Mini-Sentinel Statistical Methods Development

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

Prepared by: Ivair R Silva, PhD [1,2], Martin Kulldorff, PhD [1]

Author Affiliations: 1. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA, 2. Department of Statistics, Universidade Federal de Ouro Preto, Ouro Preto, Minas Gerais, Brazil.

January 2014

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multifaceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Statistical Methods Development

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

Table of Contents

Abstract

Using observational electronic health data, sequential statistical analysis is commonly used for prospective post-market vaccine safety surveillance, and its use 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 1 error; statistical power; the expected time to signal when the null hypothesis is rejected; and the sample size required to end surveillance without rejecting the null. Presenting a theorem, we first show that for any group sequential design there always exists a continuous sequential design that is uniformly better in the sense that it is at least as good with respect to all four criteria and better for at least one. As a corollary, it is shown that more frequent testing is always better and one should never deliberately postpone sequential testing when a new batch of data arrives.

The theorem does not state that every continuous sequential design is better than every group sequential design. Moreover, when more frequent data feeds are more costly, it is important to know how much statistical performance is lost by using less frequent group sequential designs versus continuous or more frequent group sequential designs. 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. When type 1 error, statistical power and maximum sample size are held constant, there was always a continuous sequential analysis design with shorter expected time to signal than the best group sequential design. Results are based on exact calculations.

The two key conclusions from this paper 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: Pharmacovigilance; Post-market safety surveillance; Exact sequential analysis; Expected time to signal.

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 surveillance is controlled at the desired nominal significance level (Wald, 1945, 1947). Using sequential statistical analysis, prospective post-market vaccine safety surveillance has been conducted 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). Sequential analysis has only rarely been used for post-market drug safety surveillance (Brown et al., 2007; Avery et al., 2010), but the Food and Drug Administration (FDA) is planning to greatly increase those efforts through its Mini Sentinel Initative(Platt et al., 2012). 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 paper answers that question.

Sequential statistical analysis can broadly be categorized as continuous or group sequential methods (Jennison and Turnbull, 1999). 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, 1999). 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 only been few comparative evaluation studies (Nelson et al., 2012; Zhao et al., 2012; Silva and Kulldorff, 2012).

For a standard non-sequential statistical analysis we are concerned about the type 1 error (alpha level), the statistical power and the sample size. In sequential analysis, the interest is in two aspects of the sample size: the expected sample size when the null hypothesis is rejected (expected time to signal), and the expected sample size when the null is not rejected (maximum sample size). In the rest of the paper, 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 lead 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 paper we first present a mathematical theorem that states that for any group sequential analysis design, with irregular or regularly spaced analyses and with any stopping boundary, there always exist a continuous sequential analysis design that is at least as good with respect to (i) type 1 error, (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. Moreover, for Poisson type data there

always exists a continuous sequential design that is uniformly better, by which we mean that at least one of these four criteria is better while the other three are at least as good. For example, we may have shorter expected time to signal while the alpha level, statistical power and maximum sample size are the same. As a corollary to the theorem, we also show that if we compare two group sequential designs with different number of looks at the data in such a way that one set of looks is a subset of the other, then the design with the most looks is uniformly better.

The theorem should not be interpreted to mean that every continuous sequential design is better than every group sequential design, which is not true. Neither does the theorem help us determine if a particular continuous sequential design is better than a particular group sequential design, nor which continuous sequential design is the best one to use. It simply states that the best possible design is found among continuous sequential designs, and that every group sequential design can be replaced with a better or equally good continuous sequential design. Does this mean that group sequential designs should never be used for post-market safety surveillance? The answer is no. While CDC's Vaccine Safety Datalink receives weekly near-continuous data feeds for its sequential post-market vaccine safety surveillance (Yih et al., 2011), other systems may not be as timely, and there can be logistical or financial reasons why post-market safety surveillance must be based on monthly, quarterly or annual data feeds. In such a less timely post-market safety surveillance system, group sequential analysis should sometimes be used.

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 1 error and statistical power. We then compare performance with respect to expected time to signal and maximum sample size. The comparisons are made using exact calculation of the performance metrics rather than asymptotic theory or computer simulations.

The paper is organized as follows. We first show that for any group sequential design, there is always a continuous design that is as good or better. The next section describes continuous sequential analysis with the Poisson based Maximum Sequential Probability Ratio Test (Kulldorff et al., 2011). Section 3.2 defines an equivalent likelihood ratio based group sequential design and proposes a randomized adjustment in order to obtain the nominal alpha level exactly and hence ensure a fair comparison between the two designs. Holding both the alpha level and the statistical power fixed, Section 4 compares the two methods in terms of expected time to signal and maximum sample size. An application of group and continuous sequential designs to mimic a safety monitoring after Pediarix vaccination is provided in Section 5. The paper ends with a discussion.

2 Optimality of Continuous over Group Sequential Designs

In this section we show that for any groups 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. (Group Sequential Analysis) For a set of constants $A_1, ..., A_G$, 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 $i \in [1, ..., G]$.

Definition. (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 $0 < t < L$.

Definition. (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_i 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.

Theorem. *For any non-decreasing stochastic process X_t taking non-negative integer values and 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 theorem, 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 1 error, 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 $B_t = A_i$, 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 1 error 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_c 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 $E[S_c] < E[S_g]$. □

The inclusive inequality on the performance characteristics holds for any non-decreasing stochastic process, so X_t may be distributed according to any non-negative probability distribution 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, but it does not only hold for the Poisson distribution but for many other distributions as well including the gamma and the normal. It is also worth to note that the theorem is also valid for the cases where there exists a lower boundary to stop the monitoring in order to accept the null, such as the triangular continuation regions proposed by Whitehead and Stratton (1983). The continuous sequential test constsructed in the proof would simply have exactly the same lower boundary as the reference group sequential design.

The proof provides a mechanism for how to design a uniformly better continuous sequential test given a pre-defined group sequential design. That may not necessarily be the best continuous design though. For a group sequential test, the sequence of critical value numbers $K_1, ..., K_G$ that are needed to reject the null hypothesis will typically increase by two or more at each time point when a new test is conducted. An optimal continuous design is more likely to have a gradually increasing critical value where the critical value needed to reject increases by one unit at a time.

Based on the theorem and proof, we can also conclude that it is always 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 K_i as the threshold associated to an arbitrary additional moment of testing t^*, where i is such that $t^* \in (t_{i-1}, t_i]$.

Corollary 1. *If X_t is a Poisson process, then for each group sequential design with tests at times $I = \{t_1, ..., t_G\}$, there always exists a uniformly better group sequential design with tests at times $J = \{\tau_1, ..., \tau_{K+r}\}$, where $I \in J$.*

3 Sequential Analysis for Poisson Data

3.1 Continuous Sequential Analysis

In this section we briefly describe the maximized sequential probability ratio (MaxSPRT) test statistic (Kulldorff et al., 2011). An extension 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).

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 to 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 has 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

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

when $c_t \geq \mu_t$, and $LR_t = 1$, otherwise.

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

$$LLR_t = (\mu_t - c_t) + c_t \log(c_t/\mu_t), \quad (2)$$

when $c_t \geq \mu_t$, and $LLR_t = 0$, otherwise.

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 (Silva and Kulldorff, 2012).

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 papers 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, $i = 1, ..., G$, with CV_i chosen in a way to have the desired overall type one error. 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 (Jennison and Turnbull, 1999). 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_t, c_t, μ_t, RR, LR_t 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 another 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 $LLR_t \geq CV$ 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 $M \leq 7$. Therefore, in this paper all group sequential analyses are performed using $M = 1$.

	Critical Value		Type 1 Error		Statistical Power	
T	CV_c	CV_l	$\alpha(CV_c)$	$\alpha(CV_l)$	$Pow(CV_c)$	$Pow(CV_l)$
1	1.295838	1.295837	0.02665	0.08030	0.165	0.323
1.5	1.423318	1.423317	0.04897	0.08133	0.265	0.381
2	1.581455	1.581454	0.02897	0.05829	0.256	0.383
2.5	1.752813	1.752812	0.04463	0.06290	0.332	0.428
3	1.423318	1.423317	0.04217	0.08024	0.416	0.487
4	1.581455	1.581454	0.03102	0.06195	0.438	0.506
5	1.752813	1.752812	0.03820	0.05928	0.558	0.597
6	1.667495	1.667494	0.04426	0.06061	0.598	0.681
8	1.545178	1.545177	0.04532	0.06869	0.743	0.770
10	1.520059	1.520058	0.04787	0.06464	0.807	0.856
12	1.667495	1.667494	0.04736	0.06412	0.880	0.890
15	1.650603	1.650602	0.04455	0.06094	0.926	0.933
20	1.520059	1.520058	0.04995	0.06713	0.974	0.976
25	1.776529	1.776528	0.04378	0.05232	0.987	0.991
30	1.807363	1.807362	0.04537	0.05286	0.995	0.997
40	1.775344	1.775343	0.04117	0.05102	1.000	1.000
50	1.776529	1.776528	0.04479	0.05349	1.000	1.000

Table 1: Conservative CV_c and liberal CV_l critical values, type 1 errors and statistical power for group sequential analysis with $G = 2$ and $RR = 2$.

3.3 Exact versus Conservative Type 1 Error

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 1 error. 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 1 error rather than the continuous versus group sequential nature of the methods.

In order to ensure a fair comparison between the methods, we used randomization to create group sequential analyses with the exact correct type 1 error. By using exact calculations, we first found the liberal (CV_l) and the conservative (CV_c) critical values. For $\alpha = 0.05$, $G = 2$, $RR = 2$, and several values of *T*, these are provided in the second and third columns of Table 1. As a second step, we calculated the corresponding probability of type I error for these two *CV*s, which are in the fourth and fifth columns. The statistical power in the last two columns clearly show how the performance is affected.

The conservative critical value leads to a test size smaller than α, and the liberal leads to a size larger than α. To get a test with the desired type 1 error one can randomly select either the liberal or the conservative *CV*. While we do not recommend this approach for actual surveillance, it is appropriate for a comparative methods evaluation, as it ensures that all methods have exactly the same alpha level. Let Y be a random variable distributed according to a Bernoulli probability distribution with parameter

$$\theta = \frac{\alpha - \alpha(CV_c)}{\alpha(CV_l) - \alpha(CV_c)},$$

	G=2		5		10	
T	Cons	Rand	Cons	Rand	Cons	Rand
1	0.165	0.234	0.201	0.223	0.201	0.215
1.5	0.265	0.269	0.238	0.259	0.261	0.266
2	0.256	0.347	0.317	0.322	0.293	0.301
2.5	0.332	0.360	0.336	0.367	0.336	0.344
3	0.416	0.431	0.344	0.369	0.379	0.379
4	0.438	0.480	0.475	0.487	0.472	0.483
5	0.558	0.580	0.525	0.547	0.503	0.513
6	0.598	0.627	0.605	0.615	0.539	0.555
8	0.743	0.749	0.696	0.698	0.702	0.708
10	0.807	0.813	0.776	0.776	0.764	0.767
12	0.880	0.882	0.851	0.855	0.830	0.834
15	0.926	0.928	0.911	0.913	0.889	0.889
20	0.974	0.974	0.954	0.956	0.956	0.956
25	0.987	0.990	0.984	0.987	0.979	0.980
30	0.995	0.996	0.993	0.994	0.994	0.994
40	1.000	1.000	0.999	0.999	0.999	0.999
50	1.000	1.000	1.000	1.000	1.000	1.000

Table 2: Comparison of the statistical power between the conservative (Cons) and randomized (Rand) group sequential approaches for $RR = 2$, and for $G = 2, 5$ and 10 equally spaced group sequential tests.

where $\alpha(CV_l)$ and $\alpha(CV_c)$ are the type 1 error probabilities of the tests with CV_l and CV_c, respectively.

For a given observed value y of Y, the following randomized critical value is then constructed:

$$CV_r = \begin{cases} CV_l, & \text{if } y = 1, \\ CV_c, & \text{if } y = 0\,. \end{cases} \tag{3}$$

Based on CV_r, the probability of type I error is exactly equal to α.

Table 2 compares the statistical power of conservative and randomized group sequential tests for different number of groups. When the sequential analysis has many groups, the difference in power is only marginal, and the conservative version works well for actual safety surveillance. It is only for methods comparisons that the randomized version is absolutely needed, and in the rest of the paper it is the only one considered.

4 Continuous versus Group Sequential Analysis

For 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. We do this 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 $M = 1, ..., 10$. For group sequential analysis we used G equal to 2, 5, 10, 20, 50, 80, 100 and 200. All results are exact, based on numerical calculations using code written in the R software language (R Core Team, 2012). The R functions written are published as part of the open source R package 'Sequential'.

4.1 Fixed Maximum Sample Size

When the type 1 error 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 different values of the maximum sample size $T = 1, .., 50$, the statistical power and expected time to signal are shown in Table 3. 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.

	Statistical Power						Expected Time to Signal					
	Continuous Seq.		Group Sequential			Non-	Continuous Seq.		Group Sequential			Non-
T	M=1	M=3	G=10	G=5	G=2	Sequential	M=1	M=3	G=10	G=5	G=2	Sequential
1	0.185	0.234	0.215	0.223	0.234	0.234	0.35	0.58	0.48	0.57	0.81	1
1.5	0.221	0.277	0.266	0.259	0.269	0.297	0.54	0.75	0.76	0.80	0.96	1.5
2	0.255	0.315	0.301	0.322	0.347	0.360	0.75	0.94	0.98	1.16	1.58	2
2.5	0.289	0.351	0.344	0.367	0.360	0.405	0.96	1.14	1.23	1.44	1.65	2.5
3	0.323	0.384	0.379	0.369	0.431	0.446	1.19	1.34	1.44	1.50	2.25	3
4	0.390	0.445	0.483	0.487	0.480	0.542	1.67	1.73	2.15	2.30	2.72	4
5	0.447	0.507	0.513	0.547	0.580	0.605	2.09	2.17	2.43	2.80	3.63	5
6	0.500	0.561	0.555	0.615	0.627	0.672	2.51	2.57	2.79	3.45	4.11	6
8	0.600	0.656	0.708	0.698	0.749	0.769	3.35	3.36	3.99	4.20	5.68	8
10	0.685	0.733	0.767	0.776	0.813	0.845	4.13	4.07	4.69	4.98	6.66	10
12	0.756	0.794	0.834	0.855	0.882	0.892	4.85	4.71	5.38	6.12	8.21	12
15	0.836	0.866	0.889	0.913	0.928	0.942	5.77	5.57	6.29	7.15	9.60	15
20	0.921	0.936	0.956	0.956	0.974	0.979	6.96	6.62	7.57	8.17	12.00	20
25	0.963	0.972	0.980	0.987	0.990	0.993	7.75	7.35	8.19	9.76	14.06	25
30	0.984	0.988	0.994	0.994	0.996	0.998	8.26	7.78	9.10	10.16	16.15	30
40	0.997	0.998	0.999	0.999	1.000	1.000	8.74	8.24	9.94	11.97	20.61	40
50	1.000	1.000	1.000	1.000	1.000	1.000	8.94	8.45	10.36	13.46	25.32	50

Table 3: Statistical power for *RR=2* and expected time to signal for continuous and group sequen- tial analysis when the maximum sample size *T* is fixed, that is, the upper limit on the length of surveillance expressed in terms of the expected number of events under the null. M is the mini-mum number of adverse events required to signal and G is the number of group sequential tests. The type 1 error is $\alpha = 0.05$.

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 4 shows the expected time to signal for continuous sequential analysis with a minimum of $M = 1$ to 10 adverse events required to reject the null hypothesis, and for group sequential analysis with $G = 1, 2, 3, 5, 10, 20, 50, 100,$ 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.

Figure 1 depicts some of the values in Table 4 in order to easily compare the different designs. The smallest expected time to signal is indicated with a horizontal line. The method that signal the earliest is always found among the continuous sequential designs, with 4 to 6 adverse events as the minimum requirement to signal.

Power	Continuous Sequential Analysis							Group Sequential Analysis						
(RR=2)	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	3.63	2.99	2.46	2.30	2.21	2.29	2.24
0.60	3.23	3.06	2.89	2.81	**2.79**	2.85	2.99	4.92	3.98	3.45	2.97	3.04	3.02	3.02
0.70	4.26	3.95	3.76	3.65	**3.60**	3.62	3.69	6.57	4.99	4.33	3.96	3.97	3.87	3.89
0.80	5.34	4.99	4.77	4.65	4.57	**4.56**	4.60	8.78	6.50	5.40	4.92	4.82	4.88	4.93
0.85	5.92	5.59	5.37	5.22	5.15	**5.12**	5.14	10.29	7.42	6.02	5.47	5.37	5.60	5.54
0.90	6.53	6.27	6.04	5.90	5.81	**5.77**	5.78	12.31	8.67	6.93	6.14	6.06	6.17	6.16
0.95	7.48	7.11	6.88	6.73	6.64	**6.60**	**6.60**	15.70	10.51	8.16	7.02	7.02	6.90	6.97
0.98	8.15	7.78	7.56	7.41	7.32	**7.28**	**7.27**	20.26	12.63	9.20	7.86	7.52	7.52	7.82
0.99	8.40	8.09	7.87	7.73	7.64	**7.59**	**7.59**	23.42	14.19	9.78	8.25	7.97	7.86	7.92

Table 4: Expected time to signal for $RR = 2$ and different continuous and group sequential designs, when type 1 error and statistical power is held constant. M is the minimum number of adverse events needed to reject the null hypothesis and G is the number of group sequential tests.

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, Figure 2 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.

4.3 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 1 error and the statistical power fixed.

Table 5 presents the maximum sample size T as a function of different levels of statistical power when $RR = 2$ and for different sequential study designs. Figure 3 shows the same data using histograms. 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 $RR = 1.5, 2, 3, 4, 5,$ and 10, Figure 4 shows the maximum sample size T as a function of the statistical power. For continuous sequential analyses, we considered M equal to 1 and 4, and for group sequential, we used G equal to 2, 5, 20 and 50. As expected, the group sequential

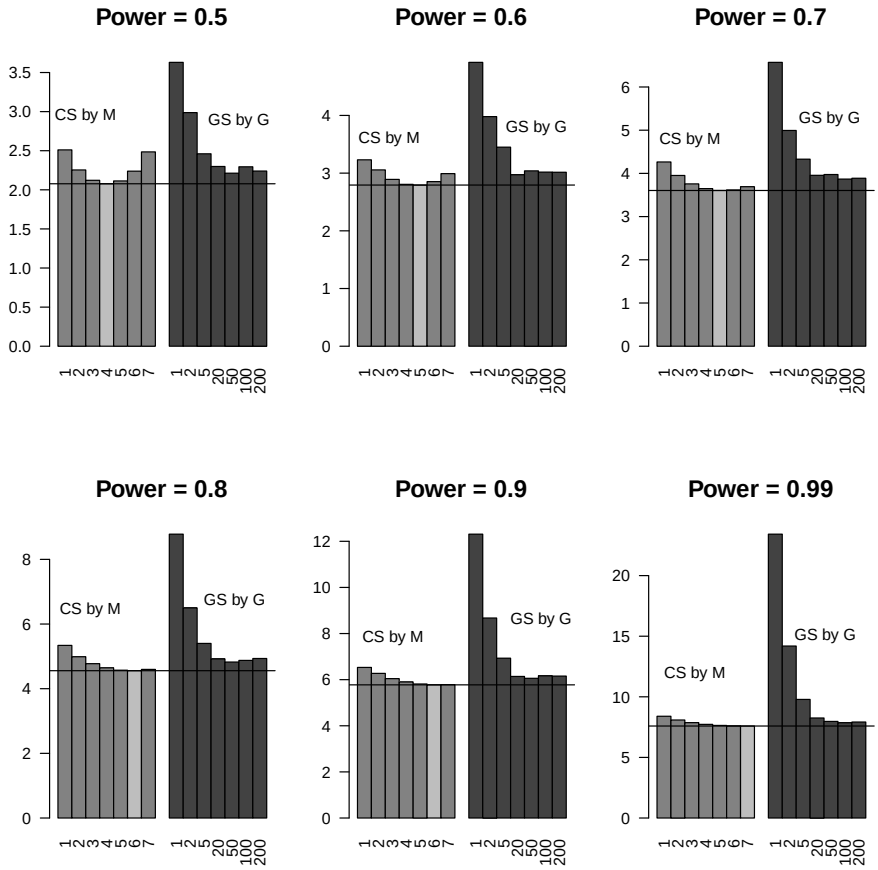

Figure 1: Expected time to signal when the statistical power is held fixed for continuous (CS) and group (GS) sequential analysis with $RR = 2$. The continuous sequential designs are represented by gray bars, with each bar representing a different value of M from 1 to 7. The black bars represent the groups sequential designs, with each bar associated with a different value of *G* from 1 to 200.

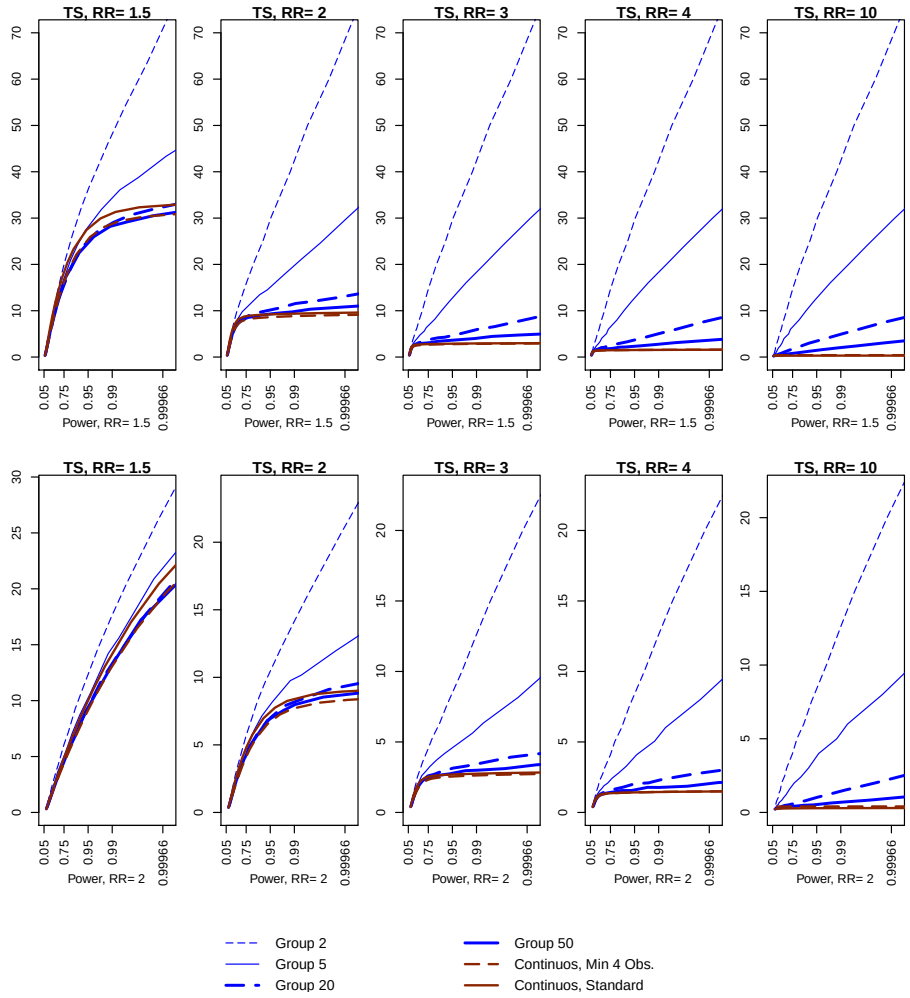

Figure 2: Comparison between continuous sequential analysis (CS) and group sequential analysis (GS) in terms of the expected time to signal. In the top row, the statistical power is held fixed for a RR of 1.5, while the expected time to signal is calculated when the true RR is 1.5, 2, 3, 4, and 10 respectively. In the bottom row, the statistical power is held fixed for a RR of 2, while the expected time to signal is calculated when the true RR is 1.5, 2, 3, 4, and 10 respectively. G is the number of sequential tests for the group designs and M is the minimum number of adverse events needed to signal in the continuous designs.

designs with fewer looks at the data have the smallest maximum sample size, but for $M = 4$, the difference is small and especially for larger RRs. The continuous sequential design with $M = 1$ has the largest maximum sample size across irrespectively of the RR.

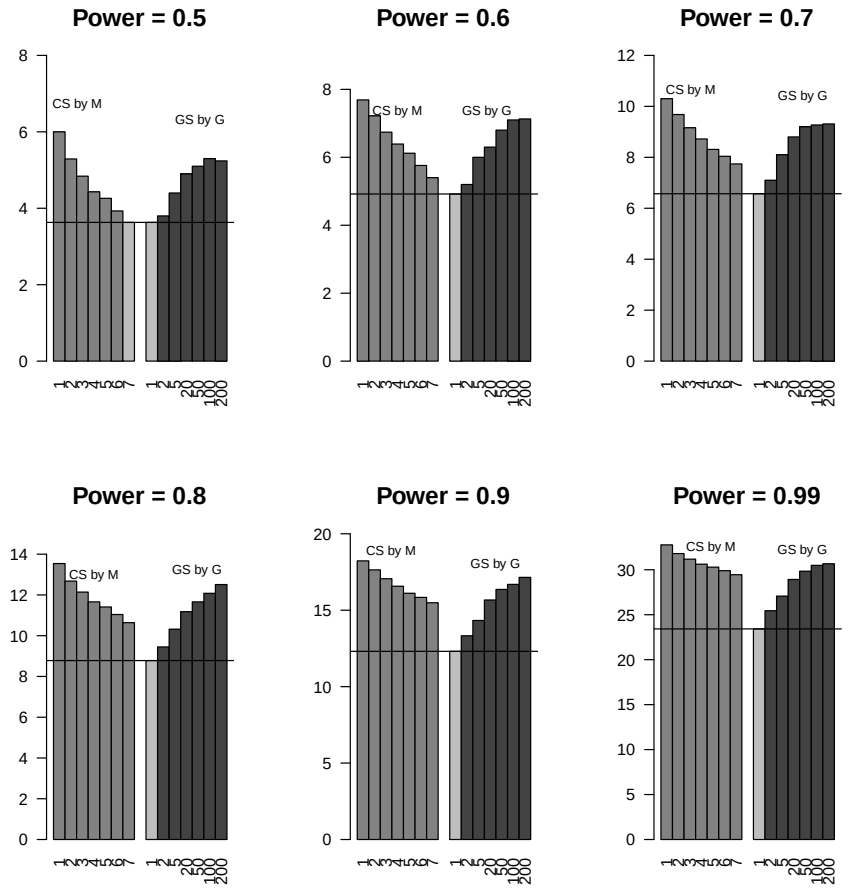

Figure 3: The maximum sample size T for continuous and group sequential analysis when the statistical power is held fixed for $RR = 2$. The continuous sequential designs (CS) are shown using gray bars with each bar representing a different value of M from 1 to 7. The black bars represent the groups sequential designs (GS) with each bar associated with a different value of G from 1 to 200.

	Continuous Sequential Analysis							Group Sequential Analysis						
Power	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**	**3.63**	3.80	4.40	4.90	5.10	5.30	5.24
0.6	7.69	7.22	6.74	6.39	6.12	5.76	5.40	**4.92**	5.20	6.00	6.30	6.80	7.10	7.13
0.7	10.30	9.68	9.16	8.72	8.31	8.04	7.74	**6.57**	7.10	8.10	8.80	9.20	9.27	9.31
0.8	13.54	12.68	12.14	11.66	11.41	11.04	10.64	**8.78**	9.45	10.32	11.18	11.66	12.08	12.51
0.85	15.56	14.79	14.24	13.74	13.39	13.09	12.69	**10.29**	11.02	11.92	12.81	13.57	14.59	14.58
0.9	18.23	17.64	17.06	16.57	16.11	15.84	15.49	**12.31**	13.33	14.33	15.67	16.36	16.69	17.15
0.95	22.92	22.18	21.58	21.07	20.60	20.29	19.95	**15.70**	16.72	17.67	19.64	21.05	21.09	21.50
0.98	28.75	27.76	27.15	26.61	26.10	25.88	25.46	**20.26**	21.55	23.57	25.08	25.40	26.09	27.28
0.99	32.77	31.80	31.18	30.62	30.29	29.90	29.45	**23.42**	25.45	27.07	28.92	29.84	30.49	30.67

Table 5: Maximum sample size (length of surveillance) for fixed statistical power when $RR = 2$ and for different and for different continuous and group sequential designs. M is the minimum number of adverse events needed to reject the null hypothesis and G is the number of group sequential tests.

4.4 Expected Time to Signal versus Maximum Sample Size

By keeping the type 1 error 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 5 (Appendix), 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 reach when requiring around 6 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. This means that whatever preference we have with respect to maximum sample size and expected time to signal, the optimal design is found among the continuous sequential designs.

5 Pediarix Vaccine and Neurological Symptoms

To evaluate and illustrate the different sequential designs on a specific exposure-outcome pair, we used historical data on neurological adverse events after Pediarix vaccine. Manufactured by GlaxoSmithKline, Pediarix is a combination vaccine that, with one injection, protects children from diphtheria, tetanus, whooping cough, hepatitis B, and Polio. The adverse event considered was any neurological symptoms during the 1 to 28 days after vaccination. The data are based on electronic health records from Kaiser Permanente Northern California and has previously been analyzed using the continuous MaxSPRT (Kulldorff et al., 2011).

We performed the evaluation mimicking weekly post-market safety surveillance, using different sequential parameter settings. For the group sequential analysis, we used $G = 1, 2, 3, 5, 10$, and 20 equally spaced group sizes in terms of the number of vaccinated children, rounded to a whole week's worth of data. For the continuous sequential design, we used $M = 4$ as the minimum required number of adverse events needed to reject the null (Silva and Kulldorff, 2012). The maximum lengths of surveillance were set in order to have 90 percent statistical power for a relative risk of 2. Results are shown in Table 6. The continuous sequential analysis and the group sequential analysis with 20 looks rejects the null hypothesis the earliest, while the other group sequential tests reject the null hypothesis 4 to 18 weeks later. Had the null not been rejected,

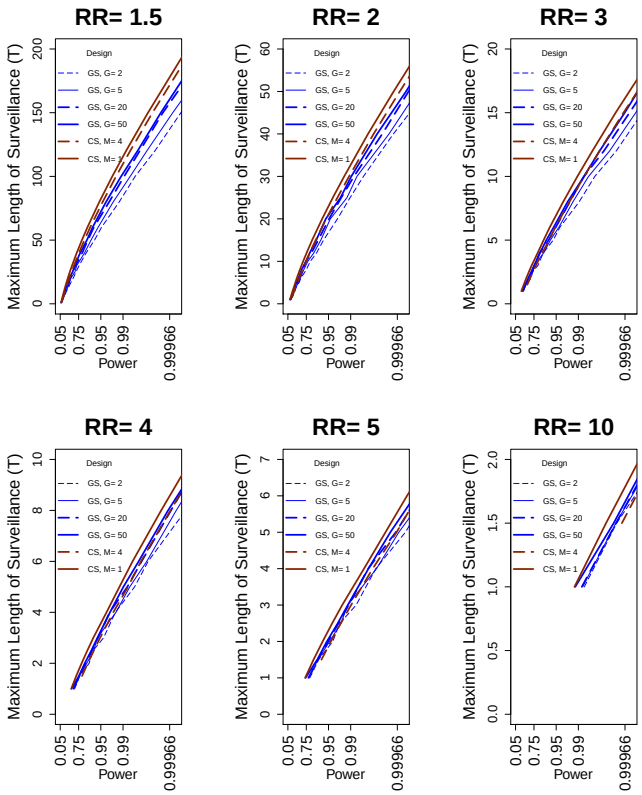

Figure 4: Comparison between continuous sequential analysis (CS) and group sequential analysis (GS) in terms of the maximum sample size. The statistical power is held fixed for a RR of 2, while the maximum sample size is calculated when the true RR is 1.5, 2, 3, 4, 5 and 10 respectively. G is the number of sequential tests for the group designs and M is the minimum number of adverse events needed to signal in the continuous designs.

then the surveillance would have ended earlier for the group sequential tests with fewer looks at the data. With approximately 7 to 8 weeks between each expected adverse event under the null, the continuous surveillance would have lasted approximately 25 weeks longer than the group sequential analysis with only two looks at the data.

The explanation for why more frequent testing tend to reject the null hypothesis earlier can be seen in Figure 6 (Appendix). Although the critical values for less frequent tests are lower than those for the more frequent testing, they have the inconvenience of being blind to what happens during the long time periods between the testing times. Adverse events will not arrive at regular intervals, but approximately as an irregular random Poisson process with naturally and randomly occurring temporal clusters of adverse events. When there is a sudden burst of adverse events, the log likelihood ratio may quickly increase. Sequential designs with more frequent tests will quickly detect that, while the less frequent designs will have to wait until the next scheduled look at the data, which could take some time. Note though that a group sequential test with fewer looks will sometimes reject the null hypothesis before a group sequential test with more looks

at the data. For example, for the group sequential design with 5 looks, the observed likelihood ratio was not large enough to reject at the first test and it took a long time until the second test came around. On the other hand, for the group sequential analysis with 3 tests, the first test came soon after the raise in the log likelihood ratio, and the null was quickly rejected. This is an artifact of particular data sets and analysis parameters rather than a general phenomena though. For example, if we had powered the analysis at 80 rather than 90 percent, the group sequential analysis with five tests would have rejected the null hypothesis before the one with three tests (data not shown).

6 Discussion

In this paper we have compared the performance of continuous and groups sequential tests for post-market drug, vaccine and device safety surveillance. Most of the literature on sequential analysis is based on asymptotic approximations or simulation studies. A strength of this paper is that all results are based on exact numerical calculations.

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 theorem 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 theorem, 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.

Data is not always available in a continuous or near continuous fashion, and it is then appropriate to use group sequential analysis. If no new data has arrived there is no need to conduct additional sequential tests, but whenever more data arrives a new test should be conducted. If the availability of continuous or near continuous data is more expensive, it is important to know how

	Continuous	Group Sequential Design, #tests				
	Design	20	10	5	3	2
Maximum Sample Size (T)	16.57	15.67	15.21	14.33	14.27	13.33
Critical Value	3.111641	2.631052	2.471350	2.116951	1.831645	1.721441
Week when Null Rejected	32	32	36	44	37	50
Adverse Events when Null Reject	10	10	10	16	11	18
Expected Events when Null Rejected	4.03	4.03	4.63	5.81	4.80	6.69
LLR when Null Rejected	3.120331	3.120331	2.325838	6.012579	2.924130	6.501149

Table 6: Sequential analysis results for neurological symptoms after Pediarix using either a continuous sequential design with a minimum of 4 adverse events required to reject the null hypothesis or a group sequential design with a different number of tests. All designs have a type 1 error of 0.05 and a 90 percent power to detect a relative risk of 2.

much performance is lost with the less frequent data feeds, in order to weigh the trade-off between performance and cost. The numerical calculations that we have done can help determine this. While these are based on a Poisson probability model and a Wald type rejection boundary that is flat with respect to the log likelihood ratio, we expect that the trade-off between continuous and group sequential designs will be similar for other probability models and other rejections boundaries.

While the continuous sequential analysis with a flat log likelihood boundary performs well, we have not shown that it is optimal in any sense of the word. In fact, it is likely that there are other continuous sequential designs that are better, depending on the particular exposure-outcome pair under surveillance. What this paper has shown is that when we search for the best sequential design for post-market safety surveillance, we can restrict that search to continuous sequential designs. If data cannot be made available in a continuous or near continuous fashion, we can restrict the search to group sequential designs that perform a sequential test as soon as a new batch of data arrives.

Post-market drug, vaccine, and device safety surveillance is important to guarantee the safety of medical products, and while serious problems are rare, it is important to detect those problems as soon as possible. For commonly used products, a few weeks earlier detection could considerably reduce mortality or morbidity when a serious problem exists. We hope that this paper will help regulatory agencies to conduct post-market safety surveillance in the most efficient manner possible.

7 Acknowledgements

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. Dr. Ivair Silva received additional support from Conselho Nacional de Desenvolvimento Científico e Tecnológico(CNPq) and from Banco de Desenvolvimento do Estado de Minas Gerais (BDMG), Brazil.

References

Abt, K. (1998), ``Poisson Sequential Sampling Modified Towards Maximal Safety in Adverse Event Monitoring,'' *Biometrical Journal*, 40, 21--41.

Avery, T., Vilk, Y., Kulldorff, M., Li, L., Cheetham, T., and Dublin., S. (2010), ``Prospective, active surveillance for medication safety in population-based health networks: a pilot study.'' *Pharmacoepidemiol Drug Saf*, 19.

Belongia, E., Irving, S., Shui, I., Kulldorff, M., Lewis, E., Yin, R., Lieu, T., Weintraub, E., Yih, W., Li, R., Baggs, J., and the Vaccine Safety Datalink Investigation Group (2010), ``Real-Time Surveillance to Assess Risk of Intussusception and Other Adverse Events after Pentavalente, Bovine-Derived Rotavirus Vaccine,'' *Pediatric Infectious Disease Journal*, 29, 1--5.

Brown, J., Kulldorff, M., Chan, K., Davis, R., Grahan, D., Pettus, P., Andrade, S., Raebel, M., Herrinton, L., Roblin, D., Boudreau, D., Smith, D., Gurwitz, J., Gunter, M., and Platt., R. (2007), ``Early

Detection of Adverse Drug Events within Population-Based Health Networks: Application of Sequential Testing Methods.'' *Pharmacoepidemiology and Drug Safety*, 16, 1275--1284.

Davis, R., Kolczak, M., Lewis, E., Nordin, J., Goodman, M., Shay, D., Platt, R., Black, S., Shinefield, H., and Chen, R. (2005), ``Active Surveillance of Vaccine Safety: A System to Detect Early Signs of Adverse Events,'' *Epidemiology*, 16, 336--341.

Jennison, V. and Turnbull, B. (1999), *Group Sequential Methods with Applications to Clinical Trials*, no. ISBN 0-8493-0316-8, London: Chapman and Hall/CRC.

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., and the Vaccine Safety Datalink (2010), ``Measles-Mumps-Rubella-Varicella Combination Vaccine and Risk of Febrile Seizures,'' *Pediatrics*, 126, e1--e8.

Kulldorff, M. (2012), ``Sequential Statistical Methods for Prospective Postmarketing Safety Surveillance,'' *Pharmacoepidemiology, John Wiley & Sons, Ltda.*, Fifth Edition, 852--867.

Kulldorff, M., Davis, R., M, K., Lewis, E., Lieu, T., and Platt, R. (2011), ``A maximized sequential probability ratio test for drug and vaccine safety surveillance,'' *Sequential Analysis*, 30, 58--78.

Lai, T. (1991), ``Asymptotic optimality of generalized sequential likelihood ratio tests in some classical sequential testing problems,'' *Handbook of Sequential Analysis*, 21, 121--144.

Lieu, T., Kulldorff, M., Davis, R., Lewis, E., Weintraub, E., Yih, W., Yin, R., Brown, J., and Platt, R. (2007), ``Real-Time Vaccine Safety Surveillance for the Early Detection of Adverse Events,'' *Medical Care*, 45, 89--95.

Nelson, J., Cook, A., Yu, O., Dominguez, C., Zhao, S., Greene, S., Fireman, B., Jacobsen, S., Weintraub, E., and Jackson, L. (2012), ``Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using eletronic observational health care data,'' *Pharmacoepidemiology and Drug Safety*, 21, 62--71.

O'Brien, P. and Fleming, T. (1979), ``A multiple testing procedure for clinical trials,'' *Biometrics*, 35, 549--556.

Platt, R., Carnahan, R., Brown, J., Chrischilles, E., Curtis, L., Hennessy, S., Nelson, J., Racoosin, J., Robb, M., Schneeweiss, S., Toh, S., and Weiner., M. (2012), ``The U.S. Food and Drug Administration□s Mini-Sentinel program: status and direction.'' *Pharmacoepidemiology and Drug Safety*, 21, 1--8.

Pocock, S. (1977), ``Group sequential methods in the design and analysis of clinical trials,'' *Biometrika*, 64, 191--199.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Shih, M., Lai, T., Heyse, J., and Chen, J. (2010), ``Sequential Generalized Likelihood Ratio Tests for Vaccine Safety Evaluation,'' *Statistics in Medicine*, 26, 2698--2708.

Siegmund, D. and Gregory, P. (1980), ``A Sequential Clinical Trial for Testing $p1 = p2$,'' *Annals of Statistics*, 8, 1219--1228.

Silva, I. and Kulldorff, M. (2012), ``Continuous Sequential Analysis with a Delayed Start,'' *To be submitted*.

Wald, A. (1945), ``Sequential Tests of Statistical Hypotheses,'' *Annals of Mathematical Statistics*, 16, 117--186.

--- (1947), *Sequential Analysis*, no. ISBN 0-471-91806-7, New York: John Wiley and Sons.

Weiss, L. (1953), ``Testing One Simple Hypothesis Against Another,'' *Annals of Mathematical Statistics*, 24, 273--281.

Whitehead, J. and Stratton, I. (1983), ``Group Sequential Clinical Trials with Triangular Continuation Regions.'' *Biometrics*, 39, 227--236.

Yih, W., Kulldorff, M., Fireman, B., Shui, I., Lewis, E., Klein, N., Baggs, J., Weintraub, E., Belongia, E., Naleway, A., Gee, J., Platt, R., and Lieu, T. (2011), ``Active Surveillance for Adverse Events: The Experience of the Vaccine Safety Datalink Project,'' *Pediatrics*, in press.

Yih, W., Nordin, J., Kulldorff, M., Lewisc, E., Lieua, T., Shia, P., and Weintraube, E. (2009), ``An assessment of the safety Datalink of adolescent and adult tetanus-diphteria-acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the Vaccine Safety Datalink,'' *Vaccine*, 27, 4257--4262.

Zhao, S., Cook, A., Jackson, L., and Nelson, J. (2012), ``Statistical performance of group sequential methods for observational post-licensure medical product safety surveillance: a simulation study,'' *Statistics and Its Interface In Press*.

8 Appendix

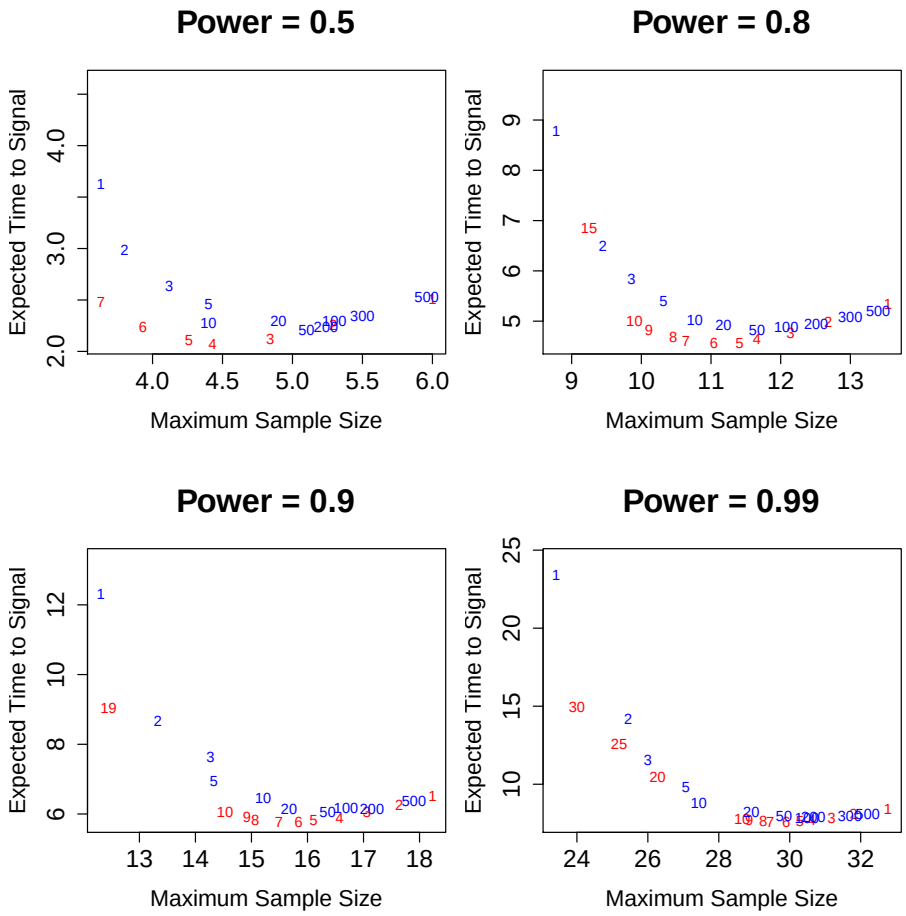

Figure 5: For fixed type 1 error of 0.05 and fixed statistical power when RR=2, the expected time to signal is plotted against the maximum sample size. The red labels refer to continuous sequential designs, with the numbers indicating the minimum number of adverse events required to reject the null hypothesis (M). The blue labels refer to group sequential designs with the numbers indicating the number of sequential tests conducted (G). The blue label with number 1 is a standard non-sequential analysis.

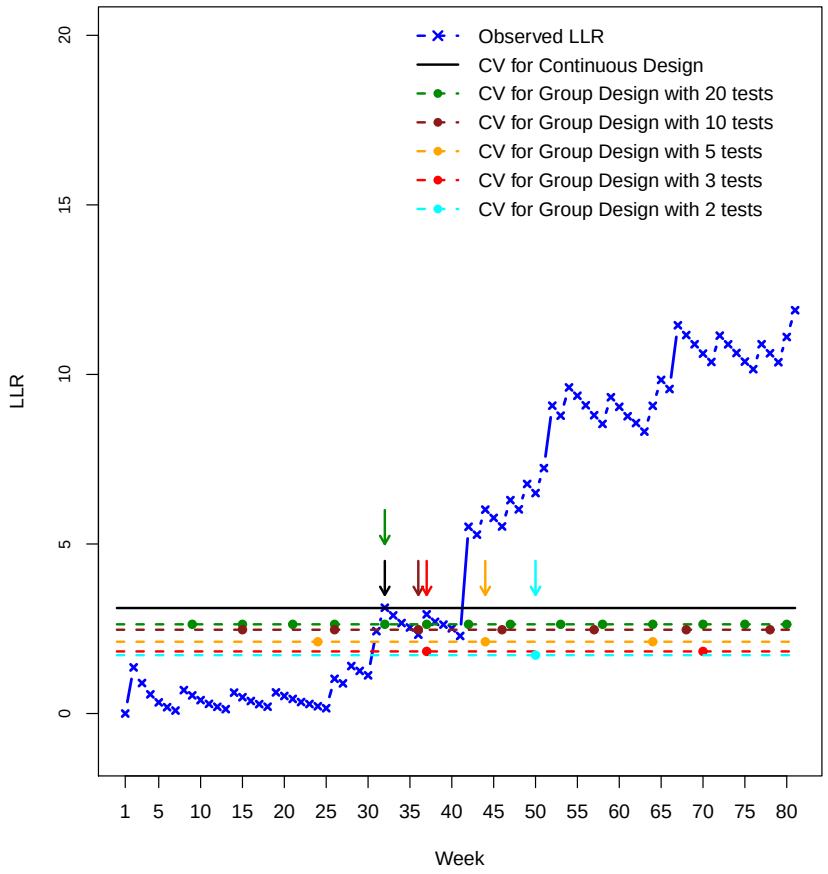

Figure 6: Continuous and group sequential analyses of neurological symptoms after Pediarix vaccine. The type 1 error is 0.05 and the statistical power is 90 percent for a relative risk of 2. For the group sequential designs, the testing moments are highlighted by solid dots, and the null cannot be rejected between those dots. For the continuous sequential design, we required at least 4 adverse events before rejecting the null. *LLR* = log likelihood ratio, *CV* = critical value.