Received: 28 December 2020 | Revised: 25 April 2021 | Accepted: 27 May 2021

DOI: 10.1002/sim.9094

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Exact sequential test for clinical trials and post-market drug and vaccine safety surveillance with Poisson and binary data

Ivair R. Silva[1] | **Judith Maro**[2] | **Martin Kulldorff**[3]

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

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA

[3]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts, USA

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

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 301391/2019-0; National Institute of General Medical Sciences, Grant/Award Number: RO1GM108999

Abstract

In sequential analysis, hypothesis testing is performed repeatedly in a prospective manner as data accrue over time to quickly arrive at an accurate conclusion or decision. In this tutorial paper, detailed explanations are given for both designing and operating sequential testing. We describe the calculation of exact thresholds for stopping or signaling, statistical power, expected time to signal, and expected sample sizes for sequential analysis with Poisson and binary type data. The calculations are run using the package *Sequential*, constructed in *R* language. Real data examples are inspired on clinical trials practice, such as the current efforts to develop treatments to face the COVID-19 pandemic, and the comparison of treatments of osteoporosis. In addition, we mimic the monitoring of adverse events following influenza vaccination and Pediarix vaccination.

KEYWORDS

adaptive design, alpha spending, continuous monitoring, group sequential

1 | INTRODUCTION

The regular practice for hypothesis testing is to conduct a single analysis based on a single data sample. Alternatively, with sequential hypothesis testing, one prospectively performs multiple hypothesis tests. Each test is performed when new data—that is, new observations—arrive, while guaranteeing the overall significance level by the end of the analysis.

The sequential approach is essential for many applications when it is urgent to reach a conclusion or decision, such as in post-market medical product safety surveillance, or when it is unethical to continue a clinical trial when there is clear evidence of benefit or harm affecting one group.

Usually, the sequential analysis is based on monitoring a test statistic in comparison to a lower and an upper signaling threshold at each of the multiples sequential looks at the data. The sequential analysis is stopped as soon as the test statistic crosses one of the thresholds. Classical methods for sequential analysis are Wald's sequential probability ratio test (SPRT),[1,2] Pocock's test,[3] O'Brien-Fleming's test,[4] and Wang-Tsiatis' method.[5] For post-market safety surveillance, recent methods are the maximized sequential probability ratio test (MaxSPRT),[6] and the conditional MaxSPRT (CMaxSPRT).[7]

Instead of thresholds given in the scale of a test statistic, sequential testing can be based on alpha spending functions.[8] The alpha spending function is a non-decreasing function taking values in the $[0, \alpha]$ interval, where α is the significance level. Therefore, the alpha spending function dictates, in advance, the amount of Type I error probability to be spent at

each of the multiple tests. This way, as an adaptive design, no matter the frequency at which the chunks of data arrive, or the cumulative sample size available at each test, the alpha spending function enables to find the thresholds accordingly.

Statistical performance evaluations and critical values calculations for sequential testing are usually obtained through asymptotic theory and/or normal distribution approximations.[8] Recent developments have shown that exact calculations are possible for many applications.[9-11] This is the approach of the present tutorial, which is devoted to offer practical examples on designing and conducting sequential hypothesis testing with binary and Poisson data. For this, we present step-by-step calculations accompanied with explanations on the underlying theory and proper interpretations of illustrative data analysis results.

The calculations for the illustrative examples are run with the *R Sequential* package.[12] *R Sequential* is an easy-to-use tool for both the design and the practical implementation of sequential analysis. All calculations are exact, based on iterative numerical procedures, rather than using asymptotic theory, computer simulations, or normal distribution approximations.

For either Poisson or binary 0/1 data, this tutorial covers the following topics:

- **Data frequency:** The number of new observations in each new data arrival does not have to be known a priori. We show how to perform sequential testing for continuous, group or mixed group-continuous sequential analysis with unpredictable data frequency.
- **Probability model:** For Poisson data, the expected counts may either be known or estimated from historical data with some uncertainty in the estimates. The binary model can be used for different studies where a dichotomous endpoint is monitored, including placebo-controlled two-arm clinical trials, self-controlled designs, and matched cohort designs.
- **Alternative hypothesis:** Unlike Wald's SPRT, here we use a composite alternative hypothesis. Both one and two-tailed tests are supported.
- **Signaling thresholds:** Signaling thresholds are calculated using Pocock's statistic, O'Brien-Fleming's statistic, Wang-Tsiatis statistic, and Wald SPRT statistic, as well as any user specified alpha spending function. Conversely, we give examples on how to calculate the alpha spending implied by any of these test statistics.
- **Optimal alpha spending function:** For a user-specified alpha level, relative risk and statistical power, we exemplify the usage of alpha spending functions that minimizes expected time to signal or expected sample size. This is done for both with or without an added requirement on the maximum length of surveillance. The optimal solution is obtained using the method proposed by Reference 13.
- **Statistical performance metrics:** Exact calculations are illustrated for statistical power, expected time of surveillance given that the null hypothesis is rejected, expected time of surveillance, and maximum maximum sample size. The latter three are calculated in the unit of sample size or number of events.

The content of this tutorial is organized in the following way: Next section presents definitions, notation and theoretical background that form the basis of this tutorial. Section 3 discusses planning and setting up sequential analysis testing according to pre-experimental statistical performance measures such as maximum sample size, statistical power, expected time to signal and expected length of surveillance. Sequential analysis designing is discussed in light of well-known test statistics (statistical measures of evidences) such as Wald's, Pocock's, O'Brien-Fleming's, and Wang-Tsiatis' tests. In addition, Section 3 shows how to calculate and interpret flat and time-variable signaling thresholds using the different test statistic scales. There, we also explain how to switch the calculations from these classical test statistic scales to the alpha spending scale, and vice-versa. Section 4 presents four examples of sequential testing for the actual analysis in practice. The first example is based on simulated data with structure inspired by the recent placebo-controlled two-arm trials on treatments for COVID-19 patients reported by References 14 and 15. The other three examples are based on real data for: (i) comparison of two treatments of osteoporosis by weighting five different adverse events in a propensity score matched cohort study, (ii) surveillance of neurological adverse events after Pediarix vaccination, and (iii) monitoring seizures after concomitant vaccination of inactivated influenza vaccine with 13-valent pneumococcal conjugate vaccine. Section 5 contains the last comments and further software considerations.

2 | EXACT SEQUENTIAL TESTING BACKGROUND

Let X_t denote a discrete stochastic process indexed by continuous or discrete time. In essence, in this article X_t is the cumulative number of events up to time t. The distributions of interest in the present work involve: (i) the cases where t

is a positive integer, here also denoted by n, and X_t is the sum of t Bernoulli outcomes with the success probability $p(RR)$, $RR > 0$, (ii) the cases where X_t is a Poisson stochastic process with parameter $RR\mu_t$, $RR > 0$, where μ_t is a known baseline rate under the null hypothesis, and (iii) X_t is a Poisson stochastic process with parameter $RR\mu_t$, μ_t unknown. Therefore, the parameter of interest is the relative risk (RR), and the underlying theory and results presented along this tutorial is applicable for any of the four pairs of hypotheses:

$$H_0 : RR \leq RR_0 \text{ against } H_1 : RR > RR_0, \tag{1}$$

$$H_0 : RR \geq RR_0 \text{ against } H_1 : RR < RR_0, \tag{2}$$

$$H_0 : RR = RR_0 \text{ against } H_1 : RR \neq RR_0, \tag{3}$$

$$H_0 : RR_{0,l} \leq RR \leq RR_{0,u} \text{ against } H_1 : RR < RR_{0,l} \text{ or } RR > RR_{0,u}, \tag{4}$$

where RR_0, $RR_{0,l}$, and $RR_{0,u}$ are specified by the user in advance. For the formats (1) to (3), a common choice is $RR_0 = 1$. Applications using $RR_0 > 1$ for format (1) are relevant too. See, for example, the public master protocol for COVID-19 vaccine active surveillance, by the U.S Food and Drug Administration (FDA).[16] In that protocol, testing margin was settled through RR_0 values of 1.25, 1.5, and 2.5, depending on the characteristics of each database.

The subset of the parameter space, implied by these hypotheses options under H_0, shall be denoted simply by Θ_0, where:

$$\Theta_0 = \begin{cases} (0,\ RR_0], \text{ for the format in (1)}, \\ (RR_0,\ \infty), \text{ for the format in (2)}, \\ \{RR_0\}, \text{ for the format in (3)}, \\ \left[RR_{0,l},\ RR_{0,u}\right], \text{ for the format in (4)}. \end{cases} \tag{5}$$

The relation of RR with the parametrization of the Bernoulli and Poisson probability models shall be further discussed in Sections 2.2,2.4, and 2.5.

Conventionally, sequential testing methods consist of comparing a test statistic, say $W(X_t)$, against pre-established signaling thresholds. The sequential testing concludes as soon as the test statistic reaches one of the thresholds. The thresholds are usually flat, such as with SPRT, MaxSPRT, Pocock's score test, and O'Brien & Fleming test, but time-variable thresholds are used too, like those elicited with the alpha spending approach. In either case, group and continuous sequential testing designs can be defined in general as following.

Definition 1 (Group Sequential Analysis). For two sets of constants, $a_1 \leq a_2 \leq \ldots, a_G$, and $b_1 \leq b_2 \leq \ldots, b_G$, given in the scale of a test statistic, $W(X_t)$, and a sequence $\{t_i\}_{i=1}^{G}$ of times taken from the set $\{1, \ldots, N\}$, where $t_G = T$ is the maximum length of surveillance, also denoted by N in the Bernoulli case, a group sequential analysis design is any procedure that ends the analysis for: (i) rejecting the null hypothesis if $W(X_{t_i}) \geq b_i$ and $W(X_{t_j}) > a_j$ for each $j \leq i$, or (ii) failing to reject the null hypothesis if $W(X_{t_i}) \leq a_i$, or $i = G$, and $W(X_{t_j}) < b_j$ for each $j \leq i$.

The rationale behind group sequential analysis is to perform tests only at pre-defined times, where, if the maximum sample size $t_G = T$ is reached without a decision, then the maximum number of tests equals G. Note that each test can have one or many events. In contrast, in continuous sequential design, outcomes arrive one-by-one, and a hypothesis test is performed after the arrival of each observation.

Definition 2 (Continuous Sequential Analysis). Let $a(t)$ and $b(t)$ denote real-valued functions such that $a(t) < b(t)$ for each $t \in (0,\ T]$, where T is also denoted by N in the Bernoulli case, and let $W(X_t)$ denote a test statistic. A continuous sequential analysis design is any procedure that ends the analysis for: (i) rejecting the null hypothesis if $W(X_t) \geq b(t)$ and $X_l > a(l)$ for each $l < t$, or (ii) failing to reject the null hypothesis if $W(X_t) \leq a(t)$, or $t = T$, and $W(X_l) < b(l)$ for each $l < t$.

The classical Wald's SPRT is an example where both, upper and lower boundaries are used as in Definitions 1 and 2. In contrast, if the analysis is allowed to be concluded before time T only under evidences against H_0, such as with MaxSPRT and CMaxSPRT, then only b_i and $b(t)$ are used in the definitions above. Another possible situation, which has not received much attention in the literature, is that where only a_i and $a(t)$ are used. In such cases, the analysis is to stop before time T only when evidences in favor of H_0 are found.

Definitions 1 and 2 are the so-called 'truncated' designs, and this is so because they have a vertical threshold, T. In practice, T represents the maximum sample size (or maximum length of surveillance described in number of data observations that have accrued) for ending the analysis without rejecting the null hypothesis. Procedures without vertical thresholds are sometimes called "open-ended designs."

For binary and Poisson counting data, as demonstrated by References 10, 11, and 13, the signaling thresholds can always be redefined in the scale of the original counting scale. For the pair of hypotheses in (1), the signaling threshold in the scale of X_t is represented by an upper boundary, that is, H_0 is rejected for large values of X_t. With the hypotheses in (2), X_t is compared against a lower boundary, then H_0 is rejected for small values of X_t. With the formats in (3) and (4), both large and small values of X_t lead to rejection of H_0.

It is important to emphasize that, for the formats (1) to (4), if the lower boundaries a_i and $a(t)$ are used to monitor $W(X_t)$, then the thresholds in the scale of X_t are represented by inner boundaries, that is, one fails to reject H_0 for X_t values consistently close to its average under H_0.

2.1 | Statistical performance measures

Three important statistical performance measures useful to compare and evaluate sequential analysis designs are: (a) statistical power, that is, the overall probability of rejecting the null hypothesis, (b) expected sample size, and (c) expected time to signal. While expected sample size is the average sample size when the analysis is stopped, irrespectively of the decision drawn about H_0, the expected time to signal is a conditional expectation, defined as the average sample size when the null hypothesis is rejected. Although in practice group and continuous sequential methods are distinct in the matter of the number of looks at the data, theoretically we can establish a unified notation for these three statistical performance measures. As demonstrated by Reference 17, each group sequential design can be rewritten in terms of a continuous sequential design holding exactly the same statistical performance. Therefore, without lost of generality, these three metrics can be expressed using the notation for the continuous sequential design.

Let τ denote the number of events when the surveillance is interrupted. The statistical power is given by:

$$\beta(RR) = Pr\left[\cup_{i=1}^{\eta} \{\tau = i\} \,|RR\right], \tag{6}$$

where η is found by evaluating the very same expression (6) iteratively, for each i, starting with $i = 1$, under the null hypothesis. That is:

$$\eta = \max\left\{ x \in \mathbb{N} : \sup_{RR^* \in \Theta_0} Pr\left[\cup_{i=1}^{x} \{\tau = i\} \,|RR^*\right] \leq \alpha \right\}. \tag{7}$$

The supreme in (7) is usually simple to evaluate for most of the probability distributions adopted in sequential analysis, and this is so because the probability argument in (7) is monotone with RR. Therefore, for hypotheses of the format in (1) to (3), the argument of the supreme is $RR^* = RR_0$. For the format in (4), the argument that solves (7) is either $RR_{0,l}$ or $RR_{0,u}$. This is valid for the probability models discussed in this article. Although the exact calculation can be performed by running a Markov Chain in i, the specific analytical expression for the probability in (7), and so in (6), are somewhat intricate. For the detailed power functions in each case, we indicate expression (13) by Reference 11 for Poisson data, expressions (8) and (29) by Reference 13 for binary data, and expressions (16) and (22) by Reference 10 for conditional Poisson data.

The monotonicity of $\beta(RR)$ with respect to RR favors to find signaling thresholds and maximum length of surveillance in order to control the statistical power for target points of the parameter space. More precisely, for a target relative risk under H_1, say RR_1, if a given sequential design leads to $\beta(RR_1) = \gamma$, then it holds that $\beta(RR^*) \geq \gamma$ for each $RR^* \geq RR_1$ with the hypotheses in (1), or for each positive $RR^* \leq RR_1$ with the hypotheses in (2). Similarly, considering two target relative risks under H_1, say $RR_l < RR_0$ and $RR_u > RR_0$ for (3), or $RR_l < RR_{0,l}$ and $RR_u > RR_{0,u}$ for (4), if $\beta(RR_l) = \gamma_l$ and $\beta(RR_u) = \gamma_u$, then $\beta(RR^*) \geq \gamma_l$ for each positive $RR^* \leq RR_l$, and $\beta(RR^*) \geq \gamma_u$ for each $RR^* \geq RR_u$.

The expected time to signal, denoted by $\mathbb{E}[\tau|H_0 \text{ rejected}, R]$, is given by:

$$\begin{aligned}\mathbb{E}[\tau|H_0 \text{ rejected}, RR] &= 1 \times Pr[\tau = 1|H_0 \text{ rejected}, RR] + 2 \times Pr[\tau = 2|H_0 \text{ rejected}, RR] + \ldots \\ &\ldots + \eta \times Pr[\tau = \eta|H_0 \text{ rejected}, RR] \\ &= 1 \times \frac{Pr[\tau = 1|RR]}{Pr[H_0 \text{ rejected}|RR]} + 2 \times \frac{Pr[\tau = 2|RR]}{Pr[H_0 \text{ rejected}|RR]} + \ldots\end{aligned}$$

$$\ldots + \eta \times \frac{Pr[\tau = \eta | RR]}{Pr[H_0 \text{ rejected} | RR]} = \frac{\sum_{i=1}^{\eta} i \times Pr[\tau = i | H_0 \text{ rejected}, RR]}{\beta(RR)}.$$

The expected sample size, denoted by $\mathbb{E}[\tau|RR]$, is given by:

$$\begin{aligned}\mathbb{E}[\tau|RR] &= 1 \times Pr[\tau = 1|RR] + 2 \times Pr[\tau = 2|RR] + \cdots + \eta \times Pr[\tau = \eta|RR] + \eta \times Pr[H_0 \text{not rejected}|RR] \\ &= \sum_{i=1}^{\eta} i \times Pr[\tau = i|RR] + \eta \times [1 - \beta(R)].\end{aligned}$$

Seeking a straightforward readability, the expected time to signal and the expected sample size will sometimes be referred through the acronyms ETS and ESS, respectively. Note that these two expectations are connected to each other, but the signaling thresholds that minimize $\mathbb{E}[\tau|H_0 \text{ rejected}, RR]$ and $\mathbb{E}[\tau|, RR]$ will usually differ. This topic shall be further discussed in Section 3.2.

2.2 | Binary data

Binary data appears in many sequential analysis problems, such as for Simon's two-stage group binomial sequential analysis,[18] and placebo-controlled two-arm clinical trials, where patients exposed to a drug are compared with matched unexposed subjects. Let C_n denote the number of exposed individuals in a total of n subjects, and assume that

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

follows a Bernoulli distribution with success probability $p_{n,RR}$, for $n = 1, 2, \ldots$, and $C_0 = 0$. In addition, $Y_1, Y_2, \ldots$ are independent, that is,:

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

for arbitrary sequences $y_1, \ldots, y_n$. The Bernoulli probability is given by:

$$p_{n,RR} = 1/(1 + z_n/RR),$$

and z_n denotes the matching ratio of the nth observation. For instance, if there are $k > 0$ controls matched to each case at the nth test, then $z_n = k$.

The Maximized Sequential Probability Ratio Test (MaxSPRT) statistic for the nth observation, in the log-scale, is given by:

$$\begin{aligned}LLR_n = I\left(\hat{RR} \notin \Theta_0\right) \times \max_{\{RR^* \in \tilde{\Theta}_0\}} \sum_{i=1}^{n} Y_i \log \frac{\hat{RR}}{\hat{RR} + z_i/RR^*} - \sum_{i=1}^{n} (1 - Y_i) \log(\hat{RR} + z_i/RR^*) - \\ - \sum_{i=1}^{n} Y_i \log \frac{1}{1 + z_i/RR^*} + \sum_{i=1}^{n} (1 - Y_i) \log(1 + z_i/RR^*),\end{aligned}$$

where $\tilde{\Theta}_0 = \{RR_0\}$ for the hypotheses in (1) to (3), and $\tilde{\Theta}_0 = \left\{RR_{0,l}, RR_{0,u}\right\}$ for the hypotheses in (4).

The subset of the parameter space under H_0, Θ_0, is defined according to (5), and $\hat{RR}$ is the maximum likelihood estimator of RR, which, in general, can be solved numerically for multiple different matching ratios over time. If the Bernoulli probability is fixed over time, that is, $z_n = z$ for each n, then the MaxSPRT statistic simplifies to:

$$\begin{aligned}LLR_n = I\left(\hat{RR} \notin \Theta_0\right) \times \max_{\{RR^* \in \tilde{\Theta}_0\}} C_n \left(\log \frac{C_n}{n} - \log \frac{1}{1 + z/RR^*}\right) + \\ + (n - C_n) \left[\log \frac{n - C_n}{n} - \log\left(1 - \frac{1}{1 + z/RR^*}\right)\right],\end{aligned}$$

where the maximum likelihood estimator of R is given by:

$$\hat{RR} = zC_n/(n - C_n).$$

The MaxSPRT was specially developed for post-market drug and vaccine safety surveillance, where the hypotheses are of the one-tailed format in (1), and a flat upper signaling threshold, cv. That is, H_0 is rejected for the first n such that $LLR_n \geq b_n = cv, n = 1, \ldots, N$, otherwise, the analysis is finalized in favor of H_0 for $n = N$. As demonstrated by Reference 6, for arbitrary significance level, $\alpha \in (0, 1)$, the exact cv can be calculated through an iterative numeric procedure by running a Markov Chain in the spirit of References 19-21.

The maximum likelihood estimator of RR can be solved numerically for multiple different matching ratios for both over time and within the same batch of data. This type of data frequently occur in post-market drug safety surveillance, where matching or stratification variables are used for confounding control. The stratification variables create risk sets of comparable subjects. In these data sets, each record would include (1) a binary variable indicating whether the adverse event or outcome of interest was exposed to the treatment or comparator exposure and (2) the proportion of treatment group-exposed patients in the risk set at the time the adverse event occurs. In this manner, the person-time data that typically populate a stratified Cox proportional hazards regression model are the same data that populate what is known as a case-centered logistic regression. The authors of Reference 22 showed that the models are mathematically identical and yield the same parameter estimates. In this manner, person-time data can be treated as a sum of binary data and all of the functions described above can be used accordingly.

2.3 | Multiple weighted binary endpoints

When there are multiple different outcomes, an important extension for binary sequential analysis is the possibility of considering weights reflecting practical interpretations, such as severity of different disease outcomes. For example, consider the case of two different outcomes, the first with weight $w = 2$, and the second with weight 1. That is, a single event of the first outcome would be equivalent to 2 independent outcomes of the second type. In general, if the first outcome type has weight w and the second has weight 1, then the first would be considered w times more severe than the second outcome type.

The authors of Reference 23 proposed a test statistic based on the weighted sum of the outcomes. For this, let $\vec{C}_1, \vec{C}_2, \ldots$, denote a sequence of D-dimensional random vectors, where $\vec{C}_i = (C_{i,1}, \ldots, C_{i,D})$, with $i = 1, 2, \ldots$, and $C_{i,j} \sim binomial(n_{i,j}, p_{j,RR_j})$, for $j = 1, \ldots, D$. Also, assume that $C_{i,j}$ is independent of $C_{i',j'}$ for each $i \neq i'$ or $j \neq j'$. To illustrate, suppose that a two armed sequential testing, where two groups, called exposure I and exposure II, are compared to each other. For the ith test, $C_{i,j}$ counts the number of individuals from group I presenting the jth endpoint. Therefore, $n_{i,j}$ is the total number of observations from endpoint j accruing in the ith test. The success probability is given by $p_{j,RR_j} = 1/(1 + z_j/RR_j)$, where z_j is the matching ratio between exposure I and exposure II with outcome j, that is, p_{j,RR_j} is the probability of having an observation from exposure I presenting outcome j. For example, if there are v exposures I matched to each exposure II for endpoint j, then $z_j = v$.

For the ith test, consider the test statistic defined as a composite endpoint, denoted by S_i, constructed as a weighted sum of cumulative outcomes, that is:

$$S_i = \sum_{g=1}^{i} (w_1 C_{g,1} + \cdots + w_D C_{g,D}), \tag{8}$$

where w_j is the weight associated to the jth outcome. Flat critical values in the scale of S_i can be solved under arbitrary significance levels using numeric procedures.[23] That is, H_0 is rejected for the first n_i such that $S_i \geq b_{n_i} = cv$, where $n_i = \sum_{g=1}^{i} \sum_{j=1}^{D} n_{g,j}$. Otherwise, the analysis is finalized in favor of H_0 for the first test such that $n_i \geq N$.

2.4 | Poisson data

Let C_t denote the number of events up to the continuous time index, t. Under H_0, C_t follows a Poisson distribution with mean μ_t, where μ_t is a known baseline function of t. Under H_1, C_t is still Poisson, but now with mean $RR\mu_t$. In this case, the MaxSPRT statistic, given in the log-likelihood ratio scale, is given by:

$$LLR_t = \max_{\{RR^* \in \tilde{\Theta}_0\}} \left[(\mu_t^* - c_t) + c_t \log c_t/\mu_t^*\right] \times I\left(\hat{RR} \notin \Theta_0\right),$$

where $\mu_t^* = RR^*\mu_t$, $\tilde{\Theta}_0 = \{RR_0\}$ for the hypotheses in (1) to (3), and $\tilde{\Theta}_0 = \left\{RR_{0,l},\ RR_{0,u}\right\}$ for the hypotheses in (4).

The subset of the parameter space under H_0, Θ_0, is defined according to (5), and the maximum likelihood estimator of RR is given by:

$$\hat{RR}(t) = c_t/\mu_t.$$

The null hypothesis is rejected as soon as $LLR_t \geq b(t) = cv$, with $t \in (0,\ T]$, where T is the maximum sample size. For continuous sequential designs,[6] shows that the exact critical value can be obtained by running a Markov Chain in the spirit of Reference 21. The solution by Reference 6 is also valid for group sequential analysis when observations are accrued following a fixed and common period of observation, which is the assumption behind the exact method by Reference 24 too. A generalized exact solution, valid for continuous, group or mixed continuous-group sequential, including the applications where the periods of observations are unpredictable, was derived by Reference 11.

2.5 | Conditional Poisson data for unknown baseline mean

The MaxSPRT statistic is a function of the data and of the Poisson rate under H_0, μ_t. Therefore, MaxSPRT can be calculated only if μ_t is known. Otherwise, a conditional Poisson distribution can be used.[7]

Let V denote the person-time in the historical sample containing c events, and let P_k denote the cumulative person-time observed until arrival of the kth event during the surveillance period. As a consequence of the Poisson process, and for a known c, V follows a Gamma distribution with shape c and scale $1/\lambda_V$.[9] Likewise, for a known k, P_k follows a Gamma distribution with shape k and scale $1/\lambda_P$. This way, the CMaxSPRT test statistic, in the scale of the log-likelihood ratio, is given by:

$$U_k = \max_{\{RR^* \in \tilde{\Theta}_0\}} \left[c \log \frac{c(1 + RR^*P_k/V)}{c+k} + k \log \frac{k(1 + RR^*P_k/V)}{(RR^*P_k/V)(c+k)}\right] \times I\left(\hat{RR} \notin \Theta_0\right),$$

where $\tilde{\Theta}_0 = \{RR_0\}$ for the hypotheses in (1) to (3), and $\tilde{\Theta}_0 = \left\{RR_{0,l},\ RR_{0,u}\right\}$ for the hypotheses in (4).

The subset of the parameter space under H_0, Θ_0, is defined according to (5). As derived by Reference 7, the maximum likelihood estimator of RR for each k is given by:

$$\hat{RR}(k) = \frac{k \times V}{c \times P_k}.$$

The null hypothesis is rejected as soon as $U_k \geq b_k = cv$, with $k = 1, \ldots, K$, where K is the maximum sample size. Calculation of the exact cv, for arbitrary tuning parameters settings, is performed through numerical procedures.[9]

2.6 | Minimum number of events before rejection of the null hypothesis

According to Reference 25, requiring a minimum number of events before allowing rejection of H_0 can reduce the expected time to signal. The idea is to establish a minimum number of events, say M, in such a way that H_0 can only be rejected after having observed at least M events after starting the analysis. Although all data observations affect the total sample size for the final decision, a decision against H_0 can only be taken after observing M events or more. This approach provides gains in terms of expected time to signal. The authors of Reference 25 showed that, depending on the tuning parameters, M values between 3 and 6 reduce the expected time to signal without affecting the overall statistical power of the sequential test. They also indicated that, in general, $M = 4$ is a good choice.

2.7 | Alpha spending

Usually, the approach for sequential testing is based on using signaling thresholds given in the scale of a test statistic. This is the case for both classical and new methods, such as the procedures of References 1, 3, 4, 6, 7, and 26. Alternatively,

one can use an alpha spending function. Denoted by $F(t)$, it is a function that establishes the amount of Type I error probability to be spent at each time t. For $t \in (0,\ 1]$, four well-known choices are:

$$
\begin{aligned}
&F_1(t) = \alpha \times t^{\rho},\ \rho > 0,\\
&F_2(t) = 2 - 2 \times \Phi(x_\alpha \times \sqrt{t^{-1}}), \text{where } x_\alpha = \Phi^{-1}(1-\alpha/2),\\
&F_3(t) = \alpha \times \log\{1 + [\exp(t) - 1] \times t\},\\
&F_4(t) = \alpha \times \left[1 - \exp\{-t\gamma\}\right] / \left[1 - \exp\{-\gamma\}\right],\ \gamma \in \mathfrak{R}.
\end{aligned}
$$

Consider, for simplicity, that only the upper limit signaling threshold ($b(t)$), according to Definition 2, is used. Under a adaptive design, where data arrive with chunks of unpredictable sample sizes. Remind that τ denotes the number of events when the surveillance is interrupted, and η is the maximum length of surveillance expressed in the scale of the number of events. For the ith observed event, make $t = i/\eta$. Note that $\eta = N$ with the Bernoulli data. The upper signaling threshold is elicited from the target alpha spending in the following way:

$$
b(t) = \max\left\{ x \in \mathbb{N} : \sup_{RR^* \in \Theta_0} Pr\left[\cup_{i=1}^{x} \{\tau = i\} \,|RR^*\right] \leq F(t) \right\}.
$$

According to References 27-29, the power-type function, $F_1(t)$, is useful to approximate Pocock's and O'Brien & Fleming's procedures. $F_1(t)$ also approximates MaxSPRT designs. It produces a line for $\rho = 1$, a convex curve for $\rho > 1$, and a concave curve for $0 < \rho < 1$. The authors of Reference 8 offers a detailed description on proper choices for ρ in order to minimize expected time of surveillance for fixed power. ρ values around 2, producing convex shapes, seems to provide small expected time of surveillance. If expected time to signal, instead of expected time of surveillance, is the target performance measure, then concave shapes ($\rho < 1$) are more appropriate.[11,30] $F_2(t)$, is shown by Reference 31 to approximate O'Brien & Fleming's procedure.[31] also explored $F_3(t)$ in order to approximate Pocock's test. The results by Reference 31 were used by the authors of Reference 32 to derived exact calculations for discrete data. $F_4(t)$, introduced by Reference 33, is also a good option to mimic Pocock's test. As shown by References 11,30, $F_1(t)$ and $F_4(t)$, under concave curves, are more appropriate choices than $F_2(t)$ and $F_3(t)$ if minimizing expected time to signal is the meaningful design criterion for binomial and Poisson data. Instead, for conditional Poisson data[10] show that a convex form for $F_1(t)$ should be used, and that $\rho = 1.5$ is a proper choice for most of the applications.

There are many proposed alpha spending functions in the literature. The authors of Reference 8 offer a rich overview and comparison of the most widely used functions.

As demonstrated by References 10,11, and 13, methods based on signaling thresholds can always be rewritten in terms of an alpha spending function, but the reciprocal is not true. Therefore, the alpha spending approach is the most general for sequential analysis. The authors of Reference 13 derived the optimal alpha spending function for binomial data for a set of target performance measures, such as power, expected sample size, and expected time to signal. Their solution is obtained through linear programming.

3 | SEQUENTIAL ANALYSIS PLANNING

For designing a sequential test procedure, it is often important to calculate statistical power, expected time to signal, expected sample size, and maximum sample size. As there are trade-offs among these metrics, the sequential plan should be defined according to the design criterion of each application. For instance, for post-market drug and vaccine safety surveillance, a large number of individuals are exposed to the drug/vaccine, then sample sizes are usually large even when the monitored event is rare. But, there is still the need for a fast identification of elevated threats from the drug, therefore minimizing the expected time to signal is a critical design criterion. Conversely, in Phase III clinical trials the number of individuals available for the study is usually of small or moderate magnitudes. Thus, minimizing the sample size by ending analysis early is of major importance since it applies that the number of affected individuals are minimized as well.

3.1 | Sample size calculations with flat thresholds

Recall that the sequential analysis ends without rejecting H_0 when the sample size reaches a pre-specified upper limit. Such upper limit must be defined in advance according to the desirable pre-experimental performance measures. To show how this can be done in practice, we start with the conventional Wald's flat-signaling threshold approach given in the scale of the log-likelihood ratio. After defining the acceptable test size through the tuning parameter α, the subset of the parameter space to form the null hypothesis (Θ_0), and the size of the effect through the target power under meaningful points of Θ, the next step is the calculation of the required sample size, N.

3.1.1 | Binomial data

For binomial data under a continuous sequential manner, Table 1 shows sample sizes (N) calculated for testing $H_0 : RR \leq 1$ under $\alpha = 0.05$ using $z = 0.25,\ 0.5,\ 0.75,\ 1,\ 2,\ 3,\ 4$, and power of 0.9 and 0.99 under $RR \geq 2$ and $RR \geq 4$.

In practice, the number of events appearing during the sequential analysis is a portion of the total number of participants receiving the treatments, then one should plan the maximum length of surveillance (N) according to a table, possibly with many other scenarios of α, RR, and z, in order to ensure a reasonable statistical power. For example, if a total of $P = 1000$ matched patients are randomized in two groups, say placebo and treatment groups, and if it is known that, under H_0, around 10% of the participants may present the monitored event, then the sequential procedure (with Wald's boundary) detects an increased relative risk of about 2 with power of 0.9 for z values in between 1 and 2 (see the first line of Table 1). One may also evaluate the reverse, that is, based on the frequency at which the events occur under H_0, one can obtain the minimum number of patients (P) needed for detection of target relative risk, power, and alpha level. Evaluations like this are also important for determining the ratio z for the number of placebo to treatment groups.

3.1.2 | Poisson data

With Poisson data the predetermined upper limit on the sample size (T) is expressed in terms of the expected number of events under the null hypothesis. For instance, the sequential test may stop as soon as the cumulative sample size is such that there are at least $T = 30$ expected events under H_0. For a power of 0.9 and 0.99 with $RR > 2$ and $RR > 4$ for testing $H_0 : RR \leq 1$. Table 2 presents the related sample sizes (maximum length of surveillance) to comply with these performance requirements.

If the baseline expected number of events (μ_0) per test is unknown, one option is to use the CMaxSPRT test as described in Section 2.5. In this case, the sample size is expressed either in terms of the ratio of the cumulative person-time in the

TABLE 1 Sample sizes (N) with binomial data ($z = 0.25,\ 0.5,\ 0.75,\ 1,\ 2,\ 3,\ 4$) for testing $H_0 : RR \leq 1$ under $\alpha = 0.05$ for power $= 0.9,\ 0.99$ under $RR = 2$ and $RR = 4$

		z						
RR	**Power**	**0.25**	**0.5**	**0.75**	**1**	**2**	**3**	**4**
2	0.9	222	147	123	112	110	120	132
2	0.99	373	245	212	194	192	216	238
4	0.9	70	44	33	30	28	25	29
4	0.99	113	73	57	53	46	47	50

TABLE 2 Sample sizes (T) with Poisson data for testing $H_0 : RR \leq 1$ under $\alpha = 0.05$ for power $= 0.9,\ 0.99$ under $RR = 2$ and $RR = 4$

RR	**Power**	***T***
2	0.9	18.32
2	0.99	32.83
4	0.9	2.77
4	0.99	5.29

TABLE 3 Sample sizes in terms of the maximum length of surveillance by doses/person-time (T) and by the cases in the surveillance data (K), for powers of 0.9 and 0.99 under $\alpha = 0.05$ and $RR = 2$

	Power = 0.9		Power = 0.99	
c	K	T	K	T
50	66	0.79	290	3.89
70	50	0.43	157	1.53
100	43	0.25	96	0.65
120	40	0.20	83	0.47
150	38	0.15	73	0.33
170	37	0.13	69	0.27
200	36	0.11	64	0.21

TABLE 4 Power, expected time to signal and expected sample size for Poisson data with known baseline rates using Wald's lower signaling thresholds 2.5, 2.6, 2.7, 2.8, and upper signaling thresholds 3, 3.1, 3.2, 3.3

RR	Power	Expected time to signal	Expected sample size
0.3	0.978	25.000	26.402
0.9	0.024	25.875	88.476
1.0	0.022	41.019	88.932
1.2	0.299	57.763	80.367
1.5	0.965	42.791	44.451

Note: The expected number of events (group sizes) used are 25, 20, 20, 25 (ie, T = 90). The calculations were ran for $RR = 0.3,\ 0.9,\ 1,\ 1.2,\ 1.5$.

surveillance population divided by the total cumulative person-time in historical population (T), or in terms of the number of events in the surveillance data (K). For instance, the monitoring may end as soon as the sample size is such that the cumulative person-time in the surveillance population is equal to the cumulative person-time in historical population, or if there are 30 events in the surveillance data.

For testing $H_0 : RR \leq 1$, sample sizes in both scales are shown in Table 3 for selected numbers of events in the historical data.

Naturally, in real data analysis the information may arrive in chunks of sizes greater than 1. But, the sample sizes in Tables 1, 2, and 3 still ensure the target power since group sequential testing is powerful than the continuous fashion under the same significance level.[17] Section 4.3 presents a real data analysis to exemplify the usage of time specific alpha spending functions for preserving statistical power with unpredictable mixed group-continuous data arrival structures.

3.2 | Performance evaluations for arbitrary thresholds

For Poisson data with known baseline rates, suppose that one desires to test:

$$H_0 : 0.9 \leq RR \leq 1.2 \ \text{ against } \ H_0 : RR < 0.9 \ \text{ or } \ RR > 1.2. \tag{9}$$

Consider the Wald's lower signaling thresholds 2.5, 2.6, 2.7, 2.8, and the upper signaling thresholds 3, 3.1, 3.2, 3.3, with expected number of events (group sizes) equal to 25, 20, 20, 25 (ie, $T = 90$), respectively. Table 4 contains the performance measures of this very specific (non-flat signaling threshold) sequential design.

Assume that one requires a statistical power of 0.9 for $RR \leq 0.3$ or $RR \geq 1.5$. From Table 4, we see that this requirement is indeed satisfied since for $RR \leq 0.3$ and $RR \geq 1.5$ the powers are about 0.98 and 0.965, respectively. Note also that the type I error probabilities for RR values between 0.9 and 1 are smaller than 0.024. However, this is not a 0.024 level test because the type I error probability is close to 0.3 for $RR = 1.2$. Therefore, the critical values should be conveniently modified in order to adjust the actual test size according to the nominal alpha level. A simplistic solution is to use the regular flat threshold approach again. For example, after checking a few scenarios, we found that the flat threshold $cv = 6.8$ promotes

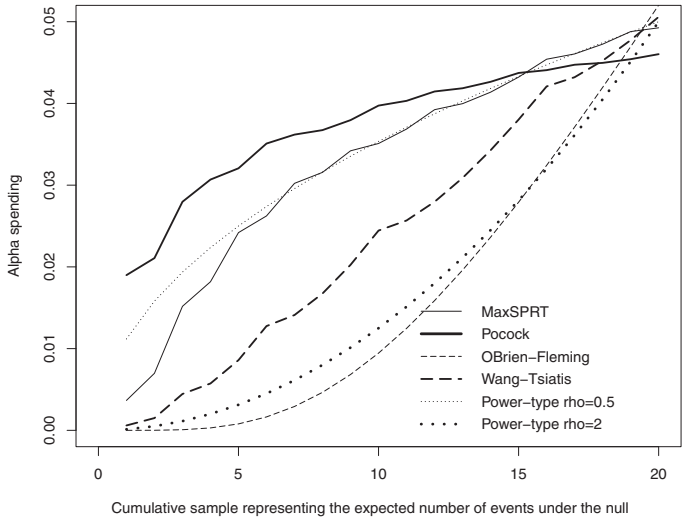

FIGURE 1 Alpha spending implied by flat signaling thresholds with Poisson data in the scale of MaxSPRT ($cv = 2.59$), Pocock ($cv = 2.83$), O'Brien-Fleming ($cv = 2.01$), and Wang-Tsiatis ($cv = 2.12$) test statistics under $\alpha = 0.05$ and $RR = 2$

	MaxSPRT	Pocock	O'Brien-Fleming	Wang-Tsiatis
Power	0.95	0.93	0.98	0.97
ETS	7.06	7.10	9.41	7.91
ESS	7.68	8.06	9.66	8.28

TABLE 5 Power, expected time to signal (ETS) and expected sample size (ESS) for MaxSPRT ($cv = 2.59$), Pocock ($cv = 2.83$), O'Brien-Fleming ($cv = 2.01$), and Wang-Tsiatis ($cv = 2.12$) test statistics under $\alpha = 0.05$ and $RR = 2$ based on $T = 20$ and samples of size 1

a 0.045 level test. But, the test no longer leads to power magnitudes greater than of 0.9 for RR around 0.3 and 1.5. Actually, the power is 0.662 for $RR = 0.3$, and 0.782 for $RR = 1.5$. This is an indication that a sample size greater than 90 is needed to satisfy the target power of 0.9.

The exercise above shows that guessing the signaling threshold is not an easy task as it demands to control desirable performance measures vis-á-vis the format of the hypotheses. A straightforward approach for planning non-flat signaling thresholds is to use the alpha spending approach. In fact, the alpha spending function is a general method as virtually any sequential procedure can be rewritten in terms of an implicit alpha spending. For example, consider a twenty-group sequential testing with test specific sample sizes all equal to 1 (ie, $T = 20$). Using $\alpha = 0.05$ for testing $R_0 : RR \leq 1$, the critical values (cv) in the scales of MaxSPRT, Pocock, OBrien-Fleming, and Wang-Tsiatis are 2.59, 2.83, 2.01, and 2.12, respectively. These critical values were obtained through a bisection procedure. The command lines are placed in the Appendix part.

Figure 1 shows the alpha spending curves for each test. We note that Wald's (MaxSPRT) and Pocock's alpha spendings are concave while O'Brien-Fleming and Wang-Tsiatis are convex. As illustrated with Table 5, such differences on the alpha spending shapes have critical implications in the pre-experimental performance measures.

According to Reference 11, for fixed power and level the expected time to signal is minimized with concave alpha spending shapes, while expected sample sizes are minimized with convex functions. The authors of Reference 11 showed that the power-type family, function $F_1(t)$ in (9), nearly-minimizes expected time to signal with Poisson data for $\rho = 0.5$. But, for minimizing expected sample size,[8] suggest that one should use ρ around 1.5 or 2.

For a continuous sequential analysis, Figure 1 shows the power-type alpha spending for $\rho = 0.5$ and $\rho = 2$. Figure 2 shows the signaling thresholds, event-by-event, in the four test statistic scales elicited from these two choices of ρ, and Table 6 shows the related statistical performance measures.

This way, no-matter the frequency of the data arrival, the user can simply use the time-specific alpha spending to calculate the critical value according to the actual amount of information (cumulative expected number of events under H_0) in hand. This type of unpredictable data frequency sequential testing shall be illustrated with real data in Section 4.4.

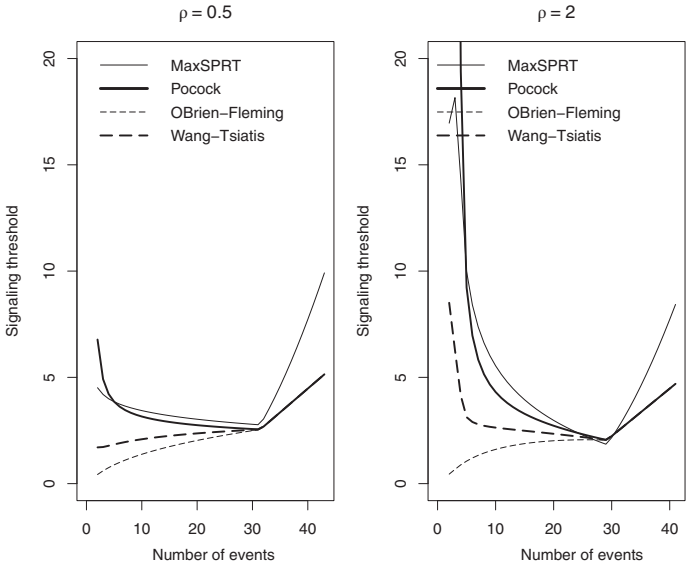

FIGURE 2 Signaling thresholds for continuous sequential testing implied by the power-type alpha spending with $\rho = 0.5$ and $\rho = 2$ for Poisson data in the scale of MaxSPRT, Pocock, O'Brien-Fleming, and Wang-Tsiatis test statistics under $\alpha = 0.05$ and $RR = 2$.

TABLE 6 Statistical performance measures by the power-type alpha spending using $\rho = 0.5,\ 2$ under $\alpha = 0.05$ and $RR = 2$ based on $T = 20$ in a continuous sequential fashion

Performance measure	$\rho = 0.5$	$\rho = 2$
Statistical power	0.95	0.97
Expected time to signal	6.90	8.48
Expected sample size	7.52	8.78

For the historical versus surveillance Poisson data (CMaxSPRT),[10] suggest to use $\rho = 1.5$ as it is near-optimal in the sense of minimizing both expected time to signal and expected sample size in most of the real data applications.

3.3 | Optimal sequential design for binomial data

The alpha spending can be specified to optimize a performance measure of interest. For example, one can elicit the alpha spending shape that minimizes either the expected time to signal or the expected sample size. This is possible with the exact optimal solution introduced by Reference 13. For $H_0 : RR \leq 1$ against $H_1 : RR > 1$, suppose that we want the alpha spending shape that minimizes the expected time to signal while guaranteeing statistical power of 0.8 for any relative risk greater than 2.

If we instead wish to minimize expected sample size. Figure 3 compares the optimal alpha spending solutions between these two different objective performance measures for $\alpha = 0.05$, and $z = 1$. The optimal expected time to signal is 32.99, and the optimal expected sample size is 39.63. The optimal samples sizes are $N = 77$ and $N = 59$, respectively. The curves in Figure 3 have different shapes. While the optimal expected time to signal is obtained under a concave alpha spending shape, the optimal expected sample size is reached with a convex alpha spending function.

For two-tailed testing, that is, when the hypotheses are of the form $H_0 : RR = 1$ against $H_1 : RR \neq 1$, one needs to specify two target powers, one to each of the two target relative risks under the alternative hypothesis. For example, suppose that we want to minimize expected time to signal, and both $RR \leq 0.5$ or $RR \geq 2$ should be detected with power of at least 0.8. Figure 4 shows the optimal alpha spending for two-tailed testing under this parametrization.

If there are restrictions on the sample size due to logistical, ethical and any other practical aspects, one can minimize expected time to signal and expected sample size while setting an upper bound on the maximum sample size. This is done through the input N on the command line above. For example, for $N = 80$, if one wishes to minimize expected time to signal under a power of 0.8 for $RR = 2$, $\alpha = 0.05$, and $z = 1$, the optimal solution leads to a minimum expected time to signal equal to 40.00, which is greater than the minimum expected time to signal, 32.99.

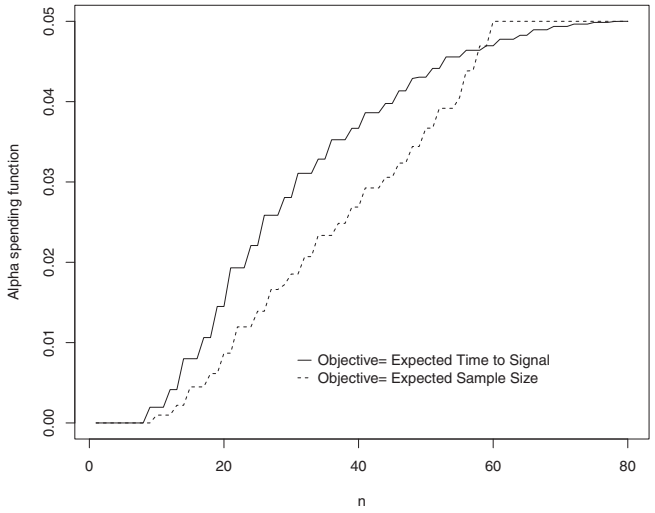

FIGURE 3 Optimal alpha spending shapes for expected time to signal and expected sample size under $\alpha = 0.05$, $z = 1$, power equal to 0.8, and $RR = 2$

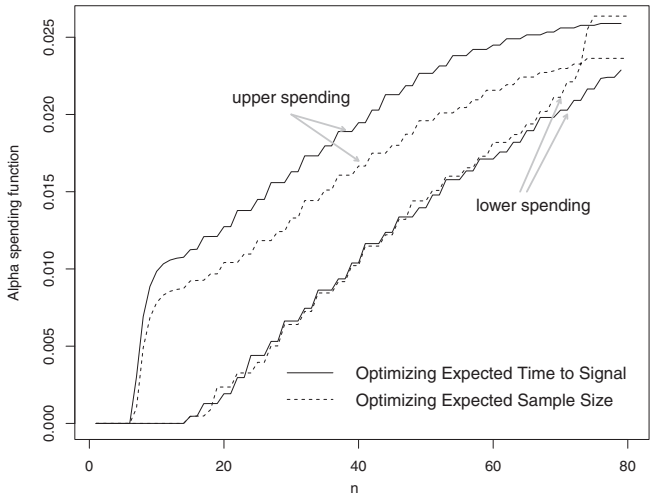

FIGURE 4 Two-tailed optimal alpha spending shapes for expected time to signal and expected sample size under $\alpha = 0.05$, $z = 1$, power equal to 0.8, and $RR_1 = 0.5$ and $RR_2 = 2$

The optimal solution proposed by Reference 13 also incorporates control on the precision of the relative risk estimate by the end of the sequential analysis through fixed-width and fixed-accuracy confidence intervals. This feature shall be exemplified with an illustrative clinical trial data in Section 4.1.

3.4 | Schematics for sequential testing planning

This section presents a synthesis of the main decision directions, so far discussed in this article, for balancing the trade-offs between statistical performance measures and the alpha spending plan. Such decisions are determinant to calculate the required sample size, that is, maximum length of surveillance.

The construction of a sampling design to collect the data takes in account ethical, logistical, and financial aspects. Concomitantly, the planning involves defining the hypotheses format, the overall alpha level, the target relative risk to detect under H_1, and the target power. The data structure, and then the underlying probability model to be used in the inference phase, results from the sampling scheme as well.

Once the data probability model (eg, binary, Poisson, conditional Poisson) is identified/defined, which we refer to as step (A), the next step is to (B) establish the meaningful statistical metric to optimize between expected sample size and expected time to signal, then (C) select the alpha spending plan, and finally (D) elicit the maximum length of surveillance in compliance with the alpha level and target power.

The actual maximum length of surveillance to be calculated in step D depends on the format of the hypotheses, significance level, and target power. For instance, the diagram in Figure 5 illustrates this steps for testing:

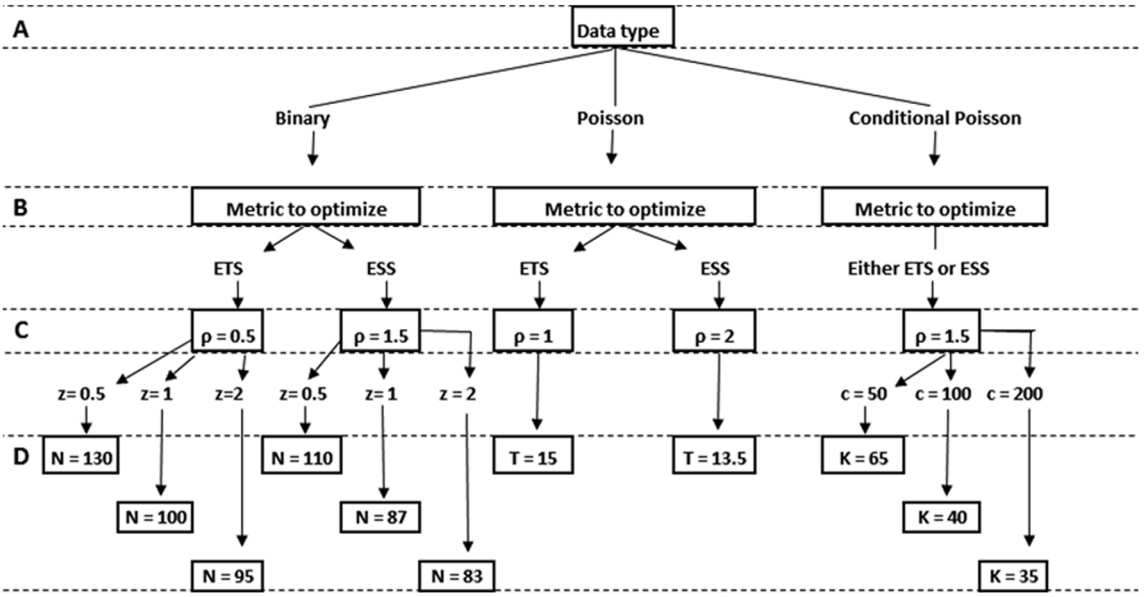

FIGURE 5 Diagram for setting a sequential analysis design for testing $H_0 : RR \leq 1$ against $H_1 : RR > 1$ under $\alpha = 0.05$, $M = 4$, and target power of 0.9 for $RR \geq 2$. For step C, values of the tuning parameter ρ are selected to fix the power-type alpha spending shape according to the performance metric to minimize between the expected time to signal (ETS) and the expected sample size (ESS)

$$H_0 : RR \leq 1 \text{ versus } H_1 : RR > 1,$$

with minimum number of events to start the surveillance equal to $M = 4$, $\alpha = 0.05$, and target power of at least 0.9 for $RR \geq 2$. Regarding the alpha spending shape, for this diagram we adopted the power-type shape, that is, $F_1(t) = \alpha \times t^{\rho}$ (see Section 2.7). By selecting the tuning parameter ρ conveniently, the power-type alpha spending works reasonably well to approximate the optimal alpha spending in each scenario.[10,11,30] In order to emphasize the possible changes in the maximum length of surveillance, three values were used for z, in the binary case, and for c, in the conditional Poisson case. All the calculations for step D were based on the continuous sequential analysis scenario. This ensures that the actual performance in terms of significance level and statistical power satisfies the nominal requirements, no matter the real frequency at which the data arrives in each application.

Naturally, the sample sizes in step D will suffer considerable changes if any of the tuning parameters are different from those used in this example, such as the format of the hypotheses and the set of the parameter space under H_0, and the target power.

It is important to note the key role of the alpha spending shape for the overall statistical performance of the sequential analysis. The values of ρ, presented in step C of Figure 5, are offered as a rule of thumb, a summary of the results found in the literature on this topic. However, ideally, the alpha spending shape should be customized according to each application, where the trade-offs appearing in the schematic presented here are confronted with the goals behind the design criterion/metrics to optimize, expected data frequency, or even intangible characteristics of each study.

4 | DATA ANALYSIS EXAMPLES

4.1 | Binomial data in placebo-controlled two-arm trial: monitoring adverse events in COVID-19 studies

The efforts to develop treatments for COVID-19 patients are urgent, therefore, important studies are currently in development in this direction. For example, the Adaptive COVID-19 Treatment Trial (ACTT), detailed by Reference 14, is a randomized placebo-controlled trial where intravenous remdesivir was administrated in 1062 adults hospitalized with COVID-19. The individuals were randomly divided in two groups, where 541 were assigned to remdesivir, and 521 to the

placebo group, then with a randomized ratio of $z = 521/541 \approx 0.96$. By the end of the analysis, the observed 131 severe adverse events from the remdesivir group and in 163 patients from the placebo group. We can also mention the study by Reference 15, where a randomized open-label trial was conducted on hospitalized adult patients of SARS-CoV-2. A total of 199 individuals were randomly divided in two groups, 99 to the lopinavir-ritonavir treatment, and the remaining 100 to the standard-care group. In this case, $z = 100/99 \approx 1.01$. Severe adverse events were observed in 19 participants from the lopinavir-ritonavir treatment, while 32 patients presented adverse events in the standard-care group.

Using a data structure similar to that described by References 14 and 15, here we mimic a clinical trial for comparing two hypothetical treatments, say Treatment A and Treatment B. Then, suppose that a randomized double-blind plan with randomized ratio $z = 1$ is conducted to a total of $P = 1000$ patients. We want to test:

$$H_0 : 0.9 \leq RR \leq 1.1, \tag{10}$$

$$H_1 : RR < 0.9 \text{ or } RR > 1.1. \tag{11}$$

Aiming to minimize the number of patients affected by severe adverse events, the optimal alpha spending that minimizes expected sample size was used restricted to $\alpha = 0.05$, and statistical power greater than or equal to 0.8 for either $RR \geq 2$ or $R \leq 0.5$. In addition, the optimal solution was constrained to a fixed-width and a fixed-accuracy 90% confidence interval for RR with the following formats:

$$\left[\hat{RR} - 1.5,\ \hat{RR} + 1.5\right] \text{ and } \left[\hat{RR}/1.5,\ 1.5\hat{RR}\right] \tag{12}$$

Figure 6 shows the optimal alpha spending solved according to the constraints above.

In practice, patients may leave the study due to many different reason other the presenting adverse events. Therefore, the number of patients in each arm, here denoted by $P_{i,A}$ and $P_{i,B}$ at the ith test, will decrease in time. For simulating

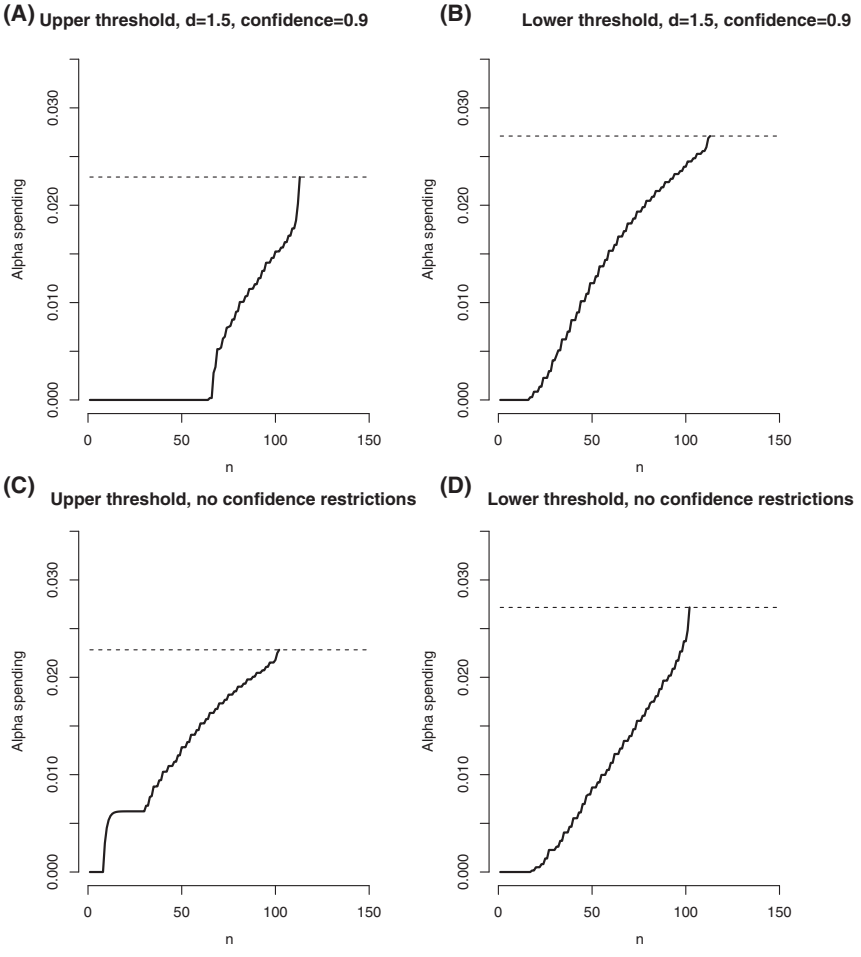

FIGURE 6 Optimal alpha spending minimizing expected sample size under $\alpha = 0.05$, power ≥ 0.8 for each $RR < 0.5$ and $RR > 2$ with hypothesis $H_0 : 0.9 \leq RR \leq 1.1$. Graphics A, and B, show the upper and lower cumulative alpha spending under fixed width and fixed accuracy 90% confidence intervals, and graphics C, and D, shows the optimal alpha spending without confidence interval constraints for comparisons

TABLE 7 Analysis results on simulated randomized, double-blind, placebo-controlled trial for monitoring severe adverse events

	# Patients				cv			$RR = 3$	
i	$P_{i,A}$	$P_{i,B}$	z_i	n_i	Lower	Upper	X_{n_i}	$\hat{RR}$	Rej. H_0
1	500	500	1	12	na	13	10	5	No
2	498	489	1.02	21	3	22	15	3.16	No
3	494	484	1.02	30	7	31	21	2.75	No
4	490	478	1.03	41	12	42	30	2.77	No
5	486	468	1.04	52	16	47	36	2.61	No
6	479	462	1.04	65	22	47	45	2.52	No
7	475	451	1.05	75	26	48	55	2.61	Yes
8	474	440	1.08	78	27	50	57	2.67	Yes
9	473	437	1.08	89	32	56	64	2.69	Yes
10	468	429	1.09	100	37	61	70	2.67	Yes
11	462	423	1.09	112	42	67	77	2.62	Yes

this effect, a discrete uniform random variable in the $\{0,\ 1,\ 2\}$ support was subtracted in both arms for each test. The number of adverse events in each test $(n_i - n_{i-1})$ was generated using a $binomial(P_i,\ 0.01)$, where $P_i = P_{i,A} + P_{i,B}$. For the first test, we used $n_1 \sim binomial(1000,\ 0.01)$. Finally, the adverse events from the Treatment B group at the ith test $(Y_n = X_{n_i} - X_{n_{i-1}})$ were generated using a $binomial(n_i - n_{i-1}, p_i)$, where $p_i = (1 + z_i/RR)^{-1}$ and $z_i = P_{i,A}/P_{i,B}$, with $R = 3$ under the alternative hypothesis.

The second and third columns of Table 7 present the number of patients in both Treatment A and B at each test. The lower and upper signaling thresholds, given in the scale of the number of events, are shown in columns 6 and 7 of Table 7. The number of adverse events from Treatment B and the maximum likelihood estimates of RR are shown columns 8 and 9. We note that the null hypothesis is rejected in the 7th look since the number of events, 55, extrapolated the upper signaling threshold, 48. By construction, a 90% confidence interval for RR is given by $[1.74,\ 3.92]$ since $2.61 - 1.5 = 1.11 < 2.61/1.5 = 1.74$ and $2.61 + 1.5 = 4.11 > 1.5 \times 2.61 = 3.92$.

The critical values in this example were elicited from the optimal alpha spending shown in Figure 6, and the relative risk estimates were obtained with the command lines in the Appendix.

4.2 | Multiple weighted binomial outcomes in propensity score matched patients: comparing two treatments of osteoporosis

The authors of Reference 23 applied the alpha spending approach to formulate a new methodology for monitoring multiple types of adverse events that are comparable through pre-specified weights. They used real data to illustrate their method for five different outcomes with the following weights: 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 ($w_5 = 0.30$). They mimicked a sequential testing for comparing 9340 patients treated with denosumab to 9340 propensity score matched patients who initiated bisphosphonates for treatment of osteoporosis, therefore $z = 1$. The goal was testing:

$$H_0 : RR_j = 1 \text{ for each } j = 1, \ldots, 5,$$
$$H_1 : RR_j \neq 1 \text{ for at least one } j \in \{1, 2, 3, 4, 5\},$$

where RR_j is the relative risk associated to the jth type of adverse event.

Table 8 was extracted from Reference 23. It contains the cumulative sample size at each test in each of the five endpoints (columns 1 to 5), and the test statistic in column 6, denoted by $U = S_{i,A}/S_{i,B}$, which is the ratio of exposure A to exposure B weighted sums, where exposure A is for denosumab treatment and exposure B is for bisphosphonates

TABLE 8 Sequential analysis results for the data of treatment of osteoporosis

Cumulative data per outcome						cv	
H-p frac.	**F. frac.**	**H. frac.**	**Infec.**	**Pneum.**	***U***	**Lower**	**Upper**
0	0	0	1	0	inf	na	na
1	2	0	5	3	1.68	0.05	19.75
2	7	0	6	7	1.22	0.17	5.84
3	9	1	10	10	0.94	0.25	3.96
6	13	1	14	14	0.65	0.34	2.98
10	17	1	20	16	0.69	0.39	2.54
12	21	1	26	21	0.58	0.44	2.28
15	29	1	28	27	0.55	0.48	2.09
20	32	2	37	32	0.70	0.52	1.92
23	44	4	48	43	0.76	0.58	1.73
26	57	6	59	55	0.66	0.62	1.62
32	72	8	79	60	0.74	0.66	1.52

Note: 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 ($w_5 = 0.30$). It was settled a maximum length of surveillance of $N = 1000$. The overall significance level used was $\alpha = 0.05$, with power-type alpha spending ($\rho = 0.5$). The critical values (cv) are shown in the scale of the test statistic given by the ratio between weighted sums from expose A to exposure B populations ($U = S_{i,A}/S_{i,B}$).

treatment. The lower and upper critical values, given in the scale of U, are shown in columns 7 and 8. This table can be reproduced with the command lines shown in the Appendix:

As the test statistic stayed in between the lower and upper signaling thresholds during the sequential monitoring, the empirical information suggests that treatments A and B do not differ in terms of the risks of these five adverse events.

4.3 | Poisson data in exposure-outcome pairs: monitoring neurological adverse events after Pediarix vaccination

This example uses the data of neurological adverse events after Pediarix vaccination from the Kaiser Permanente Northern California. With a single injection, Pediarix protects children from diphtheria, tetanus, whooping cough, hepatitis B, and Polio. The authors of Reference 6 used continuous MaxSPRT to analyze severe neurological symptoms in the period 1-28 days after the vaccination. Table 9 presents the data for the first 10 weeks out of 81 weeks of surveillance.

The second column of Table 9 contains the background rate (μ_t) of adverse events under the null hypothesis ($H_0 : RR = 1$). From this table, we see that H_0 was not rejected until the 10th test. If more data are entered in the function for subsequent chunks of data, it will run and disclose the regular output table. Figure 7 shows realized and signaling thresholds in both the cases and the MaxSPRT scales. Note how irregular are the shapes of the thresholds as the testing time evolves. The observed empirical information reached the signaling threshold in the 32th test. This signal occurred when the total amount of information reported a relative risk estimate about 2.5.

4.4 | Conditional Poisson in historical versus surveillance data: monitoring seizures after influenza vaccination

This last example uses a time-series of seizures during Days 0-1 after application of concomitant vaccination with inactivated influenza vaccine (IIV) and 13-valent pneumococcal conjugate vaccine (PCV-13). The fifth column of Table 10 contains the cumulative number of doses applied to 6-23-month-old children in the period of September 2013 to April

TABLE 9 Sequential analysis results for the first 10 weeks of surveillance of neurological adverse events after Pediarix vaccination

	Test specific		Cumulative				Alpha spending	
Week	μ_t	#Events	μ_t	#Events	cv	$\hat{RR}$	Target	Actual
1	0.04	0	0.04	0	na	1	0	0
2	0.06	1	0.10	1	3	10	0.0035	0.0015
3	0.08	0	0.18	1	3	5.56	0.0047	0.0017
4	0.10	0	0.28	1	3	3.57	0.0059	0.0020
5	0.11	0	0.39	1	3	2.56	0.0070	0.0056
6	0.12	0	0.51	1	4	1.96	0.0080	0.0060
7	0.13	0	0.64	1	4	1.56	0.0089	0.0074
8	0.12	1	0.76	2	5	2.63	0.0097	0.0075
9	0.11	0	0.87	2	5	2.3	0.0104	0.0080
10	0.12	0	0.99	2	5	2.02	0.0111	0.0089

Note: The critical values (cv) are given in the scale of the cumulative events and were settled under a maximum length of surveillance of $T = 20$ and based on the alpha spending implied by MaxSPRT, $\alpha = 0.05$.

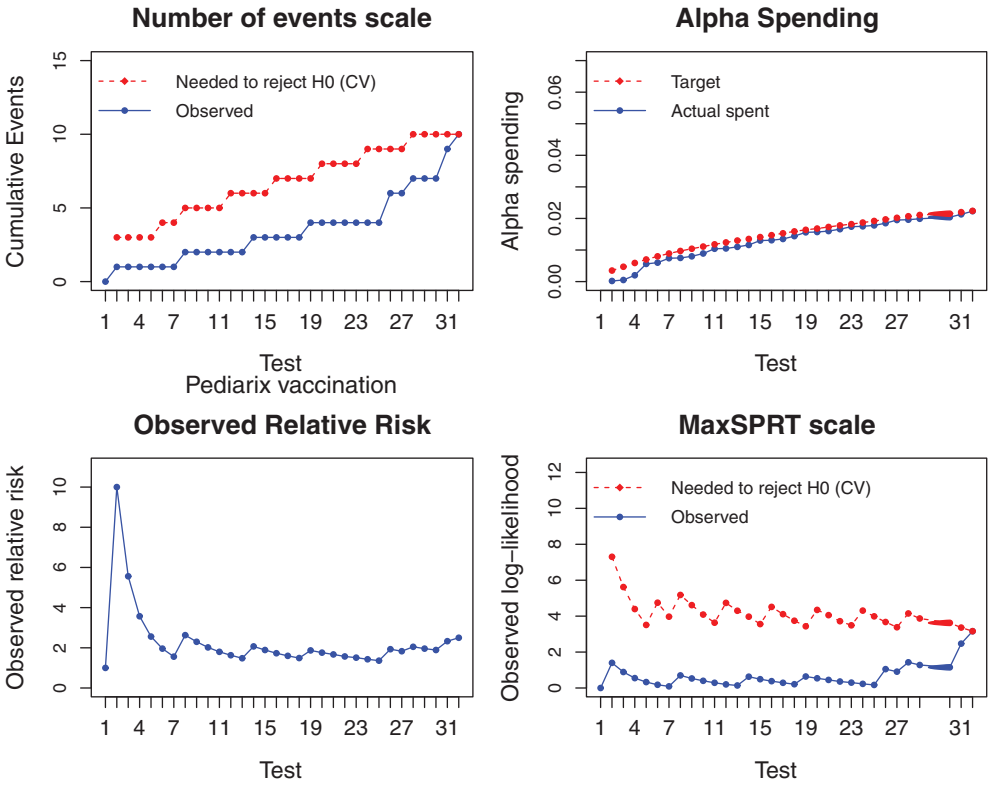

FIGURE 7 Critical values, observed data and alpha spending in the first 32 sequential tests for monitoring neurological adverse events after Pediarix vaccination. The critical values were settled under a maximum length of surveillance of $T = 20$ and based on the alpha spending implied by MaxSPRT, $\alpha = 0.05$. Evidences for rejecting the null ($H_0 : RR = 1$) occurred in week 32 [Colour figure can be viewed at wileyonlinelibrary.com]

2014. The sixth column of this table contains the cumulative number of individuals presenting seizures in the surveillance period. This data has already been used by,[10] and it come from three large U.S. health insurance or data companies ('Data Partners') participating in the U.S. Food and Drug Administration-sponsored Sentinel system. See Reference 34 for more details about the Sentinel system.

The authors of Reference 34 used Poisson MaxSPRT, with flat signaling threshold, based on the expected number of seizures estimated with the Data Partner-specific rates related to IIV vaccination in historical influenza seasons. In a different direction, instead of using estimated rates as if it is the real rate under H_0, here we use the conditional Poisson

TABLE 10 Sequential analysis results for the surveillance of seizures after influenza vaccination

	Test specific		Cumulative					
Chunk	p_k	**#Events**	P_k	**k**	P_k/v	**LLR**	**cv**	$\hat{RR}$
1	3877	0	3877	0	na	na	na	na
2	3211	0	7088	0	na	na	na	na
3	1	0	7089	0	na	na	na	na
4	25975	1	33064	1	0.04	0	6.49	0.62
5	8	0	33072	1	0.04	0	6.49	0.62
6	34760	5	67832	6	0.09	0.77	3.31	1.80
7	18497	3	86329	9	0.12	1.75	2.89	2.12
8	173	0	86502	9	0.12	1.56	2.89	2.12
9	17573	3	104075	12	0.14	2.81	2.45	2.35
10	12058	0	116133	12	0.15	2.34	na	2.10

Note: With a historical person-time information of $V = 752949$ (doses) and $c = 37$ adverse events in the historical period, the critical values (cv) are given in the scale of the log-likelihood ratio (LLR) statistic. The maximum length of surveillance is $K = 20$ under a power-type alpha spending ($\rho = 1.5$) with $\alpha = 0.05$.

approach described in Section 2.5. For this application, the cumulative number of doses reflects the person-time (P_k) for a given number k of observed events (seizures). The time-specific person-time is denoted by $p_k = P_k - P_{k-1}$, with $p_0 = 0$. The historical information is composed by $c = 37$ events in days 0-1 after $V = 752{,}949$ doses of IIV prior to licensure of PCV13.

We want to test if the relative risk is smaller than or equal to 1 (null hypothesis). Consider to setup the maximum length of surveillance as $K = 20$. It is important to check the pre-experimental statistical performance resulting from this sample size choice. Using the power-type alpha spending ($\rho = 1.5$) for learning evidences of $RR \geq 2$, we obtain a statistical power about 0.85, expected time to signal of 12.18, and expected length of surveillance equal to 13.35.

Note from the third column of Table 10 that new adverse events arrived in the surveillance period only in chunks 4, 5, 6, 7, and 9. Although one can revert the random variable taken the person-time as the amount of information and the related number of events as the monitored random measure of evidence,[9] hence the log-likelihood ratio statistic can still be calculated even when no new adverse events arrives, that would never lead to the null hypothesis rejection since the likelihood decreases when the relative person-time increases keeping the overall number of events fixed. Therefore, for futility, it is more convenient to keep the original framework by Reference 7, that is, the number of events are treated as the time index in the surveillance period, thus the monitored measure of evidence is the person-time ratio P_k/V for each fixed k. But, unlike in Reference 7, here we use the flexible group-sequential test instead of the continuous sequential manner. This ensures that alpha is spent only when new events arrive, increasing the overall statistical power of the sequential analysis.

After having twelve adverse events in only 14% of the person-time in the surveillance period relatively to the historical period, from Table 10 one can conclude in the 9th test that there are empirical evidences against H_0, which occurred when the relative risk estimate was about 2.35.

5 | CONCLUDING REMARKS AND SOFTWARE CONSIDERATIONS

This manuscript discuss sequential analysis hypothesis testing where the goal is to do inference of intrinsic characteristics of the analyzed phenomenon. Therefore, the ideas are not directly convertible for quality control problems, where the goal is to detect problems that suddenly appear during the surveillance, causing an abrupt increase in the relative risk.

Another point to emphasize is that we focused on exact calculations only. For those interested on easy-to-use tools based on the conventional practice of using approximations based on the asymptotic statistical theory or Monte Carlo

simulations, we recommend the following R packages: `ldbounds`, `Binseqtest`, `gsDesign`, `PwrGSD`, `seqDesign`, `seqmon`, `OptGSand`, `sglr`.

Regarding the R `Sequential` package, a limitation is that it is only applicable for analyzing binomial, Poisson, or conditional Poisson data. For other probability distributions, such as Gaussian or exponential processes, we recommend `GroupSeq` and `SPRT`.

The reason we opted to use R `Sequential` in this tutorial is its flexibility to dealing with the data structure that usually appears in real applications. For instance, `Sequential` is used to conduct the sequential monitoring of adverse events according to the master protocol for COVID-19 vaccine active surveillance in the United States.[16]

Unlike the R package `Sequential`, most R packages for sequential analysis are only designed for group sequential analysis. A review on available packages for group sequential designs is offered by Reference 35. Among the few alternative options for continuous sequential analysis, `Binseqtest` works only for binomial data, while `SPRT` only works for simple alternative hypotheses. Besides, although some of the packages cited above are able to provide statistical power and expected sample size calculations, such as `OneArmPhaseTwoStudy`, `ph2rand`, to the best of our knowledge, R `Sequential` is the only freely-available package that also calculates expected time to signal. The calculation of sample size for a given target power and relative risk, the calculation of signaling thresholds for a given alpha spending, and the calculation of alpha spending for given pre-experimental signaling thresholds, are also unique features of R `Sequential`.

It merits to remark that the package performs near-automated data analysis. The functions automatically signal when the rejection thresholds are reached. After each analysis, the package delivers the results in the form of both tables and graphs. The package has a detailed user guide explaining how different features and parameter options are specified. For people unfamiliar with the *R* language, there is a web interface (http://www.sequentialanalysis.org) where users only have to specify the input data and the analysis parameters. With this online tool, users can perform many of the analyses shown in this article without having to install or write code in *R*.

This article only used the main features of *R Sequential* for planning and performing sequential analyzes with continuous, group, or mixed group-continuous Poisson and binomial data. Further features, explanations, and examples are available in the PDF user guide, which accompanies the R *Sequential* package. All calculations in this article were run using version 3.3.1 from http://CRAN.R-project.org/package=Sequential.

ACKNOWLEDGEMENTS

This research was funded by the National Institute of General Medical Sciences, USA, grant #RO1GM108999. Additional support was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, Brazil, process #301391/2019-0.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as the illustrative data examples are either fictitiously generated through Monte Carlo simulation or taken from first author's previous publications.

ORCID


Ivair R. Silva https://orcid.org/0000-0003-2701-8924

REFERENCES

1. Wald A. Sequential tests of statistical hypotheses. *Ann Math Stat*. 1945;16:117-186.
2. Wald A. *Sequential Analysis*. New York, NY: John Wiley and Sons; 1947.
3. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64:191-199.
4. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549-556.
5. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trails. *Biometrics*. 1987;43:193-200.
6. 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: Design Methods and Applications*. Vol 30; 2011;(1):58-78.
7. Li L, Kulldorff M. A Conditional maximized sequential probability ratio test for pharmacovigilance. *Stat Med*. 2009;29:284-295.
8. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. London, UK: Chapman and Hall; 2000.
9. Silva IR, Li L, Kulldorff M. Exact conditional maximized sequential probability ratio test adjusted for covariates. *Seq Anal*. 2019;38(1):115-133.
10. Silva IR, Lopes WM, Dias P, Yih WK. Alpha spending for historical versus surveillance Poisson data with CMaxSPRT. *Stat Med*. 2019;28(12):2126-2138.

11. Silva IR. Type I error probability spending for post-market drug and vaccine safety surveillance with Poisson data. *Methodol Comput Appl Probab*. 2018;20(2):739-750.
12. Silva I.R., Kulldorff M. Sequential: exact sequential analysis for Poisson and binomial data. R foundation for statistical computing-contributed packages. R Package Version 3.1. Vienna, Austria; 2019.
13. Silva IR, Kulldorff M, Yih WK. Optimal alpha spending for sequential analysis with binomial data. *J R Stat Soc Ser B*. 2020;82(4): 1141-1164.
14. Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of Covid-19; final report. *N Engl J Med*. 2020;383: 1813-1826.
15. Cao B, Wang Y, Wen D, et al. A trial of Lopinavir–Ritonavir in adults hospitalized with severe Covid-19. *N Engl J Med*. 2020;382(19):1787-1799.
16. CBER Surveillance program. COVID-19 vaccine safety surveillance: active monitoring master protocol; 2021. https://www.bestinitiative.org/wp-content/uploads/2021/02/C19-Vaccine-S%afety-Protocol-2021.pdf. [Online Accessed April 08, 2021].
17. Silva IR, Kulldorff M. Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. *Biometrics*. 2015;71(3):851-858.
18. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10:1-10.
19. Lin DY, Wei LJ, DeMets DL. Exact statistical inference for group sequential tests. *Biometrics*. 1991;47:1399-1408.
20. Causey BD. Exact calculations for sequential tests based on Bernoulli trials. *Commun Stat Simul Comput*. 1985;14:491-495.
21. Aroian LA. Sequential analysis, direct method. *Technometrics*. 1968;10:125-132.
22. Fireman B, Lee J, Lewis N, Bembom O, Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. *Am J Epidemiol*. 2009;170(5):650-656.
23. Silva IR, Kulldorff M, Gagne J, Najafzadeh M. Exact sequential analysis for multiple weighted binomial end points. *Stat Med*. 2020;39(3):340-351.
24. Grayling MJ, Wason JMS, Mander AP. Exact group sequential designs for two-arm experiments with Poisson distributed outcome variables. *Commun Stat Theory Methods*. 2021;50(1):18-34.
25. Kulldorff M, Silva IR. Continuous post-market sequential safety surveillance with minimum events to signal. *Revstat Stat J*. 2017;15: 1-21.
26. Gombay E, Li F. Sequential analysis: design methods and applications. *Seq Anal*. 2015;34:57-76.
27. 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.
28. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach(with disussion). *J Royal Stat Soc B*. 1989;51: 305-361.
29. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Stat Sci*. 1990;5: 299-317.
30. Silva IR. Type error 1 probability spending for post-market drug and vaccine safety surveillance with binomial data. *Stat Med*. 2018;37(1):107-118.
31. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663.
32. Stallard N, Todd S. Exact sequential tests for single samples of discrete responses using spending functions. *Stat Med*. 2000;19: 3051-3064.
33. Hwang IK, Shih WJ, DeCani JS. Group sequential designs using a family of type I error probability spending functions. *Stat Med*. 1990;9:1439-1445.
34. Yih WK, Kulldorff M, Sandhu SK, et al. Prospective influenza vaccine safety surveillance using fresh data in the sentinel system. *Pharmacoepidemiol Drug Saf*. 2016;25:481-492.
35. Grayling MJ, Wheeler GM. A review of available software for adaptive clinical trial design. *Clin Trials*. 2020;17(3):323-331.

How to cite this article: R. Silva I, Maro J, Kulldorff M. Exact sequential test for clinical trials and post-market drug and vaccine safety surveillance with Poisson and binary data. *Statistics in Medicine*. 2021;40:4890–4913. https://doi.org/10.1002/sim.9094

APPENDIX

Here we deliver the command lines for the calculations shown along the article using the *Sequential* package.

Command lines for Section 3

Table 1 was produced with command lines in similarity with the following:

```
SampleSize.Binomial(RR=c(2,4),alpha=0.05,
```

```
power=c(0.9,0.99),z=0.25,Tailed="upper")
```

Table 2 was produced with the following command line:

```
SampleSize.Poisson(alpha=0.05,power=c(0.9,0.99),M=1,D=0,
RR=c(2,4),Tailed="upper")
```

The following command line exemplifies how Table 3 was produced:

```
SampleSize.CondPoisson(cc=50,D=0,alpha=0.05,
power=c(0.9,0.99),RR=2)
```

Table 4 was produced with the following command line and outputs:

```
res<- Performance.Threshold.Poisson(SampleSize=90,
CV.lower=c(2.5,2.6,2.7,2.8),CV.upper=c(3,3.1,3.2,3.3)
GroupSizes=c(25,20,20,25),Tailed="two",
Statistic="MaxSPRT",Delta="n",RR=c(0.3,0.9,1,1.2,1.5))
res
$AlphaSpend_lower
[1] 0.006467484 0.006467484 0.006467484 0.006467484
$AlphaSpend_upper
[1] 0.00569632 0.01033923 0.01306292 0.01533008
$events_lower
[1] 14 30 47 68
$events_upper
[1] 39 63 87 116
$Performance
RR Power ESignalTime ESampleSize
[1,] 0.3 0.97843535 25.00000 26.40170
[2,] 0.9 0.02376536 25.87450 88.47603
[3,] 1.0 0.02179756 41.01881 88.93233
[4,] 1.2 0.29882956 57.76333 80.36673
[5,] 1.5 0.96483674 42.79107 44.45109
```

Critical values calculation in Section 3.2:

```
cv1<- 0; cv2<- 10; cvm<- (cv1+cv2)/2; alpha<- 0.05; alphain<- 0; count<- 0;
aux<- log(10/(10^(-6)))/log(2)
while(abs(alpha-alphain)>10^6)&count<aux){
count<- count+1
alphain<- Performance.Threshold.Poisson(SampleSize=20,
CV.upper=cvm,GroupSizes=rep(1,20),Tailed="upper",
Statistic="Pocock",RR=1)$Performance[[2]]
if(alphain>alpha)cv1<- cvmelse{cv2<- cvm}; cvm<- (cv1+cv2)/2}
resPocock<- Performance.Threshold.Poisson(SampleSize=20,
CV.upper=cvm,GroupSizes=rep(1,20), Tailed="upper",
Statistic="Pocock",RR=c(1,2))
```

The object `resPocock$AlphaSpend` contains the alpha spending implied by Pocock's test.

Figure 1 and Table 6 were obtained with the command lines similar to:

```
resM<- Performance.AlphaSpend.Poisson(SampleSize=20, alpha=0.05,RR=2,
alphaSpend=1, rho=0.5,gamma="n",Statistic="MaxSPRT",
Delta="n",Tailed="upper")
Performance.Threshold.Poisson(SampleSize=20,CV.lower="n",
CV.upper=resM$cvs,GroupSizes="n", Tailed="upper",Statistic="MaxSPRT",
Delta="n",RR=c(1,2))
```

The solutions for constructing Figure 3 can be obtained with the following command lines:

```
Optimal.Binomial(Objective="ETimeToSignal",N="n",z=1,
alpha=0.05,power=0.8,RR=2,GroupSizes="n",Tailed= "upper")
Optimal.Binomial(Objective="ESampleSize",N="n",z=1,
alpha=0.05,power=0.8,RR=2,GroupSizes="n",Tailed= "upper")
```

The content of Figure 4 was calculated with the following command lines:

```
Optimal.Binomial(Objective="ETimeToSignal",N="n",z=1,
p="n",alpha=0.05,power=c(0.8,0.8),RR=c(0.5,2),
GroupSizes="n",Tailed= "two")
```

For calculating the optimal solution of Section 3.3:

```
Optimal.Binomial(Objective="ETimeToSignal",N="80",z=1,
alpha=0.05,power=0.8,=2,GroupSizes="n",Tailed= "upper")
```

Command lines for Section 4

The lower and upper critical values presented in Table 8 of Section 4.2 can be reproduced with the following command lines:

```
AnalyzeSetUp.wBinomial(name="Treatments_AxB",N=1000,
alpha=0.05,M=1,rho=0.5,
title="Treatment A vs Treatment B comparison",
address="C:/Users/Example",Tailed="two")
Analyze.wBinomial(name="Treatments_AxB",test=1,
z=c(1,1,1,1,1),w=c(0.05,0.08,0.09,0.11,0.3),
ExposureA=c(0,0,0,0,1,0),ExposureB=c(0,0,0,0,0,0))
```

The critical values in the scale of the number of events and related alpha spending in Section 4.3 were obtained with command lines in similarity to:

```
AnalyzeSetUp.Poisson(name="PediarixVaccine",SampleSize=20,
alpha=0.05,M=1,AlphaSpendType="power-type",rho=0.5,
title="Pediarix vaccination",address="C:/Users/Example")
Analyze.Poisson(name="PediarixVaccine",test=1,mu0=0.04,
events=0)
Analyze.Poisson(name="PediarixVaccine",test=2,mu0=0.06
,events=1)
⋮
Analyze.Poisson(name="PediarixVaccine",test=10,mu0=0.12
,events=0)
```

The calculations in Section 4.4 were obtained with the following command lines:

```
Performance.AlphaSpend.CondPoisson(K=20,cc=37,alpha=0.05,
AlphaSpend=1,GroupSizes="n",rho=1.5,gamma="n",
Tailed="upper",RR=2)
AnalyzeSetUp.CondPoisson(name="INFLUENZA",
SampleSizeType="Events",K=20, cc=37,alpha=0.05,
M=1,AlphaSpendType="power-type",rho=1.5,
title="n",address="C:/Users/Example")
Analyze.CondPoisson(name="INFLUENZA",test=1,events=1,
PersonTimeRatio=0.044)
Analyze.CondPoisson(name="INFLUENZA",test=2,events=5,
PersonTimeRatio=0.046)
Analyze.CondPoisson(name="INFLUENZA",test=3,events=3,
PersonTimeRatio=0.025)
Analyze.CondPoisson(name="INFLUENZA",test=4,events=3,
PersonTimeRatio=0.024)
```

Command lines for the relative risk estimates in Table 7

```
MLE_R<- function(cases,SampleSizes,z)
  {
  # cases: number of new adverse events from Treatment B in each test until the ith
test.
  # SampleSizes: number (A+B) of new adverse events in each test until the ith test.
```

```
  # z: matching ratio in each test until the ith test.
  recand<- matrix(seq(0.01,10,0.01)"1)
  lr<- function(rr){
  return(prod(choose(SampleSizes,cases)*((1/(1+z/rr))^(cases))*((1-1/(1+z/rr))^
(SampleSizes-cases))))
  }
  veccand<- apply(recand,1,lr)
  Rhat<- seq(0.01,10,0.01)[veccand==max(veccand)]
  return(Rhat)
  }
  Rparciais<- rep(0,length(cases))
  for(i in 1:length(cases)){
  casesh<- cases[1:i]
  SampleSizesh<- ni[1:i]
  zh<- z[1:i]
  Rparciais[i]<- MLE_R(casesh,SampleSizesh,zh)
  }
```