Proper inference from Simon’s two-stage designs

Tatsuki Koyama*† and Heidi Chen

Department of Biostatistics, Vanderbilt University School of Medicine, 571 Preston Building, Nashville, TN 37232-6848, U.S.A.

SUMMARY
Simon’s two-stage designs are very popular for phase II clinical trials. A literature review revealed that the inference procedures used with Simon’s designs almost always ignore the actual sampling plan used. Reported $P$-values, point estimates and confidence intervals for the response rate are not usually adjusted for the design’s adaptiveness. In addition, we found that the actual sample size for the second stage is often different from that planned. We present here a method for inferences using both the planned and the actual sample sizes. The conventional and the preferred inference procedures usually yield similar $P$-values and confidence intervals for the response rate. The conventional inference, however, may contradict the result of the corresponding hypothesis testing. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: clinical trials; adaptive design; conditional power; sample size modification; $P$-values

1. INTRODUCTION

Two-stage designs are commonly used in phase II clinical trials and especially in cancer clinical trials. Simon [1] has proposed two criteria, minimax and optimal, for selecting sample sizes and critical values for these two-stage designs. The maximum sample size and the expected sample size under $H_0$ are minimized in the minimax and optimal designs, respectively. Simon’s two-stage designs have been gaining influence; his 1989 paper [1] has been cited more than 700 times. Since it was cited twice in 1991, the number of citations per year has increased steadily to 50 times in 2001 and over 100 times in 2005 and 2006.

There have been a number of extensions proposed for Simon’s designs. These include consideration of toxicity [2–5], inclusion of more than one treatment [6–9], addition of the third stage [10], consideration of partial and complete responses [11, 12] and consideration of multiple strata [13]. Banerjee and Tsiatis [14] proposed to extend a Simon’s design by allowing different stage 2 sample sizes for different stage 1 results. Their more flexible designs have advantages over the...
original Simon’s designs with respect to the expected sample sizes under the null and alternative hypotheses. Others have considered optimality in two-stage designs and proposed to improve on Simon’s designs [15–17].

Owing to the adaptive nature of the design, the inference procedures for Simon’s designs are not straightforward. A maximum likelihood estimator of the response rate, number of positive responses/total number of patients, is biased [18, 19]. Confidence interval and P-value should not be computed as if the data were obtained in a single stage. There are relatively few papers that discuss the inference procedures used with two-stage designs in phase II clinical trials. Whitehead [18] has studied the bias of the maximum likelihood estimator and has proposed a bias-reduced estimator that was further studied by Chang et al. [19]. Jung and Kim [20] have provided a comprehensive study of estimators from a multi-stage design. For general group sequential designs, the inference procedures have been considered by Fairbanks and Madsen [21], Tsiatis et al. [22], Lin et al. [23], Yi and Yu [24], Fan and DeMets [25], among others.

It could be argued that these two-stage designs are primarily for decision making and that estimation is a secondary objective of a phase II clinical trial. It is preferable, however, to compute a P-value, a confidence interval and an estimate of the response rate at the termination of the trial. The latter two are especially useful when the design of a new phase III clinical trial is based on the findings of a phase II trial. As a matter of fact, our literature review revealed that the majority of studies using Simon’s two-stage designs report an estimate of the response rate with a confidence interval.

Our literature review also revealed that the actual sample size of stage 2 is often different from the planned stage 2 sample size. It is not straightforward to conduct a hypothesis testing when the stage 2 sample size is changed in a Simon’s design. We will further extend the inference procedures to handle those cases in which the actual sample size is different from the one planned.

A subtle but critical issue is that the decision to use a different sample size must be made blinded to any part of the data including the stage 1 result. A legitimate situation in which a sample size is different from that planned can occur when the sample size is planned in anticipation of a certain number of non-informative drop-outs, but a different number of patients actually finish the study. On the other hand, it is not permissible to decide at the end of stage 1 (or during stage 2) to extend the study because there are fewer positive responses than expected or to shorten the study simply because there are more positive responses than expected. There may be designs that allow such adaptive changes and still protect type I error rates. In order for such designs to make proper inferences additional assumptions are required; that of the sample size which would have been used if a different number of successes had been observed in stage 1 [26, 27]. We will only consider non-informative sample size change in this paper.

Because we only consider non-informative drop-outs, we can make the actual stage 1 sample size always equal to the same as the planned stage 1 sample size. Missing data are treated as missing completely at random and non-informative, and they are not included in the analysis. Stage 1 terminates when we obtain data on exactly the planned number of patients. On the other hand, we often do not have the same level of flexibility at the end of stage 2. If we have obtained more observations than planned, they are all included in the final analysis, and if we have fewer than the planned number of samples the trial is sometimes terminated due to outside circumstances such as budgetary restrictions, even if an extension would otherwise be possible. In this paper, we assume that the actual stage 1 sample size is the same as that planned. Green and Dahlberg [28] and Chen and Ng [15] have considered the cases in which the actual stage 1 sample size is different from the planned sample size.
In Section 2, a brief introduction of Simon’s designs is given. We apply the inference procedures developed for multi-stage group sequential designs to Simon’s designs in Section 3. Then in Section 4, we present general inference procedures for Simon’s designs when the stage 2 sample size is different from the one planned.

2. SIMON’S DESIGNS

In a study with a Simon’s design, the null hypothesis is concerned with a response rate, \( \pi \). Without loss of generality we assume that a higher value of \( \pi \) is more favorable and write \( H_0: \pi \leq \pi_0 \). The power of the study is set at some \( \pi_1 \) that is greater than \( \pi_0 \).

A Simon’s design is usually indexed by four numbers that represent the stage 1 sample size \( (n_1) \), stage 1 critical value \( (r_1) \), final sample size \( (n_f) \) and final critical value \( (r_f) \). In stage 1, a sample of size \( n_1 \) is taken. Let \( X_1 \) be the number of successes in stage 1. If \( X_1 \leq r_1 \), the trial is stopped for futility; otherwise, an additional sample is taken until a total of \( n_f \) observations are obtained. Let \( X_2 \) be the number of successes in stage 2, and let \( X = X_1 + X_2 \). If \( X_1 \leq r_1 \), futility is concluded; otherwise efficacy is concluded by rejecting \( H_0 \).

We depart from convention and use \( R_1 = r_1 + 1 \) and \( R_f = r_f + 1 \) to denote the critical values of a Simon’s design. This new notation is a simpler one in certain extensions that we will consider in later sections. It will also be necessary to consider the stage 2 critical value, and we will use the notation \( R_2(x_1) \), which is a function of \( X_1 \) and can be expressed simply as \( R_2(x_1) = R_1 - x_1 \) for \( R_1 \leq X_1 < R_f \), and 0 for \( R_f \leq X_1 \) in a usual Simon’s design.

To present the power function of a Simon’s design, we first introduce the conditional power of stage 2 given the stage 1 result, \( X_1 = x_1 \). It is \( P_\pi [X_2 \geq R_2(x_1)] \) for \( X_1 \geq R_1 \), where \( X_2 \sim \text{Binomial}(n_2, \pi) \). Throughout the paper, the notation \( P_\pi [E] \) represents the probability of the event \( E \) at a specific \( \pi \). We use the notation \( A(x_1, n_2, \pi) \) to denote the conditional power, which can be expressed as

\[
A(x_1, n_2, \pi) = \sum_{x_2=R_2(x_1)}^{n_2} \binom{n_2}{x_2} \pi^{x_2} (1-\pi)^{(n_2-x_2)}
\]

(1)

Inclusion of \( n_2 \) as one of the arguments of the conditional power is necessary because we will consider changing \( n_2 \) from the one planned in the later sections.

If the trial is terminated in stage 1 because the results indicated futility \( (X_1 < R_1) \), we let \( A(x_1, n_2, \pi) = 0 \). We note that the conditional power at \( \pi = \pi_0 \) is the conditional type I error rate.

The power function of a Simon’s design is

\[
\beta(\pi) = P_\pi [\text{Reject } H_0] = \sum_{x_1=R_1}^{n_1} P_\pi [X_1 = x_1] A(x_1, n_2, \pi)
\]

(2)

A design \( (n_1, R_1, n_2, R_f) \) is usually chosen so that \( \beta(\pi_0) \leq \alpha \) and \( \beta(\pi_1) \geq 1 - \beta \).

3. INFERENCEx PROCEDURES

Inference procedures for multi-stage designs have been discussed by many [21–25, 29], and we will review them in this section and extend them in the subsequent section.
3.1. P-value

Suppose that the number of successes is $X_1 = x_1 \geq R_1$ and $X_2 = x_2$ for stages 1 and 2, respectively. Note that $x_t = x_1 + x_2$. It is a common practice to compute a P-value at the end of stage 2, incorrectly assuming that the data are collected in a single stage. Let us call this a conventional P-value and denote it by $p_c$. We can express $p_c = \sum_{x_1=0}^{n_1} P_{\pi_0}[X_1 = x_1] P_{\pi_0}[X_2 \geq x_t - x_1]$ (3)

The summand of the right-hand side of the above equation includes impossible sample paths in which $X_1 < R_1$ and $X_2 = x_t - x_1$. A preferred P-value may be $p_p = \sum_{x_1=R_1}^{n_1} P_{\pi_0}[X_1 = x_1] P_{\pi_0}[X_2 \geq x_t - x_1]$ (4) which does not include these impossible sample paths. Clearly, $p_c \geq p_p$.

The P-value is the probability of obtaining the result that is at least as extreme as the observed one under the null hypothesis. In (3) and (4), we are implicitly assuming that the larger value of $X_t$, regardless of $X_1$ and $X_2$, is more extreme. When $X_1 < R_1$ and futility is concluded in stage 1, the P-value is $P_{\pi_0}[X_1 \geq x_1]$. Because we use the total number of successes to order the possible outcomes in (4), it is applicable only if $n_2$ is a constant for all $X_1$. In certain extensions of Simon’s design in which the stage 2 sample size is not a constant in $X_1$ (e.g. Banerjee and Tsiatis [14]), the P-value in (4) is not applicable. We will also show in Section 4 that, when the realized stage 2 sample size is different from that planned in a simple design with a constant $n_2$, (4) may not be applicable.

3.2. Confidence interval

A 95 per cent two-sided confidence interval is reported in many of the papers that we have studied. It is not wrong to report a 95 per cent two-sided confidence interval. If, however, it is desirable that a confidence interval and hypothesis testing be consistent, because Simon’s design is used in one-sided hypothesis testing, then a 95 per cent one-sided confidence interval of the form $(\pi_L, 1]$ for $\alpha = 0.05$ is required. The method described below can be used to form a more common two-sided confidence interval of the form, $(\pi_L, \pi_U)$. We will use a 90 per cent two-sided confidence interval to allow consistency with the one-sided hypothesis testing at $\alpha = 0.05$. In other words, we can interpret this confidence interval in a usual way: the two statements, ‘$\pi_0$ is not contained in a confidence interval’ and ‘$H_0$ is rejected’, are equivalent.

We can compute a P-value for testing $H_0: \pi \leq \pi_0'$ using (4) for any $\pi_0'$. A 90 per cent two-sided confidence interval that we use is a collection of $\pi_0'$ such that the corresponding P-value is within $[0.05, 0.95]$. This is a simple application of the method based on ‘stage-wise ordering’ of group sequential methodology and is a methodology that produces an interval [29, 30].

3.3. Point estimate of the response rate

The maximum likelihood estimator of the response rate, $\hat{\pi} = x_t / n_t$ if $x_1 \geq R_1$ or $\hat{\pi} = x_1 / n_1$ if $x_1 < R_1$, underestimates the true response rate in Simon’s designs that we are considering. When $x_1 / n_1$ is larger than $\pi$, the trial tends to proceed to stage 2, and the upward bias tends to be corrected.
On the contrary, when \( x_1/n_1 \) is smaller than \( \pi \), the trial tends to be terminated without a chance for the downward bias to be corrected.

A bias-reduced estimator due to Whitehead [18], denoted here by \( \hat{\pi}_w \), is the solution to \( E_{\hat{\pi}}[\hat{\pi}] = \hat{\pi} \) [19]. The characteristics of this and other estimators have been studied previously [20, 31]. Generally, it is more favorable than the maximum likelihood estimator in terms of the mean square errors.

### 3.4. Example

We will consider as an example a design \( (n_1=10, R_1=2, n_2=29, R_2=6) \) that is the minimax design for \( \pi_0=0.1, \pi_1=0.3 \) with \( \alpha=0.05, \beta=0.2 \). Suppose that we observe \( X_1=2 \) and \( X_2=6 \). This sample path leads to the rejection of \( H_0 \) in stage 2. We can compute for this example, \( p_c=0.064 \) and \( p_p=0.047 \) using (3) and (4). A 90 per cent confidence interval is \((0.094, 0.368)\) with (3) and \((0.102, 0.401)\) with (4). The conventional \( P \)-value (3) yields a \( P \)-value that is greater than \( \alpha \) and a confidence interval that contains \( \pi_0 \) although \( H_0 \) is rejected. And \( \hat{\pi}=0.207, \hat{\pi}_w=0.243 \).

### 4. EXTENDING OR SHORTENING THE STUDY

Our literature review reveals that the actual sample size is often different from the planned sample size in many studies that use a Simon’s design. Depending on the predicted and actual accrual rates and drop-out rates, the actual sample size may be larger or smaller.

Extending or shortening a study is simple in a single-stage design since it is easy to recalculate a new critical value or the correct \( P \)-value and make the correct decision regarding \( H_0 \). With a Simon’s two-stage design, however, extending or shortening a study is not as straightforward. A common practice to conduct a hypothesis testing is to compute the conventional \( P \)-value that is based on incorrect distributional assumption and use it to make a decision. Equivalently, the critical value is re-computed, as if the new sample size were originally planned in a single-stage design. Unlike the simple two-stage designs with no sample size change, where the preferred \( P \)-value in (4) is always smaller than or equal to the conventional one in (3), it is now possible to make an incorrect decision and inflate the type I error rate. We will discuss in this section how to calculate the new critical value and how to make a new inference based on the actual sample size.

As a notational convention, we use a prime to indicate that the sample size has changed. For example, \( n'_2 \) is the new sample size for stage 2 and \( x'_2 \) is the number of successes among \( n'_2 \) in stage 2.

#### 4.1. Hypothesis testing

We have introduced the term conditional power in Section 2 and have denoted it by \( A(x_1, n_2, \pi) \), which is the probability of rejecting \( H_0 \) in stage 2 given the result of stage 1. Rejecting \( H_0 \) if \( X_2 \geq R_2(x_1) \) and rejecting \( H_0 \) if \( \text{cp}(x_1, x_2, n_2, \pi_0) \equiv P_{\pi_0}[X_2 \geq x_2 | X_1 = x_1] \leq A(x_1, n_2, \pi_0) \) are equivalent. We call \( \text{cp}(x_1, x_2, n_2, \pi_0) \) the conditional \( P \)-value, which is the \( P \)-value of stage 2 given the result of stage 1. The conditional \( P \)-value is compared with the conditional type I error rate in decision making at the termination of stage 2.

The conditional \( P \)-value can be computed regardless of the sample size of the actual stage 2, and the decision to reject \( H_0 \) can be made by comparing the conditional \( P \)-value and conditional
type I error rate evaluated at the observed \( x_1 \). The same conclusion may be based on the new critical value, \( R'_2(x_1) \), which is defined as the largest \( R \), such that

\[
P_{\pi_0}[X'_2 \geq R | X_1 = x_1] \leq P_{\pi_0}[X_2 \geq R_2(x_1) | X_1 = x_1]
\]

(5)

where \( X'_2 \sim \text{Binomial}(n'_2, \pi) \). With the new critical value, \( R'_2(x_1) \), we define \( A'(x_1, n'_2, \pi) = P_{\pi}[X'_2 \geq R'_2(x_1) | X_1 = x_1] \). Because we set the new critical value so that the new conditional type I error rate is always equal to or smaller than the original conditional type I error rate, the overall unconditional type I error rate is controlled.

A simplistic but inappropriate method of hypothesis testing may be a re-computation of the critical value as if the new sample size, \( n'_2 \), was planned originally. This involves a simple numerical search of the critical value that produces a type I error rate of less than \( \alpha \) given \( n_1 \), \( R_1 \) and \( n'_2 \). This method only attempts to control the unconditional type I error rate even though the stage 1 has already completed. If we apply the same reasoning and only attempt to control the unconditional type I error rate, we could have as much conditional type I error rate at the observed \( X_1 = x_1 \) as \( \min[1, x/P_{\pi_0}[X_1 = x_1]] \) and 0 elsewhere. Then the result of computation of (2) at \( \pi = \pi_0 \) with \( A(x_1, n_2, \pi_0) \) replaced by \( A'(x_1, n'_2, \pi_0) \) would be at most \( x \). The type I error rate, however, is not actually controlled because we would have used different critical values if \( X_1 \) were different. In other words, this method fails to protect the conditional type I error rate, which is relevant and needs to be protected after stage 1. In this method, attempted control of the unconditional type I error rate does not guarantee control of the conditional type I error rate which is necessary for ensuring ultimate control of the unconditional type I error rate. The second method (assuming that \( n'_2 \) was planned originally) should be avoided because it does not guarantee that the unconditional type I error rate will not exceed \( x \).

When the new critical value, \( R'_2(x_1) \), is obtained from (5), the total number of positive responses, \( x_1 + R'_2(x_1) \), necessary to reject \( H_0 \) may be different for different values of \( X_1 \) as the following example demonstrates:

Consider the design \((n_1 = 19, R_1 = 7, n_2 = 39, R_2 = 17)\). This is the minimax design for \( \pi_0 = 0.3, \pi_1 = 0.5 \) with \( \alpha = 0.05 \) and \( \beta = 0.2 \). Suppose that we observe \( X_1 = 7 \) in stage 1. Then \( H_0 \) would be rejected if the stage 2 \( P \)-value is smaller than the conditional type I error rate, \( A(7, 20, 0.3) = P_{\pi_0}[X_2 \geq 10] = 0.0480 \). Further suppose that the stage 2 sample size is increased from the planned 20 to 23. If we observe 11 successes in these 23 stage 2 observations, the conditional \( P \)-value is \( P_{\pi_0}[X'_2 \geq 11] = 0.0546 \), and the null hypothesis should not be rejected. The critical value at \( X_1 = 7 \) with \( n_2 = 23 \) is \( R'_2(7) = 12 \) from (5).

In a different scenario, suppose that we observe \( X_1 = 10 \) in stage 1; the conditional type I error rate is \( A(10, 20, 0.3) = 0.3920 \). With the new sample size, \( n'_2 = 23 \), the critical value is \( R'_2(10) = 8 \) from (5). Thus in the first scenario, \( H_0 \) is rejected with 19 or more positive responses, but in the second scenario, the number of necessary positive responses is 18.

In the second scenario above, the conditional type I error rate is 0.3920 at \( X_1 = 10 \), and \( H_0 \) would be rejected with a stage 2 \( P \)-value that is less than or equal to 0.3920. When compared with the usual unconditional type I error rate, \( \alpha \), the conditional type I error rate sometimes seems very high. It may be non-intuitive that \( H_0 \) is rejected with such a high (conditional) type I error rate. The unconditional type I error rate can be viewed in (2) as a weighted average of the conditional type I error rate. Thus, when \( P_{\pi_0}[X_1 = x_1] \) is very small for a particular \( x_1 \), the conditional type I error rate at this \( x_1 \) may be very high, and the unconditional type I error rate is still protected.
4.2. Inference when sample size is changed

As shown in Section 4.1, even when the stage 2 sample size is changed, hypothesis testing can be conducted with the conditional \( P \)-value and conditional type I error rate. These conditional quantities may not be intuitive because they cannot be compared directly with the usual unconditional type I error rate, \( \alpha \). Together with the original motivation for computing unconditional \( P \)-value, confidence interval and an estimate of \( \pi \), this compels us to make unconditional inference when the stage 2 sample size is changed.

The formula for \( P \)-value in (4) needs to be extended because this formula is only valid when the stage 2 sample size is not changed. The unconditional \( P \)-value may be expressed as follows:

\[
p_p = \sum_{x_1=R_1}^{n_1} P_{\pi_0}[X_1 = x_1] cp(x_1, x_2, n_2, \pi_0) \tag{6}
\]

where \( cp(x_1, x_2, n_2, \pi_0) \), introduced in Section 4.1, is the conditional \( P \)-value for testing \( H_0: \pi \leq \pi_0 \). It is only observed at one \( (x_1, x_2) \), but it needs to be extended to the entire range of \( X_1 \in [R_1, n_1] \) to evaluate (6). When the stage 2 sample size is constant and not changed (Section 3.1), this extension is based on \( X_t \). When, however, the stage 2 sample size is changed, we cannot extend the conditional \( P \)-value based on \( X_t \). It is demonstrated in Section 4.1 that the same value of \( X_t \) may or may not lead to the rejection of \( H_0 \) depending on the sample paths. Moreover, as per the work of Banerjee and Tsiatis [14], if the planned stage 2 sample size is not constant for different values of \( X_1 \) extending the conditional \( P \)-value based on \( X_t \) would not make sense.

We propose the following approach for extending \( cp(x_1, x_2, n_2, \pi) \) to the potential values of \( X_1 \). We find a conditional power function, \( A(x_1, n_2, \pi^*) \), for some \( \pi^* \) that goes through the observed conditional \( P \)-value. It requires solving numerically for \( \pi^* \) such that \( A(x_1, n_2, \pi^*) = cp(x_1, x_2, n_2, \pi_0) \). This \( A(x_1, n_2, \pi^*) \) function can be extended to the potential values of \( x_1 \) using (1) with the original \( n_2 \) and the original \( R_2(X_1) \). We propose that the sample paths that lead to the same \( \pi^* \) have the same magnitude of evidence against \( H_0 \). It then becomes possible to order different sample paths with different \( x_1 \) and the realized sample size for stage 2 by comparing \( \pi^* \). The smaller the \( \pi^* \), the more extreme the evidence against \( H_0 \). This ordering is coherent with the hypothesis testing procedure described in Section 4.1; the \( P \)-value based on this ordering is smaller than the type I error rate if and only if the null hypothesis is rejected. To compute the \( P \)-value, we replace \( cp(x_1, x_2, n_2, \pi_0) \) in (6) by \( A(x_1, n_2, \pi^*) \) so that

\[
p_p = \sum_{x_1=R_1}^{n_1} P_{\pi_0}[X_1 = x_1] A(x_1, n_2, \pi^*) \tag{7}
\]

Suppose that we observe in the same example that \( X_1 = 7 \) and \( X_2' = 10 \) (\( n_2 = 23 \)). Then the conditional \( P \)-value is 0.1201, and \( H_0 \) is not rejected because the conditional type I error rate at \( X_1 = 7 \) is 0.0480. We compute to find that \( \pi^* = 0.3491 \). We can then extend the conditional power to the different potential values of \( X_1 \). This is represented by the bold line, \( A(x_1, n_2, 0.3491) \), in Figure 1 which also shows the conditional type I error rate, \( A(x_1, n_2, 0.3) \), and the conditional power at the original alternative, \( A(x_1, n_2, 0.5) \). Finally, we find the \( P \)-value using (7) to be \( p_p = 0.0828 \).

A confidence interval and a reasonable point estimate of \( \pi \) can be obtained by ‘inverting the hypothesis testing’ as described in Sections 3.2 and 3.3. A 90 per cent confidence interval is a collection of \( \pi_0' \) such that \( H_0: \pi = \pi_0' \) would be rejected by the sample path. We can use (7) with \( \pi_0 = \pi_0' \) to compute a \( P \)-value for testing \( H_0: \pi = \pi_0' \). We note that different \( \pi^* \) are used for different
Figure 1. Extension of the observed conditional $P$-value to other $X_1$ values based on $\pi^*$. The solid line is the conditional type I error rate, and the dotted line is the conditional power under the alternative. The conditional $P$-value is indicated by the solid square. The bold line is the conditional power function that goes through this point.

values of $\pi_0'$. And the value of $\pi_0'$ that makes the $P$-value $=0.5$ can be used as a reasonable estimate of $\pi$. This method can be applied even when the stage 2 sample size is not a constant and is changed from the planned one, but the actual coverage probability of such confidence intervals and the properties of such an estimator are largely unknown and would need further exploration.

For the current example, $X_1=7$ ($n_1=19$) and $X_2'=10$ ($n_2'=23$), a 90 per cent confidence interval for $\pi$ is $\left(0.282, 0.546\right)$. And the value of $\pi_0'$ which gives a $P$-value of 0.5 in this example is 0.405.

5. DISCUSSION

The simplicity of Simon’s designs may account for their popularity. Our literature review revealed that, frequently, the inference procedures used with Simon’s designs are often not corrected to account for these designs’ adaptive nature.

Also, when the sample size is changed, the inference procedures become more complicated. In this paper, we have shown how to make a preferred inference taking into account the planned
and actual sample sizes when a Simon’s design is used. When the actual sample size is the same as the planned sample size, the inference procedure is rather simple. However, when the actual sample size is different from that planned, we need to take into consideration both the realized and the planned sample sizes in order to compute the proper P-value and confidence interval of the response rate.

The concept of conditional power is well studied in the context of adaptive phase III clinical trials, and it is applicable to the phase II methodologies that we have considered in this paper. When the sample size is changed, the critical value needs to be updated so that the conditional type I error rate is not inflated. As shown in this paper, more obvious methods of updating the design may not control the type I error rate.

Computing a confidence interval or a reasonable point estimate has always been a challenge in the two-stage methodology. In phase III clinical trial methodology, interval and point estimates have been well studied for simple group sequential methods [29]; however, when the sample size for stage 2 is not a constant or when it is changed from the planned one, computing a confidence interval or a reasonable point estimate is not straightforward. We have presented one method which can be applied even when the stage 2 sample size is not a constant and sample size is changed, but more research is needed to study the characteristics of these estimates.

Finally, we have developed and made available a web-based program to compute a P-value, a point estimate and a confidence interval for the response rate from data obtained from a Simon’s design. The website address is http://biostat.mc.vanderbilt.edu/TwoStageInference.

ACKNOWLEDGEMENTS

The authors wish to thank the associate editor and two referees for their careful and thoughtful comments, which greatly improve the content and presentation of our work. The research of Koyama was supported in part by the VUMC Discovery Grant Program.

REFERENCES