length of surveillance, N . The function $F(t)$ serves to dictate, in advance, the amount of Type I error probability to be spent at each look at the data. To ensure a statistically valid sequential analysis, the Type I error spending can only depend on the total number of events (exposure A + exposure B) at prior tests and the total number of events in the current test, but never in the number of events from exposure A or from exposure B.

Given the cumulative sample size of the i th test,

$$N_i = \sum_{g=1}^i \sum_{j=1}^K n_{gj}, \quad (9)$$

define the punctual Type I error spending, $f(N_i)$, as follows:

$$f(N_i) = F(N_i) - F(N_{i-1}), \text{ where } F(N_0) = 0. \quad (10)$$

For two-tailed testing, the assigned $f(N_i)$ spending can be arbitrarily divided in two fractions, say, $\alpha_i^l \geq 0$ and $\alpha_i^u \geq 0$, where $\alpha_i^l + \alpha_i^u = f(N_i)$. For one-tailed testing, either α_i^l or α_i^u must be settled equal to zero. With right-sided testing, we have $\alpha_i^l = 0$ and $\alpha_i^u = f(N_i)$. With left-sided testing, we have $\alpha_i^l = f(N_i)$ and $\alpha_i^u = 0$.

An important advantage of the present method is that it allows for unpredictability about the sample size of the last test. Although the maximum length of surveillance, N , is set *a priori*, it is more realistic to assume that the last chunk of data, the one leading to the attainment of N , will by chance, and probably, accumulate more than N observations instead of exactly N . However, this is by no means a problem. When the cumulative sample size reaches/exceeds N , which we denote by N_{end} , and for a given number g of previous tests, the level of the termination test is given by the remaining Type I error probability still available:

$$f(N_{end}) = \alpha - F(N_{i-1}). \quad (11)$$

Naturally, even the very first test can present a sample size exceeding N , but then the natural solution is $f(N_{end}) = \alpha$.

It merits mention that the test statistic S_i has a discrete support. Thus, let m_i denote the total number of different states of S_i , and define

$$\mathcal{A}_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,m_i}\}, \quad (12)$$

the set of ordered points from the support of S_i , ie, $s_{i,1} < s_{i,2} < \dots < s_{i,m_i}$ are the possible states of S_i . Note that $s_{i,1} = 0$ and $s_{i,m_i} = \sum_{g=1}^i \sum_{j=1}^K w_j n_{g,j}$. Analogously, for the partial sum

$$S_{i,j} = \sum_{g=1}^i (w_1 X_{g1} + \dots + w_K X_{gj}), j = 1, \dots, K, \quad (13)$$

take the associated set of ordered states:

$$\mathcal{A}_{i,j} = \{s_{i,j,1}, s_{i,j,2}, \dots, s_{i,j,m_{i,j}}\}. \quad (14)$$

For the very first test, ie, $i = 1$, the lower- and upper-limit critical values are given by

$$cv_1^l = \max_r \{s_{1,r} : Pr[S_1 < s_{1,r} | H_0] \leq \alpha_1^l\}, \quad (15)$$

$$cv_1^u = \min_r \{s_{1,r} : Pr[S_1 > s_{1,r} | H_0] \leq \alpha_1^u\}. \quad (16)$$

The probabilities in the constraints above are calculated as the following: for a point $a \in \mathcal{A}_1$, we have

$$Pr[S_1 > a | H_0] = \sum_{\{r^*: s_{1,r^*} > a\}} Pr[S_1 = s_{1,r^*} | H_0] = \sum_{\{r^*: s_{1,r^*} > a\}} P_1(s_{1,r^*}). \quad (17)$$

For calculating the terms of the sum in (17), let s denote a generic point in $\mathcal{A}_1$, then

$$\begin{aligned} P_1(s) &= Pr[S_1 = s | H_0] = Pr[S_{1,1} = s | H_0] + Pr[S_{1,1} \neq s, S_{1,2} = s | H_0] + \dots \\ &+ Pr[\cap_{j=1}^{K-1} \{S_{1,j} \neq s\}, S_{1,K} = s | H_0] \\ &= Pr[X_{1,1} = s/w_1 | H_0] + \sum_{\substack{0 \leq x \leq n_{1,1} \\ w_1 x \neq s}} Pr[X_{1,2} = (s - w_1 x)/w_2 | H_0] Pr[X_{1,1} = x | H_0] + \dots \\ &+ \sum_{x \in A_1} Pr[X_{1,K} = \frac{s - w_{K-1}x}{w_K} | H_0] Pr[X_{1,K-1} = x | H_0] \prod_{j=1}^{K-1} \{S_{1,j} \neq s\}, \end{aligned}$$

where $A_1 = \{0 \leq x \leq n_{1,K-1} : w_{K-1}x \leq s\}$.

For the second test, ie, $i = 2$, only the alive states of S_1 , points of $\mathcal{A}_1$ from cv_1^l to cv_1^u , are needed to calculate critical values. For $i > 2$, define the partial weighted sum

$$Y_i = \sum_{j=1}^K w_j X_{i,j}. \quad (18)$$

Thus, for a fixed point y of the Y_i support, the probability function of Y_i under H_0 , $P_{i,Y}$, is given by

$$\begin{aligned} P_{i,Y} &= Pr[Y_i = y|H_0] = Pr[Y_{i,1} = y|H_0] + Pr[Y_{i,1} \neq y, Y_{i,2} = y|H_0] + \dots \\ &+ Pr\left[\cap_{j=1}^{K-1} \{Y_{i,j} \neq y\}, Y_{i,K} = y|H_0\right] \\ &= Pr[X_{i,1} = y/w_1|H_0] + \sum_{\substack{0 \leq x \leq n_{i,1} \\ w_1 x \neq y}} Pr[X_{i,2} = (y - w_1 x)/w_2|H_0] Pr \\ &\left[X_{i,1} = x|H_0\right] + \dots + \sum_{x \in A_i} Pr\left[X_{i,K} = \frac{y - w_{K-1}x}{w_K}|H_0\right] Pr \\ &\left[X_{i,K-1} = x|H_0, \cap_{j=1}^{K-1} \{Y_{i,j} \neq y\}\right], \end{aligned} \quad (19)$$

where $A_i = \{0 \leq x \leq n_{i,K-1} : w_{K-1}x \leq y\}$.

Thus, the lower and upper limits for the second test are calculated as follows:

$$cv_2^l = \max_r \{s_{2,r} : Pr[S_2 < s_{2,r}, cv_1^l \leq S_1 \leq cv_1^u|H_0] \leq \alpha_2^l\}, \quad (20)$$

$$cv_2^u = \min_r \{s_{2,r} : Pr[S_2 > s_{2,r}, cv_1^l \leq S_1 \leq cv_1^u|H_0] \leq \alpha_2^u\}, \quad (21)$$

where, for a generic point $s \in \mathcal{A}_i$,

$$\begin{aligned} Pr[S_2 = s, cv_1^l \leq S_1 \leq cv_1^u|H_0] &= \sum_{\substack{s_1 \in \mathcal{A}_1 \\ cv_1^l \leq s_1 \leq cv_1^u}} Pr[Y_2 = s - s_1|H_0] Pr \\ [S_1 = s_1|H_0]. \end{aligned} \quad (22)$$

Generalizing, the probabilities for the next round will then be calculated conditioned on the terminal versus alive states and their corresponding probabilities. For a straightforward notation of such conditioned probabilities, define

$$S_i^{(-s)} = S_i | \cap_{g=1}^{i-1} \{S_g \neq s\}, i = 2, \dots, \quad (23)$$

which represents the conditional random variable S_i given that state s was not observed with previous tests.

Finally, the lower and upper limits for the i th test, with $i > 1$, are given by

$$cv_i^l = \max_r \{s_{i,r} : Pr[S_i < s_{i,r}, \cap_{g=1}^{i-1} \{cv_g^l \leq S_g \leq cv_g^u\} | H_0] \leq \alpha_i^l\}, \quad (24)$$

$$cv_i^u = \min_r \{s_{i,r} : Pr[S_i > s_{i,r}, \cap_{g=1}^{i-1} \{cv_g^l \leq S_g \leq cv_g^u\} | H_0] \leq \alpha_i^u\}, \quad (25)$$

where, for a generic point $s \in \mathcal{A}_i$,

$$Pr[S_i = s, \cap_{g=1}^{i-1} \{cv_g^l \leq S_g \leq cv_g^u\} | H_0] = \sum_{a \in \mathcal{A}_{i-1}^*} Pr[Y_i = s - a | H_0] Pr[S_{i-1}^{(-s)} = a | H_0], \quad (26)$$

where $\mathcal{A}_{i-1}^*$ is the set of alive states from test $(i-1)$, that is,

$$\mathcal{A}_{i-1}^* = \{s^* \in \mathcal{A}_{i-1} : cv_{i-1}^l < s^* < cv_{i-1}^u\}. \quad (27)$$

Note that the probabilities in the right-hand side of (26) are simple to compute because they only demand running the probabilities given in (19) and combining with the probabilities $Pr[S_{i-1}^{(-s)} = a | H_0]$ already available from the $(i-1)$ th test.

For user convenience, and to compute critical values under different tuning parametrization, all these expressions are implemented and available under open access with the **R** *Sequential* package.

5 | CHOOSING THE TYPE I ERROR SPENDING SHAPE

The Type I error probability spending approach dates back to the early 1980s.^{22–27} As a rule, these traditional proposals are motivated by an attempt to approximate Pocock's and O'Brien and Fleming's sequential methods. A more flexible function, which embraces Pocock's and O'Brien and Fleming's methods just as particular cases, is the so-called “power-type” shape

$$F(t) = \alpha \times t^\rho, \quad (28)$$

with $\rho > 0$ and $t \in \{1/N, 2/N, \dots, 1\}$.

As emphasized by the work of Kim and Demets²⁴ and Jennison and Turnbull,^{25,26} for specific values of ρ , the function in (28) offers a good approximation for Pocock's and O'Brien and Fleming's sequential testing. However, it can lead to alternative shapes as a function of ρ . For univariate binomial stochastic processes, the work of Silva²⁸ suggests to

choose ρ values around 0.5 for minimization of the expected time to signal. However, if the expected sample size is the most important statistical performance measure, then Jennison and Turnbull³ suggest adoption of ρ values around 1.5 or 2.

6 | REAL DATA EXAMPLE

We illustrate the proposed method by mimicking a statistical sequential analysis using real data containing five different outcomes. This case study examined increased risk of adverse events in 9340 patients who initiated denosumab compared to 9340 propensity score matched patients who initiated bisphosphonates for treatment of osteoporosis. Hip and pelvis fracture ($w_1 = 0.05$), forearm fracture ($w_2 = 0.08$), humerus fracture ($w_3 = 0.09$), serious infection ($w_4 = 0.11$), and pneumonia ($w_5 = 0.30$) were five outcomes that were considered as end points in this analysis. Monitoring for adverse events took place monthly over a 12-month period.

The matching ratio of each outcome is $z = 1$, and the hypotheses are of the form

$$\begin{aligned} H_0 : R_j &= 1 \text{ for each } j = 1, \dots, 5, \\ H_1 : R_j &\neq 1 \text{ for at least one } j \in \{1, 2, 3, 4, 5\}. \end{aligned} \quad (29)$$

For this example, we adopted the Type I error spending of the power-type with $\rho = 0.5$. A maximum length of surveillance of $N = 1000$ was selected, and the overall significance level used was $\alpha = 0.05$. For a more intuitive interpretation of results, the test statistic is rewritten in terms of the ratio of exposure A to exposure B weighted sums, denoted by $S_{i,A}/S_{i,B}$. Note that there is a one-to-one correspondence between this new metric and $S_{i,A}$, so the test is in essence the same.

Table 1 presents the cumulative sample size, test by test, for each of the five outcomes, in columns 2 to 6. This table also shows the observed test statistic, column 7, the lower and upper critical values in columns 8 and 9, respectively, and the target Type I error spending, column 10. Due to the discrete nature of the test statistic, the target Type I error spending is rarely exactly achieved, but, instead, the actual spending, shown in column 11, is always a number smaller than or equal to the target value.

Because $z = 1$, under H_0 , the ratio $S_{i,A}/S_{i,B}$ tends to be around 1 for each test, and this is what we observe in most of the cases shown in Table 1. In some tests, the ratio differed a little bit from 1, such as in test 2 ($S_{2,A}/S_{2,B} = 1.68$) and in test 8 ($S_{8,A}/S_{8,B} = 0.55$), but not enough to reject H_0 . Another thing to note is how the distance between the lower and upper critical values becomes narrower and narrower while the surveillance advances, which is due to the gain in information while sample size increases in the course of the surveillance.

Hip and pelvis fracture ($w_1 = 0.05$), forearm fracture ($w_2 = 0.08$), humerus fracture ($w_3 = 0.09$), serious infection ($w_4 = 0.11$), and pneumonia ($w_5 = 0.30$)

7 | CHOICE OF WEIGHTS

The choice of weights reflects the practical impact of each adverse event on patients' health or quality of life. The weights that have been used in our case study were based on health-related quality of life measures associated with each of the five adverse events that have been reported in the published literature. The weights indicate the impact of each end point on patients' quality-adjusted life years (QALYs).

7.1 | The impact of weights on time to signal

Different results are observed if different weights are used for the same data set, as it should. To illustrate, consider a new data set, in this case, containing only two outcomes, myocardial infarction ($w_1 = 2.2$), and major bleeding ($w_2 = 0.04$), obtained from a study comparing risk of myocardial infarction and gastrointestinal bleeding among 23 395 patients who initiated rofecoxib and 23 395 propensity score-matched patients who initiated non-selective non-steroidal anti-inflammatory drugs (ns-NSAIDs), which leads to a matching ratio of $z = 1$ for both outcomes. Monthly data were collected during a 20-month follow-up period. The hypotheses in this case are

$$\begin{aligned} H_0 : R_1 = R_2 = 1, \\ H_1 : R_1 \neq 1 \text{ or } R_2 \neq 1, \end{aligned} \quad (30)$$

and consider the following tuning parametrization: $\rho = 0.5$, $N = 1000$, and $\alpha = 0.05$. Table 2 presents the cumulative data and results of the sequential analysis based on this new data set. We see that a signal for rejecting H_0 occurs at the 15th test. However, this signal would occur at a different time if different weights were used. Moreover, even the final test decision could change depending on the weight choice. With only two outcomes, what really matters is the weight ratio, $w = w_2/w_1$. In the present example of rofecoxib vs NSAID, we could use $w_1 = 1$ and $w_2 = 55$ because it leads to the same weight ratio, ie, $w = 2.2/0.04 = 55$, and the results of the sequential analysis would be exactly the same.

To assess the behavior of stopping time as a function of the weight ratio for this specific data set, we applied the sequential procedure to this same data set several times, each for a different value for weight $w = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100$. For the first five values of w , 1, 1.5, 2, 2.5, and 3, the null hypothesis was not rejected; hence, the monitoring would have continued until test 21 in these cases. For the other values, $w = 3.5, 4, \dots, 100$, the stopping time for H_0 rejection varied among tests 15, 16, and 17, with the result being constant at test 15 for each $w \geq 5.5$.

Figure 1 shows the test moment when the surveillance was stopped for values in the range $[2, 2.5, \dots, 7]$. The practical implications of such variation would depend on how many additional people entered in the study between tests 15 and 17. However, in the situations $w = 1, 1.5, 2, 2.5, 3$, where H_0 would not be rejected, the fact that the conclusion changes, from rejection to nonrejection, suggests that the choice of weights can directly impact the signal generation and rejection of null hypothesis.

The nonmonotonicity observed between times 15 and 17 is due to the discrete nature of the test statistic because the actual alpha spending varies with the cardinality of the sample space. Weights having more decimal places tend to lead to sample spaces that are more diluted, slightly influencing time to signal and power.

The practical interpretation of the weights as measures of seriousness of each outcome relative to the others must be emphasized. In this two-outcome example, if the weight ratio is close to zero, it means that only gastrointestinal bleeding is important in the analysis; hence, myocardial infarction would have almost zero influence in the test decision. Otherwise, if weight ratio goes to ∞ , then only myocardial infarction is taken in account during the analysis, while gastrointestinal bleeding could not affect the test decision. This interpretation is applicable for analyses in general.

7.2 | Weights versus computational complexity

With multiple weights, the computational complexity will be larger compared to the special case where all end points have the same weight. Likewise, non-integer weights can lead to more time-consuming computations than integer weights. To see this fact, suppose that we have two different types of outcomes, O_1 with weight 1, and O_2 with weight 2. Suppose that, in total, there were 20 observations from O_1 and 15 from O_2 . Hence, there is a total of $(1 \times 20) + (2 \times 15) + 1 = 51$ possible values of $S_{N_{end}}$. If we had uniform weights of 1 for both outcomes, then the state space would be in the range $[0,35]$, with 36 possible values, so there is not a big difference between these two choices for the weights. Of course, if we had weights 1 and 5, the range would be $[0,95]$, with 96 possible values, which is just more than double the number of values. Hence, the rate of increase in possible values is slow when the weights are all integer values and not too far apart. If instead we had weights 1 and $\sqrt{2}$, then the state space would be one-dimensional, with $(1 + 20) \times (1 + 15) = 336$ possible states. This number will quickly proliferate as data arrive from new patients. For instance, if we had 50 individuals from O_1 and 45 from O_2 , with weights 1 and $\sqrt{2}$, respectively, then we would have $(1 + 50)(1 + 45) = 2346$ possible states. However, if we use 2 in place of $\sqrt{2}$ for the weight of O_2 , then the cardinality of the state space reduces from 2346 to 146. Therefore, in practice, it would be preferable to select weights, when possible, that are integer values and within a $[1,10]$ or $[1,100]$ range, to make sure that computing times are kept reasonable.

8 | CONCLUDING REMARKS

The proposed test statistic combines formality with intuition and practicality. While severe outcomes are objectively prioritized over less severe outcomes, the method still enables control of the overall performance in terms of Type I error probability and time to signal.

The method performs reasonably for early detection of differences in net benefits. As we propose exact calculations of critical values, a key point is the execution time of the proposed method. As already emphasized, the weight profile is a determinant of complexity and, hence, computing time. In the context of the data analysis examples in this paper, we showed that the method may be applicable for most real-data scenarios. For example, using a regular PC (Windows 10, Intel(R) Core(TM) i7-6700HQ CPU, 2.60 GHz, 32.0 GB of RAM), the method ran in less than 10 seconds for each test of Tables 1 and 2.

Alternative to the exact signaling threshold calculation proposed in this paper, one could use a normal distribution approximation consistent with the conventional sequential analysis theory. While the asymptotic theory can potentially improve the computation time of the proposed method, it can also lead to inflated test sizes. Therefore, a challenge for future studies is to evaluate the pros and cons of using asymptotic theory instead of exact calculations for finding signaling thresholds on the scale of the weighted sum of binomial end points.

ACKNOWLEDGEMENTS

This research was funded by the National Institute of General Medical Sciences, USA, under grant #R01GM108999. Additional support was provided by Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Minas Gerais, Brazil (FAPEMIG). The authors are grateful to Kristina Noemi Stefanini for editing support and for the important improvements suggested by the referees.

Funding information

National Institute of General Medical Sciences, Grant/Award Number: RO1GM108999; Fundação de Amparo à Pesquisa do Estado de Minas Gerais

REFERENCES

1. Wald A Sequential tests of statistical hypotheses. *Ann Math Stat.* 1945;16:117–186.
2. Armitage P Sequential methods in clinical trials. *Am J Public Health.* 1958;48(10):1395–1402.
3. Jennison C, Turnbull BW. *Group Sequential Methods With Applications to Clinical Trials.* London, UK: Chapman and Hall/CRC; 2000.
4. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the vaccine safety datalink project. *Pediatrics.* 2011;127(1):54–64.
5. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–1087. [PubMed: 6534410]
6. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics.* 1987;43:487–498. [PubMed: 3663814]
7. Tang D, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics.* 1993;49:23–30. [PubMed: 8513104]
8. Ye Y, Li A, Liu L, Yao B. A group sequential Holm procedure with multiple primary endpoints. *Statist Med.* 2013;32(7):1112–1124.
9. Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. *Contr Clin Trials.* 1997;18:557–560.
10. Prentice RL. Discussion: on the role and analysis of secondary endpoints in clinical trials. *Contr Clin Trials.* 1997;18:561–567.
11. Kosorok MR, Shi Y, DeMets D. Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics.* 2004;60:134–145. [PubMed: 15032783]
12. Hamasaki T, Asakura K, Evans SR, Sugimoto T, Sozu T. Group-sequential strategies in clinical trials with multiple co-primary outcomes. *Stat Biopharm Res.* 2015;7(1):36–54. [PubMed: 25844122]
13. Jennison C, Turnbull BW. Exact calculations for sequential t , χ^2 and F -tests. *Biometrika.* 1991;78:133–141.
14. Freedman L, Anderson G, Kipnis V, et al. Approaches to monitoring the results long-term disease prevention trials: examples from the Women's Health Initiative. *Control Clin Trials.* 1996;17:509–525. [PubMed: 8974210]
15. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics.* 1993;49:741–752. [PubMed: 8241370]

16. Cook RJ, Farewell VT. Incorporating surrogate endpoints into group sequential trials. *Biometrical Journal*. 1996;38(1):119–130.
17. Zhang J, Quan H, Ng J, Stepanavage M. Some statistical methods for multiple endpoints in clinical trials. *Control Clin Trials*. 1997;18:204–221. [PubMed: 9204221]
18. Liu A, Hall WJ. Unbiased estimation following group sequential test. *Biometrika*. 1999;86:71–78.
19. Hung HMJ, Wang S-J, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *J Biopharm Stat*. 2007;17:1201–1210. [PubMed: 18027226]
20. Conaway MR, Petroni GR. Bivariate sequential designs for phase ii trials. *Biometrics*. 1995;51(2): 656–664. [PubMed: 7662852]
21. Silva IR, Kulldorff M. Sequential-package. Contributed packages. Vienna, Austria: R Foundation for Statistical Computing; 2018 <http://www.R-project.org>
22. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3): 659–663.
23. Gordon Lan KK, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659–663.
24. 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.
25. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach (with discussion). *J R Stat Soc B*. 1989;51:305–334.
26. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science*. 1990;5:299–317.
27. Stallard N, Todd S. Exact sequential tests for single samples of discrete responses using spending functions. *Statist Med*. 2000;19:3051–3064.
28. Silva IR. Type error 1 probability spending for post-market drug and vaccine safety surveillance with binomial data. *Statist Med*. 2018;37(1):107–118.

FIGURE 1.

Stopping time per weight choice in the range $[2, 2.5, \dots, 7]$ based on the rofecoxib vs NSAID comparison. These results are obtained under tuning parametrization of $\rho = 0.5$, $N = 1000$, and $\alpha = 0.05$

TABLE 1

Sequential analysis results for the data of treatment of osteoporosis. The outcomes are hip and pelvis fracture ($w_1 = 0.05$), forearm fracture ($w_2 = 0.08$), humerus fracture ($w_3 = 0.09$), serious infection ($w_4 = 0.11$), and pneumonia ($w_5 = 0.30$). Maximum length of surveillance was $N = 1000$. The overall significance level used was $\alpha = 0.05$, with power-type alpha spending ($\rho = 0.5$). The critical values are shown in the scale of the test statistic given by the ratio between weighted sums from exposure A to exposure B populations ($T = S_{i,A}/S_{i,B}$)

Cumulative data per outcome					Critical value		Alpha spending		
H-p frac.	F. frac.	H. frac.	Infec.	Pneum.	T	Lower	Upper	Target	Actual
0	0	0	1	0	inf	na	na	0.0016	0
1	2	0	5	3	1.68	0.05	19.75	0.0052	0.0039
2	7	0	6	7	1.22	0.17	5.84	0.0074	0.0069
3	9	1	10	10	0.94	0.25	3.96	0.0091	0.0088
6	13	1	14	14	0.65	0.34	2.98	0.0110	0.0109
10	17	1	20	16	0.69	0.39	2.54	0.0126	0.0126
12	21	1	26	21	0.58	0.44	2.28	0.0142	0.0141
15	29	1	28	27	0.55	0.48	2.09	0.0158	0.0157
20	32	2	37	32	0.70	0.52	1.92	0.0175	0.0174
23	44	4	48	43	0.76	0.58	1.73	0.0201	0.0201
26	57	6	59	55	0.66	0.62	1.62	0.0225	0.0225
32	72	8	79	60	0.74	0.66	1.52	0.0250	0.0250

TABLE 2

Sequential analysis results for the rofecoxib vs ns-NSAID comparison. This application has two different outcomes, myocardial infarction ($w_1 = 2.2$), and gastrointestinal bleeding ($w_2 = 0.04$). The tuning parameters used are $N = 1000$, $\alpha = 0.05$, and power-type Type I error spending with $\rho = 0.5$. The critical values are shown on the scale of the ratio between weighted sums from exposure A to exposure B populations ($T = S_{i,A}/S_{i,B}$)

Test	Cumulative data($\hat{R}$)			Critical value		Alpha spending	
	Gastro. bleed.	Myoc. inf.	T	Lower	Upper	Target	Actual
1	22 (0.85)	24 (1.20)	0.85	0.34	2.97	0.0107	0.0089
2	45 (1.03)	59 (1.14)	1.04	0.48	2.07	0.0161	0.0161
3	59 (1.20)	90 (1.11)	1.19	0.55	1.80	0.0193	0.0191
4	74 (1.38)	133 (1.11)	1.37	0.62	1.60	0.0227	0.0227
5	92 (1.36)	170 (0.88)	1.36	0.65	1.56	0.0256	0.0255
6	108 (1.28)	196 (0.86)	1.27	0.68	0.148	0.0276	0.0275
7	122 (1.25)	225 (0.94)	1.25	0.69	1.44	0.0295	0.0294
8	135 (1.27)	252 (0.99)	1.27	0.71	1.42	0.0311	0.0310
9	146 (1.23)	286 (1.00)	1.27	0.72	1.38	0.0329	0.0328
10	156 (1.25)	319 (1.05)	1.24	0.74	1.36	0.0345	0.0344
11	168 (1.24)	340 (1.05)	1.24	0.74	1.34	0.0356	0.0356
12	177 (1.23)	353 (1.08)	1.23	0.75	1.34	0.0364	0.0364
13	181 (1.27)	368 (1.06)	1.27	0.65	1.34	0.0370	0.0370
14	189 (1.30)	384 (1.03)	1.30	0.76	1.32	0.0378	0.0378
15	199 (1.32)	395 (1.03)	1.32	0.76	1.32	0.0385	0.0385
16	206 (1.32)	408 (1.04)	1.32	na	na	na	na
17	218 (1.31)	428 (1.02)	1.31	na	na	na	na
18	236 (1.27)	457 (0.98)	1.27	na	na	na	na
19	247 (1.24)	478 (0.96)	1.24	na	na	na	na
20	258 (1.24)	485 (0.96)	1.23	na	na	na	na
21	265 (1.23)	491 (0.92)	1.23	na	na	na	na