Published in final edited form as:

Biometrics. 2015 September ; 71(3): 851–858. doi:10.1111/biom.12324.

Continuous Versus Group Sequential Analysis for Post-Market Drug and Vaccine Safety Surveillance

I. R. Silva^{1,2,*} and M. Kulldorff¹

¹Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts 02215, U.S.A

²Department of Statistics, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil

Summary

The use of sequential statistical analysis for post-market drug safety surveillance is quickly emerging. Both continuous and group sequential analysis have been used, but consensus is lacking as to when to use which approach. We compare the statistical performance of continuous and group sequential analysis in terms of type I error probability; statistical power; expected time to signal when the null hypothesis is rejected; and the sample size required to end surveillance without rejecting the null. We present a mathematical proposition to show that for any group sequential design there always exists a continuous sequential design that is uniformly better. As a consequence, it is shown that more frequent testing is always better. Additionally, for a Poisson based probability model and a flat rejection boundary in terms of the log likelihood ratio, we compare the performance of various continuous and group sequential designs. Using exact calculations, we found that, for the parameter settings used, there is always a continuous design with shorter expected time to signal than the best group design. The two key conclusions from this article are (i) that any post-market safety surveillance system should attempt to obtain data as frequently as possible, and (ii) that sequential testing should always be performed when new data arrives without deliberately waiting for additional data.

Keywords

Exact sequential analysis; Expected time to signal; Post-market safety surveillance; Uniformly better sequential design

1. Introduction

In prospective post-market drug and vaccine device safety surveillance, the goal is to detect serious adverse reactions as early as possible without too many false alarms. Sequential statistical methods allow investigators to repeatedly analyze the data as it accrues, while ensuring that the probability of falsely rejecting the null hypothesis at any time during the

*; Email: ivairest@gmail.com

6. Supplementary Materials

Web Appendices, and Figures referenced in Sections 2, 4.2, 4.2.1, and 4.3 and details about how to reproduce the calculations using the R package “Sequential” are available with this paper at the *Biometrics* website on Wiley Online Library.

surveillance is controlled at the desired nominal significance level (Wald, 1945, 1947). Prospective post-market vaccine safety surveillance has been conducted using sequential statistical analysis for most newly approved vaccines in order to detect potential adverse vaccine reactions that were too rare to find during phase 3 clinical trials (Lieu et al., 2007; Yih et al., 2009; Belongia et al., 2010; Klein et al., 2010; Yih et al., 2011). It is then important to know what sequential analysis designs works best. In particular, one hotly debated issue has been whether to perform sequential tests as soon as data arrives or whether it can be advantageous to delay testing in order to improve statistical power. This article answers that question.

Sequential statistical analysis can broadly be categorized as continuous or group sequential methods (Jennison and Turnbull, 2000). The former allows the investigator to perform a test as often as the investigator desires, including continuous monitoring. With group sequential methods, the data is analyzed at regular or irregular discrete time intervals after a group of subjects enter the study. Group sequential statistical methods are commonly used in clinical trials, where a trial may be stopped early due to either efficacy or unexpected adverse events (Jennison and Turnbull, 2000; Proschan, Lan, and Wittes, 2006). There is a large amount of literature comparing different group sequential methods in this context, but the interest in sequential analysis for post-market safety surveillance is more recent (Abt, 1998; Davis et al., 2005; Lieu et al., 2007; Shih et al., 2010; Kulldorff et al., 2011; Yih et al., 2011), and there have been few comparative evaluation studies (Nelson et al., 2012; Zhao et al., 2012; Kulldorff and Silva, 2015).

For a standard non-sequential statistical analysis we are concerned about the type I error probability, the statistical power and the sample size. In sequential analysis, the interest is in two aspects of the sample size: the expected number of events when the null hypothesis is rejected (expected time to signal), and the expected number of events, under the null hypothesis, to interrupt the surveillance when the null is not rejected (maximum length of surveillance). In the rest of the article, we will often use the shorter term in parenthesis in place of the longer more formal definition.

There are two key differences between clinical trials and post-market safety surveillance that require a different approach to sequential analysis. In clinical trials, it is often expensive to increase the maximum sample size since it may be expensive to recruit new patients to the study. Hence, the maximum sample size is a key design criterion. Expected time to signal when the null is rejected may be less important, since the number of people taking the drug/vaccine is limited in the pre-market setting. In contrast, it is typically easy to increase the maximum sample size in post-market safety surveillance. Once the surveillance system is up and running, it is typically easy and cheap to run the system for a few additional months. The expected time to signal is instead a much more important criterion, since there are many people exposed to the drug/vaccine in the post-market setting, and only a few of them may be part of the surveillance system (Kulldorff, 2012).

When the alpha level and the maximum sample size are held fixed, continuous sequential analysis has by default less statistical power than group sequential analysis, which in turn has less statistical power than a standard non-sequential analysis. This fact together with

published computer simulations studies (Zhao et al., 2012; Nelson et al., 2012) have led some to conclude that it can be advantageous to use a group sequential design even when the data is available continuously, or, that it can be advantageous to use a group sequential design with fewer looks at the data even though the data arrives more frequently. In fact, that is never the case.

In this article, we first present a mathematical proposition to prove that for any group sequential analysis design, with irregular or regularly spaced analyses and with any stopping boundary, there always exists a continuous sequential analysis design that is at least as good with respect to (i) type I error probability, (ii) statistical power, (iii) expected time to signal and (iv) maximum sample size. This result is very general in that it holds for a wide variety of sequential designs with different probability distributions and sequential designs, including Poisson data with historical or concurrent controls as well as binomial data with concurrent or self controls.

To allow users to balance the better performance of more frequent data feeds with the additional financial cost that may incur, we have compared continuous sequential analysis with different parameter settings versus group sequential analysis with different frequency of testing. We do this (i) for a Poisson probability model with observed and expected counts; (ii) with a Wald type upper rejection boundary that is flat with respect to the likelihood ratio; (iii) without a lower acceptance boundary; and (iv) with some upper limit on the length of surveillance at which time the sequential analysis ends without rejecting the null hypothesis. While this is only one special case of many possible sequential designs, it will give a general understanding and intuition regarding the performance lost by having less frequent data feeds. To allow for easy comparison, we fix the two most important performance characteristics: the type I error probability and statistical power. We then compare performance with respect to expected time to signal and maximum sample size. Exact calculations were possible by using the package “Sequential” (Silva and Kulldorff, 2013), programed in the R language. The article ends with a discussion.

2. Optimality of Continuous over Group Sequential Designs

In this section, we show that for any group sequential design there is always a continuous sequential design that is at least as good, and for Poisson type data, there is always a continuous sequential design that is better. We first define the notion of a uniformly better sequential design in terms of four key performance characteristics. Let X_t be a non-negative integer valued stochastic process describing the number of adverse events that occur during time $[0, t]$ time window.

Definition 1. (Group Sequential Analysis)

For a set of constants $A_1, \dots, A_G$, and a sequence $\{t_i\}_{i=1}^G$ of times, a group sequential analysis design is any procedure that rejects the null hypothesis if $X_{t_i} \geq A_i$ for some $i \in [1, \dots, G]$.

Definition 2. (Continuous Sequential Analysis)

For a function $B(t)$, a continuous sequential analysis design is any procedure that rejects the null hypothesis if $X_t \geq B(t)$ for some $0 < t \leq L$.

Definition 3. (Uniformly Better Sequential Design)

Let D_1 and D_2 be two sequential analysis designs. For D_j , denote the vector with the performance characteristics by $(\alpha_j, \beta_j, E[S_j], E[L_j])$, where α_j is the probability of type I error (alpha level, e.g. 0.05); β_j is the probability of type II error (the statistical power = $1 - \beta_j$); S_j is the random variable representing the sample size when the null hypothesis is rejected (expected time to signal), and L_j is the sample size at the time when the surveillance ends without the null hypothesis being rejected (maximum sample size). L_j may either be a random variable or a constant. The sequential design D_1 is at least as good as the sequential design D_2 if $\alpha_1 \leq \alpha_2, \beta_1 \leq \beta_2, E[S_1] \leq E[S_2]$, and $E[L_1] \leq E[L_2]$. If at least one of the four inequalities is a strict inequality, then D_1 is uniformly better than D_2 .

In words, one sequential design is uniformly better than a second one if it is at least as good on all the four characteristics defined above and if it is better on at least one of them.

Proposition 1

For any non-decreasing stochastic process X_t indexed by continuous or discrete time, and for any group sequential design that rejects the null for large values of X_t , there always exists a continuous sequential design that is at least as good. If X_t follows a Poisson distribution and if there exists an i and $t_i < m < t_i + 1$ such that $E[X_m] - E[X_{t_i}] > 0$, then there exists a continuous sequential design which is uniformly better.

The last inequality states that there is at least one instance in which data arrives in between the group sequential looks. To prove the Proposition, we construct a continuous sequential design that is identical to the group sequential except that it looks at the data in between the group sequential looks, and it rejects the null hypothesis as soon as we have seen the number of adverse events that are needed to reject the null at the next group sequential test. Since the number of events is non-decreasing, the type I error probability, power and maximum sample size will be unchanged, at the same time as the expected time to signal is smaller. This may seem like a trivial observation, and we think that it is, but it is also fundamentally important in that it refutes a belief among some that it can be advantageous not to conduct a test every time new data arrives in order to reduce the number of group sequential tests and increase statistical power.

Proof—Based on the group sequential design, consider the continuous sequential design where $L = t_G$, $t_0 = 0$, and $B(t) = A_r$ for $t \in (t_{i-1}, t_i]$. This continuous sequential design rejects H_0 if and only if the group sequential does. Because this assertion holds whether H_0 is true or false, the group and the continuous sequential designs have the same type I error probability and the same statistical power. Also, since $L = t_G$, they also end the surveillance at the same time when the null is not rejected. Now, let S_g be the random variable at which time the continuous sequential design rejects the null hypothesis. That is, it is the minimum t such that $X_t \geq B(t)$. Let S_g be the random variable at which time the group sequential design

rejects the null hypothesis. That is, it is the smallest t_i for which $X_{t_i} \geq A_i$. By construction, $S_c \leq S_g$ for any realization of the stochastic process X_t and hence, $E[S_c] \leq E[S_g]$, implying that there is always a continuous design that is at least as good as any group sequential design. Moreover, if there exists an $t_i < m < t_{i+1}$ such that $E[X_m] - E[X_{t_i}] > 0$, and since X_t is a Poisson process, we know that $P(X_m - X_{t_i} > k) > 0$ for any value of k . This means that the probability that the continuous sequential analysis rejects the null hypothesis at time m is $P(X_{t_i} < A_i, X_m \geq A_{i+1}) > 0$, which means that $E[S_c] < E[S_g]$. $\square$

This proof provides a mechanism for how to design a uniformly better continuous sequential test given a pre-defined group sequential design. Note that even the non-sequential test, $G = 1$, can be improved by applying the reasoning above. This constructive approach may not necessarily be the preferable continuous design though. It is likely that there are better continuous designs, in terms of time to signal, than those that are constructed directly from a group sequential design.

The inclusive inequality on the performance characteristics holds for any non-decreasing stochastic process, so X_t may be distributed according to any non-decreasing random variable and there can be any form of dependence between X_t and X_s .

The strict inequality does not hold for any non-decreasing stochastic process because it depends on shape of the threshold constants $A_1, \dots, A_G$. For example, assume that X_t follows a binomial distribution with t trials, that is, discrete time, and, under the null hypothesis, a success probability equal to 0.5. Consider a hypothetical group design with $G = 2$ and having thresholds as following: $t_1 = 20$, $t_2 = 25$, $A_1 = 15$, and $A_2 = 19$. Under the null, the probability of stopping the monitoring at t_1 is 0.0207, and this probability at t_2 is 0.0012, what gives an overall type I error probability of 0.0219. We see that, if the null is not rejected at time 20, then it is not possible to stop the monitoring before time 25 because, once X_{20} is at most 14, then there is no chance of having 19 successes before time 25. Thus, a continuous design following the steps in the proof of the Proposition will have the same performance than the reference group design.

It is important to emphasize that the strict inequality holds for other distributions than the Poisson if: (i) there are i and l , with $t_i < l < t_{i+1}$, such that $P(X_l > A_{i+1}) > 0$, and (ii) the test statistic X_t is non-decreasing with t . For a friendly example, consider the binomial design of the last paragraph again. If instead of 19 we had $A_2 = 18$, then there would exist a chance to stop the monitoring earlier at time 24. Thus, a quite general conclusion can be stated as following.

Remark 1

For any group sequential design, if X_t is a non-decreasing stochastic process, then there always exists a continuous sequential design which is at least as good. If there are i and l , with $t_i < l < t_{i+1}$, such that $P(X_l > A_{i+1}) > 0$, where $t_0 = 0$, then there exists a continuous sequential design which is uniformly better.

Observe that, in practice, the original random variable measured in the population, say Y_t , may not have a monotonous behavior with respect to t , but it is not rare that the test statistic,

X_t defined as a real valued function of Y_t , turns out to be non-decreasing in t . This can occur for some well-known sequential analysis methods as well including the Pocock's test (Pocock, 1977) and the O'Brien & Fleming's test (O'Brien and Fleming, 1979). For example, let $Y_1, Y_2, \dots$, be a sequence of independent and identically distributed observations, all with the same known variance σ^2 and some continuous distribution. Consider the one-sample Pocock's group sequential method for testing $H_0: E[Y_t] = \mu_0$ versus $H_A: E[Y_t] > \mu_0$. For simplicity, consider that a test will be performed after each chunk of m observations. For the i th test, Pocock's method uses the standardized statistic:

$$Z_i = \frac{1}{\sqrt{im\sigma^2}} \left(\sum_{j=1}^{im} Y_j - im\mu_0 \right) \\ = \frac{1}{\sqrt{im\sigma^2}} (X_{im} - im\mu_0), \text{ with } i=1, \dots, G.$$

The surveillance is interrupted to reject H_0 at the i th test if $Z_i > c_i$, where c_i is the critical

value for the i th test, i.e., stop the surveillance if $X_{t_i} > A_i = c_i + \frac{t_i\mu_0}{\sqrt{t_i\sigma^2}}$, where $t_i = im$ and $X_{t_i} = \sum_{j=1}^{t_i} Y_j$. Note that, in this example, t is discrete and represents the arrival order of the observations. The critical values c_i , $i = 1, \dots, G$, are obtained by assuming that Y_t follows a normal distribution. Suppose that, due to the nature of the problem, Y_t cannot assume negative values. Thus, X_t is non-decreasing with t and then a continuous design could be used without any loss in terms of the four performance characteristics of Definition 3, but probably with a certain gain in terms of time to signal.

In some problems the reverse occurs, that is the test statistic is not monotonous with t , but the original random variable is, by its turn, non-decreasing with t . Then the applicability of the Remark is trivial. This is shown for the Poisson process in the Web Appendix B, where the threshold of the uniformly better continuous design is illustrate using the test statistic scale. Based on the Proposition and Remark, we can also conclude that it is possible to insert additional tests in a pre-defined group sequential design in order to obtain another uniformly better group sequential design. This follows from the fact that one can use the group critical value A_{i+1} as the threshold associated to an arbitrary additional moment of testing I , where i is such that $I \in (t_i, t_{i+1})$, and $t_0 = 0$.

Corollary 1

If X_t is a non-decreasing stochastic process, and if there are i and l , with $t_i < I < t_{i+1}$, such that $P(X_I > A_{i+1}) > 0$, then for each group sequential design with tests at times $I = \{t_1, \dots, t_G\}$, there always exists a uniformly better group sequential design with tests at times $J = \{t_1, \dots, t_K\}$, where $I \subset J$. This is applicable, for example, when X_t is the Poisson process.

This constructive way of rewriting a group sequential test by using a continuous design is sometimes referred in the clinical trials literature as non-stochastic curtailment. For example, in the clinical trials context, Proschan et al. (2006, Section 3.2) describes the curtailment approach to interrupt the experiment at time t if it is impossible to obtain a result in favor to H_0 by the end of the trial. In general, the basic idea of a non-stochastic curtailment is to

terminate the surveillance as soon as the final result of the next test becomes inevitable. Although the curtailment strategy was not inspired by the necessity of discussing the comparison of group versus continuous designs, other authors have already pointed out the possibility of improving a group sequential design by considering a continuous sequential design like, for example, the works of Tan and Xiong (1996), Posch et al. (2004), Chi and Chen (2008) and Kunz and Kieser (2012). There is also the possibility of applying a stochastic curtailment for a early stopping when X_t follows a normal distribution. Some efficient methods are described by Jennison and Turnbull (2000, Chap. 10). Here, we do not explore the stochastic curtailment because our attention is directed to the cases where X_t follows a Poisson process, what is supported by the Proposition.

3. Sequential Analysis for Poisson Data

In this section, we briefly describe the maximized sequential probability ratio (MaxSPRT) test statistic (Kulldorff et al., 2011). An extension of the well known Sequential Probability Ratio Test (SPRT) proposed by Wald (1945), the MaxSPRT is both a “generalized sequential probability ratio test” (Weiss, 1953) and “sequential generalized likelihood ratio test” (Siegmund and Gregory, 1980; Lai, 1991). Unlike the standard SPRT, the MaxSPRT is defined for a composite rather than a simple alternative hypothesis. It was developed for the prospective rapid cycle vaccine safety surveillance implemented by the Centers for Disease Control and Prevention sponsored Vaccine Safety Datalink project (Lieu et al., 2007).

3.1. Continuous Sequential Analysis

Let C_t be the random variable that counts the number of patients who received the vaccine before time t and who had the adverse event between 1 and W days after receiving the vaccine. Let c_t be the corresponding observed number of patients with the adverse event. For rare adverse events, it is reasonable to model C_t as a Poisson process. Under the null hypothesis, C_t has a Poisson distribution with mean μ_t , reflecting a known background rate of the adverse event, adjusting for age, gender and other covariates. Under the alternative hypothesis, C_t has a Poisson distribution with mean $RR\mu_t$, where RR is the unknown increased relative risk due to the vaccine. The MaxSPRT statistic is defined as (Kulldorff et al., 2011):

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

The argument that solves the last term of expression (1) is c_t/μ_t . Then

$$LR_t = e^{\mu_t - c_t} (c_t / \mu_t)^{c_t}, \text{ when } c_t \geq \mu_t, \text{ and } LR_t = 1, \text{ otherwise.} \quad (2)$$

Equivalently, the MaxSPRT can be defined in terms of the log likelihood ratio as:

$$LLR_t = (\mu_t - c_t) + c_t \log(c_t / \mu_t), \text{ when } c_t \geq \mu_t, \text{ and } LLR_t = 0, \text{ otherwise.} \quad (3)$$

In the continuous sequential surveillance approach, the LLR_t is monitored for all values of $t > 0$, and the surveillance ends with H_0 being rejected the first time when LLR_t is greater than a rejection boundary CV , or, when $\mu_t = T$, in which case the null is not rejected. T is defined a priori as the upper limit on the sample size, defined in terms of the expected number of events under the null hypothesis. In this definition, even a single adverse event can reject the null hypothesis if it occurs sufficiently early. An alternative version of the MaxSPRT requires a minimum number of adverse events, M , before one can reject H_0 , which can simultaneously increase the statistical power and decrease the expected time to signal (Kulldorff and Silva, 2015). The possibility of having a performance improvement through a delayed start strategy has already been considered by Simon (1989). In his study, Simon finds out that some of the best strategies for a group sequential testing are implied by a very small, but greater than 1, sample size for the first group sequential test. For the MaxSPRT method, exact critical values (CV), statistical power, expected time to signal and maximum sample size can be calculated using iterative numerical calculations (Kulldorff et al., 2011).

3.2. Group Sequential Analysis

Group sequential analysis is used in a wide variety of scientific areas, but its development has primarily been stimulated by clinical trials. The literature is vast. The strategy to define the groups is an important aspect of the group sequential design. In the articles by Pocock (1977) and O'Brien and Fleming (1979), a maximum number of groups, G , and group size, n , are fixed a priori. The analysis consists of comparing a test statistic, based on accumulating data, against a critical value, CV_i , after each group of k_i observations, $i = 1, \dots, G$, with CV_i chosen in a way to have the desired overall type I error probability. Group sequential methods have been extended in various directions in order to account for different group sizes, critical value functions, and to embrace different probability distributions for the outcome. An excellent review of group sequential methods for clinical trials has been written by Jennison and Turnbull (2000). The strategy to define the group sizes a priori is linked to clinical trial context, where the number of subjects and the time to perform the sequential tests are easy to control. This is not the case for post-market safety surveillance, where new batches of different size data from different data providers may arrive at different frequencies.

For group sequential analyses for Poisson data, we use the same definitions of C_b , c_b , μ_b , RR , LR_b , and LLR_t as in the prior section. The only difference is that for group sequential analysis, the LLR_t test statistic is only evaluated a finite number of times. This can be done using regular or irregular time intervals that may or may not be pre-defined before the sequential analysis commences. The group sizes are defined in terms of the sample size, expressed as the expected number of adverse events under the null hypothesis. In this comparative evaluation we use equally spaced tests with equal group sizes. In other words, LLR_t is compared against the critical value CV at each T/G time interval, where G is the maximum number of group sequential tests that will be performed. Surveillance ends when $LLR_t > CV$ for some t that is a multiple of T/G or when $\mu_t = T$. The exact CV is obtained through numerical calculations to ensure that the overall probability of type I error is less or

equal to α . While equal group sizes are unlikely to be used in practice for post-market safety surveillance, it serves as a good benchmark for methods evaluation.

Just as for the continuous MaxSPRT, it is possible to require a minimum number of M adverse events before rejecting H_0 when using group sequential analysis. Such a requirement does not improve the performance though unless the time between looks is very small. With true relative risks equal to 1.5, 2, 3, 4, and 5, and a large set of values for G and M , we verified that the group sequential test takes the smallest expected time to signal when M is equal to 1 (data not shown). The intuition behind this is that the group sequential approach already incorporates an inability to reject the null after only a few events, as a certain sample size is required at the first look. To illustrate, let $T = 50$, $\alpha = 0.05$, and $G = 20$, which implies a group size of $50/20 = 2.5$ expected adverse events under the null. When the expected count is 2.5, the number of adverse events needed to reject H_0 is 7. Thus, the statistical power and expected time to signal is the same for any value of $M \geq 7$. Therefore, in this article all group sequential analyses are performed using $M = 1$.

4. Continuous Versus Group Sequential Analysis

By considering Poisson type data, we use exact calculations to compare continuous and group sequential analysis with respect to statistical power, expected time to signal, and maximum sample size for the MaxSPRT statistic. Assuming a flat threshold in the scale of the likelihood function, we do this comparison for different values of the true relative risk, the number of group sequential looks (G) and the minimum number of events required to signal (M). For continuous sequential analysis, we used $M = 1, \dots, 10$. For group sequential analysis we used G equal to 2, 5, 10, 20, 50, 80, 100, and 200.

4.1. Fixed Maximum Sample Size

When the type I error probability and the maximum sample size are fixed, the statistical power is, by default, higher for group versus continuous sequential analysis, higher for group sequential analysis with fewer looks and the highest for a standard non-sequential analysis. For expected time to signal, the opposite ranking generally holds. Hence, for a fixed maximum sample size, the choice of continuous versus group sequential analysis is a trade-off between higher statistical power versus shorter expected time to signal. As we will see in the next section though, this is not a necessary trade-off for post-market safety surveillance.

4.2. Expected Time to Signal for Fixed Statistical Power

Clinical trials are often restricted in terms of the maximum sample size, due to the high cost of recruiting patients, and the focus on sequential analysis is typically to maximize statistical power within a limited maximum sample size while still having some ability to terminate the study early if needed. Post-market safety surveillance is very different. Once the surveillance system is up and running, it is easy and cheap to extend the surveillance for a few more months to achieve the desired statistical power. The only exceptions are products that are only used for a limited time, such as influenza vaccines, for which there is a new vaccine each season.

The best way to evaluate and compare sequential designs for post-market safety surveillance is to fix the alpha level and the statistical power, which are the two most important design criteria, and then compare methods in terms of the expected time to signal and the maximum sample size. Table 1 shows the expected time to signal for continuous sequential analysis with a minimum of $M = 1-7$ adverse events required to reject the null hypothesis, and for group sequential analysis with $G = 1, 2, 5, 20, 50, 100$, and 200 tests. This is done for different levels of the statistical power for the alternative hypothesis of $RR = 2$. This means that each entry in the table is based on different values of the maximum sample size T , chosen to provide the desired power. These maximum sample size values are evaluated in the next section. Note that for $G = 1$, there is just one look, which is equivalent to a standard non-sequential analysis.

From Table 1 we see that the method that signal the earliest is always found among the continuous sequential designs, with 4–6 adverse events as the minimum requirement to signal.

The sequential designs that have equal statistical power when $RR = 2$ do not necessarily have equal statistical power when $RR = 1.5$, even though both designs will have lower power for $RR = 1.5$ than for $RR = 2$. In order to extend the comparison, we also provide an additional figure placed in the Web Figure 1, which shows the expected time to signal as a function of the power for different relative risks. Continuous sequential analysis with a minimum of four adverse events required to signal performs well across the board.

4.2.1. Exact versus conservative type I error probabilities—Unlike the continuous MaxSPRT, the group sequential version will seldom have a critical value for which the probability of type I error is exactly equal to the desired α level. This is because of the discrete nature of the group sequential looks at the data. For example, for $T = 20$ and $G = 2$ the probability of type I error is 0.04995 when $CV = 1.520059$ while it jumps to 0.06713 for $CV = 1.520058$. Hence, a group sequential approach is almost always conservative with respect to the type I error probability. This may not be a major problem in practice, but when comparing various continuous and group sequential methods it is, as we do not know if differences in statistical power and time to signal are due to the different type I error probabilities rather than the continuous versus group sequential nature of the methods.

In order to ensure a fair comparison between the methods, we performed a study about the effects on the expected time to signal after ensuring the same type I error probability and power in both, continuous and group designs. For our investigations, we used $\alpha = 0.05$, $M = 1, \dots, 7$, $G = 1, 2, 5, 20, 50, 100$, and target statistical powers of 0.5, 0.8, and 0.9 for fixed true relative risk $RR = 2$. For each G , we found the value of T needed to achieve given power. Then, the exact type I error probability obtained from this combination of G and T was used to find out the value of T needed to achieve the same power in the continuous MaxSPRT design. For example, for $G = 5$ and a target power of 0.9, it is necessary to set $T = 14.7518$, what gives a type I error probability of 0.0457 and an expected time to signal equal to 7.0857. For $\alpha = 0.0457$ and $M = 6$, the value of T needed to ensure a power of 0.9 in the continuous MaxSPRT is 16.1551, what gives an expected time to signal equal to 5.9869. The results for the others scenarios are not shown here. But, even under this disadvantageous

significance level, the expected time to signal of the continuous design is in general better than the group approach. This corroborates with the results obtained in the last section. A step by step description of how to obtain the values described above are presented in the Web Appendix C. The script uses the functions of the R package “Sequential.”

4.3. Maximum Sample Size for Fixed Statistical Power

In this section we compare the continuous and group sequential methods with respect to the maximum sample size required to end the surveillance without rejecting the null hypothesis. Again, this is done while holding both the type I error probability and the statistical power fixed.

Table 1 presents the maximum sample size T as a function of different levels of statistical power when $RR = 2$ and for different sequential study designs. Irrespectively of the statistical power, the maximum sample size is minimized with a non-sequential design ($G = 1$). For group sequential designs the maximum sample size is generally smaller with fewer groups and for continuous sequential analysis it is always smaller with larger values of M , the minimum number of adverse events required to signal.

For $RR = 1.5, 2, 3, 4$, and 10 , G equal to $2, 5, 20$, and 50 for the group designs, and M equal to 1 and 4 for continuous sequential designs, we compared the maximum sample sizes between group and continuous. In the Web Figure 2 of the Web Appendix material we present a visual resume of the results. As expected, the group sequential designs with fewer looks at the data have the smallest maximum sample size, but for $M = 4$, especially for larger RR s, the difference is small. The continuous sequential design with $M = 1$ has the largest maximum sample size across irrespectively of the RR .

4.4. Expected Time to Signal Versus Maximum Sample Size

By keeping the significance level and the statistical power fixed, the choice of sequential design is a trade-off between the expected time to signal and the maximum sample size. To better understand this trade-off it is interesting to plot the two against each other. For $RR = 2$, this is done in Figure 1, with the maximum sample size on the x -axis and the expected time to signal on the y -axis. For continuous sequential analysis there is a parabolic relationship when the expected time to signal is expressed as a function of the maximum sample size. The minimum expected time to signal is reached when requiring around six adverse events in order to reject the null, depending on the statistical power. It is interesting to note that the curve for the continuous sequential analysis is consistently below the curve for the group sequential analysis.

Observe that the continuous sequential approach can always be improved in a way to present a smaller expected signal time and the same maximum length of surveillance than the non-sequential test. As already mentioned in Section 2, if A is the critical value in the scale of the number of events with the non-sequential test, then an uniformly better continuous design can be obtained by setting the signaling threshold equal to A , corresponds to set $M = A$. For $RR = 2$, the threshold A is equal to 7 for a power of 0.5 , and it is equal to 19 for a power of 0.9 . It means that whatever preference we have with respect to maximum sample size and expected time to signal, the better performance is found among the continuous sequential

designs. But, another implication of this fact is that, if we set a M value greater than A , which means a significance level smaller than α , then both, power and expected time to signal, are simultaneously reduced in comparison to $M = A$. Therefore, A is an upper bound for the choice of M .

5. Discussion

There has been some uncertainty as to whether group sequential analysis may be better than continuous sequential analysis for post-market safety surveillance, and whether statistical power could be increased by not doing a sequential test every time new data arrives (Nelson et al., 2012; Zhao et al., 2012). Providing both a general mathematical proof and numerical calculations for specific scenarios, we have shown that continuous sequential analysis performs better than group sequential analysis and that more frequent group sequential analyses perform better than less frequent group sequential analyses. Hence, group sequential analysis should never be deliberately applied to post-market safety surveillance when the data is available in a continuous or near continuous fashion. Moreover, the goal should always be to obtain post-market safety surveillance data as frequently as possible and new sequential tests should be performed as soon as new data arrives. Based on the Remark, this conclusion is not limited to the Poisson based MaxSPRT with a flat boundary with respect to the log likelihood ratio, but valid for any rejection boundary and probability distribution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by the United States Food and Drug Administrations Center for Biologics Evaluation and Research, through the Mini-Sentinel Post-Rapid Immunization Safety Monitoring (PRISM) program. Additional support was received from the National Institute of General Medical Sciences, grant #R01GM108999, from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and from Banco de Desenvolvimento do Estado de Minas Gerais (BDMG), Brazil. We wish to thank the reviewers and Editors for their valuable comments and suggestion that helped to improve the paper.

References

- Abt K. Poisson sequential sampling modified towards maximal safety in adverse event monitoring. *Biometrical Journal*. 1998; 40:21–41.
- Belongia E, Irving S, Shui I, Kulldorff M, Lewis E, Yin R, Lieu T, Weintraub E, Yih W, Li R, Baggs J, the Vaccine Safety Datalink Investigation Group. Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. *Pediatric Infectious Disease Journal*. 2010; 29:1–5. [PubMed: 19907356]
- Chi Y, Chen C. Curtailed two-stage designs in phase II clinical trials. *Statistics in Medicine*. 2008; 27:6175–6189. [PubMed: 18816510]
- Davis R, Kolczak M, Lewis E, Nordin J, Goodman M, Shay D, Platt R, Black S, Shinefield H, Chen R. Active surveillance of vaccine safety: A system to detect early signs of adverse events. *Epidemiology*. 2005; 16:336–341. [PubMed: 15824549]
- Jennison, V.; Turnbull, B. *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman and Hall/CRC; 2000.

- Klein N, Fireman B, Yih W, Lewis E, Kulldorff M, Ray P, Baxter R, Hambidge S, Nordin J, Naleway A, Belongia E, Lieu T, Baggs J, Weintraub E, the Vaccine Safety Datalink. Measles-mumps-rubella-varicella combination vaccine and risk of febrile seizures. *Pediatrics*. 2010; 126:e1–e8. [PubMed: 20587679]
- Kulldorff, M. *Pharmacoepidemiology*. 5th. Wiley-Blackwell; 2012. Sequential Statistical Methods for Prospective Postmarketing Safety Surveillance; p. 852–867.
- Kulldorff M, Davis RMK, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*. 2011; 30:58–78.
- Kulldorff M, Silva IR. Continuous post-market sequential safety surveillance with minimum events to signal. 2015; arxiv 1503.01978 [stat.ap].
- Kunz C, Kieser M. Curtailment in single-arm two-stage phase II oncology trials. *Biometrical Journal*. 2012; 54:445–456. [PubMed: 22610516]
- Lai T. Asymptotic optimality of generalized sequential likelihood ratio tests in some classical sequential testing problems. *Handbook of Sequential Analysis*. 1991; 21:121–144.
- Lieu T, Kulldorff M, Davis R, Lewis E, Weintraub E, Yih W, Yin R, Brown J, Platt R. Real-time vaccine safety surveillance for the early detection of adverse events. *Medical Care*. 2007; 45:89–95.
- Nelson J, Cook A, Yu O, Dominguez C, Zhao S, Greene S, Fireman B, Jacobsen S, Weintraub E, Jackson L. Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. *Pharmacoepidemiology and Drug Safety*. 2012; 21:62–71. [PubMed: 22262594]
- O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
- Pocock S. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199.
- Posch M, Timmesfeld N, Knig F, Mller H. Conditional rejection probabilities of Student's t-test and design adaptations. *Biometrical Journal*. 2004; 46:389–403.
- Proschan, M.; Lan, K.; Wittes, J. *Statistical Monitoring of Clinical Trials: A Unified Approach Series Statistics for Biology and Health*. New York: Springer; 2006.
- Shih M, Lai T, Heyse J, Chen J. Sequential generalized likelihood ratio tests for vaccine safety evaluation. *Statistics in Medicine*. 2010; 26:2698–2708. [PubMed: 20799244]
- Siegmund D, Gregory P. A sequential clinical trial for testing $p_1 = p_2$. *Annals of Statistics*. 1980; 8:1219–1228.
- Silva, I.; Kulldorff, M. *Sequential-Package*. Vienna, Austria: R Foundation for Statistical Computing-Contributed Packages; 2013.
- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
- Tan M, Xiong X. Continuous and group sequential conditional probability ratio tests for phase II clinical trials. *Statistics in Medicine*. 1996; 15:2037–2051. [PubMed: 8896138]
- Wald A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*. 1945; 16:117–186.
- Wald, A. *Sequential Analysis*. New York: John Wiley and Sons; 1947.
- Weiss L. Testing one simple hypothesis against another. *Annals of Mathematical Statistics*. 1953; 24:273–281.
- Yih W, Kulldorff M, Fireman B, Shui I, Lewis E, Klein N, Baggs J, Weintraub E, Belongia E, Naleway A, Gee J, Platt R, Lieu T. Active surveillance for adverse events: The experience of the Vaccine Safety Datalink Project. *Pediatrics*. 2011; 127:S54–S64. [PubMed: 21502252]
- Yih W, Nordin J, Kulldorff M, Lewisc E, Lieua T, Shia P, Weintraube E. An assessment of the safety datalink of adolescent and adult tetanus-diphtheria-acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the Vaccine Safety Datalink. *Vaccine*. 2009; 27:4257–4262. [PubMed: 19486957]

Zhao S, Cook A, Jackson L, Nelson J. Statistical performance of group sequential methods for observational post-licensure medical product safety surveillance: A simulation study. *Statistics and Its Interface*. 2012; 5:381–390.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

For a fixed significance level of 0.05 and statistical power when $RR = 2$, the expected time to signal is plotted against the maximum sample size. The bold labels refer to continuous sequential designs, with the numbers indicating the minimum number of adverse events required to reject the null hypothesis (M). The gray labels refer to group sequential designs with the numbers indicating the number of sequential tests conducted (G).

Table 1

Expected time to signal and maximum sample size (length of surveillance) for $RR = 2$ and different continuous and group sequential designs, with significance level of 0.05 and statistical power from 0.5 to 0.99

Power	Expected time to signal													
	Continuous sequential analysis							Group sequential analysis						
	M=1	2	3	4	5	6	7	G=1	2	5	20	50	100	200
0.50	2.51	2.25	2.12	2.08	2.11	2.24	2.49	3.83	3.16	2.55	2.29	2.27	2.26	2.23
0.60	3.23	3.06	2.89	2.81	2.79	2.85	2.99	6.28	4.06	3.63	3.11	3.02	3.00	3.02
0.70	4.26	3.95	3.76	3.65	3.60	3.62	3.69	7.31	5.69	4.67	4.05	4.06	3.88	3.89
0.80	5.34	4.99	4.77	4.65	4.57	4.56	4.60	9.62	6.66	5.39	5.05	4.90	4.90	4.93
0.85	5.92	5.59	5.37	5.22	5.15	5.12	5.14	11.19	8.24	6.22	5.51	5.46	5.64	5.68
0.90	6.53	6.27	6.04	5.90	5.81	5.77	5.78	12.95	8.96	7.09	6.15	6.04	6.17	6.17
0.95	7.48	7.11	6.88	6.73	6.64	6.60	6.60	16.29	11.26	8.18	7.14	7.03	6.90	6.97
0.98	8.15	7.78	7.56	7.41	7.32	7.28	7.27	21.74	13.37	9.67	8.01	7.55	7.52	7.82
0.99	8.40	8.09	7.87	7.73	7.64	7.59	7.59	24.51	14.69	10.19	8.37	7.97	7.86	7.92

Power	Maximum sample size (T)													
	Continuous sequential analysis							Group sequential analysis						
	M=1	2	3	4	5	6	7	G=1	2	5	20	50	100	200
0.5	6.00	5.29	4.84	4.43	4.26	3.93	3.63	3.83	4.23	4.41	4.84	5.15	5.21	5.22
0.6	7.69	7.22	6.74	6.39	6.12	5.76	5.40	5.24	5.61	5.98	6.60	6.78	7.05	7.13
0.7	10.30	9.68	9.16	8.72	8.31	8.04	7.74	7.31	7.73	8.09	9.15	9.28	9.27	9.31
0.8	13.54	12.68	12.14	11.66	11.41	11.04	10.64	9.48	9.62	10.82	11.18	11.90	12.08	12.51
0.85	15.56	14.79	14.24	13.74	13.39	13.09	12.69	11.19	11.65	12.31	13.19	13.57	14.59	14.58
0.9	18.23	17.64	17.06	16.57	16.11	15.84	15.49	12.95	14.38	14.74	16.06	16.12	16.69	17.15
0.95	22.92	22.18	21.58	21.07	20.60	20.29	19.95	16.29	17.88	18.11	19.70	21.22	21.09	21.50
0.98	28.75	27.76	27.15	26.61	26.10	25.88	25.46	21.74	22.19	23.56	24.97	25.44	26.09	27.28
0.99	32.77	31.80	31.18	30.62	30.29	29.90	29.45	24.51	25.55	28.25	28.72	29.66	30.49	30.67