Open Access Journal of Clinical Trials

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

The effects of heterogeneity on Simon Phase II clinical trial design operating characteristics

This article was published in the following Dove Press journal:
Open Access Journal of Clinical Trials
3 July 2010
Number of times this article has been viewed

Christopher N Barnes[1]
Shesh N Rai[2]

[1]Department of Bioinformatics and Biostatistics, University of Louisville, [2]Biostatistics Shared Facility, James Graham Brown Cancer Center, Louisville, Kentucky, USA

Abstract: The homogeneity assumption of a Simon Phase II clinical trial is commonly violated due to excess variation in the response known as response heterogeneity. Using a general framework to model heterogeneity, we investigate its effects on the operating characteristics of the Simon trial design using the standard practice of averaging responses. We show that, under heterogeneity and averaging, the Simon designs have higher than expected errors which may result in false negative and false positive Phase II outcomes.

Keywords: Simon Trial, Phase II, heterogeneity

Introduction

Simon Phase II designs are single arm trial designs used to estimate efficacy for an experimental treatment. The primary assumption for this type of trial is the assumption of response homogeneity. Response homogeneity is defined as the variance of the response being bounded by the variance of a binomial distribution given a response rate, π.[1] In practice, this assumption may be violated with the response variation greater than expected. An explanation for the increase in response variation or response heterogeneity is the existence of unique subgroups in the population.[2–4] The standard practice in Phase II trials to deal with heterogeneity has been to conduct separate subgroup trials or to ignore the heterogeneity by conducting a single trial using an averaged subgroup response.[2,5–8] Though recent designs have included methodology to handle heterogeneity, the application of the methodology is still in its infancy and Simon designs account for the majority of Phase II trials.[9] This paper examines the effects of heterogeneity on the Simon designs using the practice of averaging subgroup response rates.

In contrast to a single response rate for a homogeneous population, the response rate for a heterogeneous population can be deconstructed as a response profile, a vector of subgroup response rates. Using the response profile, three scenarios, simple averages, weighted averages, and weighted averages with accrual differences, will be used to combine the profile into a single response rate and the effects of heterogeneity will be examined using a systematic evaluation platform.

The paper is organized as follows. Section 2 provides an overview of Simon's Phase II designs. We describe general models to accommodate response heterogeneity in section 3. Section 4 presents the set of operating characteristics used to measure the effects of heterogeneity. In the penultimate section, we study the properties of the

Correspondence: Christopher Barnes
University of Louisville, Clinical and Translational Research Center, 505 S. Hancock #209, Louisville, KY 40202, USA
Tel +1 502 852 4030
Tel +1-502-852-3731
Fax +1 502 852 2356
Email chris.barnes@louisville.edu

design parameters using a limited simulation study with discussion and concluding remarks in section 6.

Simon trial designs

Simon Phase II trial designs use two stages to allow for futility of the alternative hypothesis between stages; minimizing the number of patients subjected to a non-efficacious treatment.[1] Let n_i be the sample size for stage $i = 1, 2$, where the total sample size is $n = n_1 + n_2$; and x and y be the sum of positive responses to a treatment in stage 1 and stage 2, respectively, for the experimental treatment. The Simon design assumes that the sum of responses follows a binomial distribution: $x \sim b(n_1,\pi_1)$ and $y \sim b(n_1,\pi_1)$, with variance $Var(x) = n\pi_1(1-\pi_1)$ and $Var(y) = n\pi_1(1-\pi_1)$ with response rate π_1. If the sum of responses for the experimental treatment in the first stage is not larger than a critical value r_1, the trial is stopped for futility; otherwise the trial proceeds to stage two enrolling an additional n_2 patients. Once all of the patients have been evaluated, the sum of responses over both stages, $x + y$, is compared to a second critical value, r. If the sum of responses is not larger than r, then the treatment is estimated to not have the desired effect; otherwise the experimental treatment is estimated to be efficacious with a response rate of $\pi_1 > \pi_1$.[1,10] The values for the critical values are computed given target nominal values for the operating characteristics, the type I error, eg, false positive error, and type II error, eg, false negative error.

Heterogeneity model

Let π_i be the response probability for the *ith* subgroup for $i = 1,2\ldots,g$ mutually exclusive subgroups. A subgroup response profile can be constructed, $\boldsymbol{\pi}_T = (\pi_{T1}, \pi_{T2}, \pi_{Tg})$, for $T = \{0,1\}$ corresponding to the null and alternative treatments respectively, where $\boldsymbol{\pi}_T$ is a vector composed of g subgroups, and there exists $\pi_i \neq \pi_{i'}$ for some $i \neq i'$. Additionally, let $\mathbf{w} = (w_1, w_2,\ldots, w_g)$ define the population subgroup proportions or weights.

Let the historical baseline response rate be denoted by $\pi_0^* = \min_g\left(\pi_{0i}\right)$ and the baseline treatment effect be denoted by $\delta^* = \text{mean}_g\left(\delta_{1i}\right)$. Note that separate location measures, minimum, and mean are used for the baseline response and baseline treatment effect. The use of the minimum response rate ensures that the historical fixed effects of heterogeneity in the model are positive. For estimation purposes, we define $\delta^* = \delta$ such that δ is the target treatment effect used in a Simon trial design ignoring subgroups.

Furthermore, let η_i be the prognostic response heterogeneity between subgroup i and the baseline historical response such that $\eta_i \geq 0$ and let τ_i be the predictive heterogeneity of the treatment effect over the baseline treatment effect such that $\tau_i \in$ R. Then,

$$\pi_{Ti} = \pi_0^* + \eta_i + (\delta^* + \tau_i)I(T = 1) \quad \text{where } 0 \leq \pi_{Ti} \leq 1 \tag{1}$$

defines a model for heterogeneity with indicator function $I(\cdot)$. Response heterogeneity can be divided into three classes, historical response heterogeneity (HRH), assumed response heterogeneity (ARH), and general response heterogeneity (GRH), based on the source of the response heterogeneity.[11] For all $i \neq i'$,

$$\pi_{0i} \neq \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i \neq \eta_{i'} \text{ and } \tau_i = \tau_{i'} = 0 \text{ such that } \delta_i = \delta_{i'}, \tag{2}$$

defines the HRH class and

$$\pi_{0i} = \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i = \eta_{i'} = 0 \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'} \tag{3}$$

defines the ARH class. In both classes, each experimental treatment subgroup response rate is unique. The variation in the experimental response in (2) is attributed to variation in the standard response rate through the known historical differences, η_i, between subgroup i and the baseline subgroup with a homogeneous treatment effect. The variation in (3) is attributed to variation in the treatment effect, τ_i, through a treatment-subgroup interaction when comparing the treatment effect in subgroup i to the baseline treatment effect with a homogeneous historical response rate.

The third class, GRH, relaxes the unique response rate constraint. A mixture of prognostic and predictive heterogeneity can result in nonunique experimental response rates. The etiology of each subgroup's heterogeneity is the basis for the subgroup construction and is assumed to be unique. GRH is defined as follows. There exists some $i \neq i'$ for which

$$\pi_{0i} \neq \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i \neq \eta_{i'} \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'}. \tag{4}$$

the variation in (4) is attributed to both heterogeneity in the standard response rates and the treatment effects.

To simulate the full range of each class of heterogeneity, a second component to heterogeneity, heterogeneity imbalance, is needed. Heterogeneity imbalance is a measure of the mean difference between subgroup population proportions or between accrual weights,

$$\hat{I} = \begin{cases} |w_i - w_{i'}| & g = 2 \\ \left(\sum_{i=1}^{C_{g,2}} |w_i - w_{i'}| \right) / C_{g,2} & g \geq 3 \end{cases}$$

where $C_{g,2}$ is the combination of g pairwise elements. The simplest case is balanced population proportions where $\hat{I} = 0$. To distinguish between population heterogeneity and accrual heterogeneity, $\hat{I}a$ will be used to denote accrual heterogeneity.

To estimate the response profiles, three scenarios, simple averages, weighted averages, and weighted averages with accrual differences, will be used. The simple and weighted averages are respectively defined by

$$\begin{aligned} \bar{\pi}_T &= \sum_{i=1}^{g} \left(\pi_i + \delta_i I\left(T = 1 \right) \right) / g \quad \text{and} \\ \bar{\bar{\pi}}_T &= \sum_{i=1}^{g} w_i \left(\pi_i + \delta_i I\left(T = 1 \right) \right). \end{aligned} \tag{5}$$

To allow for uncertainty in the true population profiles or differences in accrual when the weighted average method is applied, given a population profile with imbalance $\hat{I}$, the accrual profile, a = $(a_1, a_2, \ldots, a_s)$, is generated such that $Ia \in (I \pm \partial a)$ where ∂a is the estimated divergence from the true population proportion. For a given divergence, the resulting accrual profiles maintain the same mean population heterogeneity imbalance.

Operating characteristics

In a Simon Phase II trial, the hypothesis of interest is to test $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$ eg, $\pi = \pi_1$ where $\pi_1 = \pi_0 + \delta^*$ under the type I and II errors. In terms of observed positive responses, (x, y) and critical value, r, the type I and type II errors are defined by

$$\begin{aligned} \alpha &= E\left[x + y > r \,|\, \pi = \pi_0 \right] \quad \text{and} \\ \beta &= E\left[x + y \leq r \,|\, \pi = \pi_1 \right]. \end{aligned} \tag{6}$$

Under heterogeneity, the construction of the conditioning response rates in (6), though averaging, is not a unique process; multiple combinations of response and population profiles can result in a common response rate. Thus, an additional condition must be placed on the error construction, a specific combination of weights and response profile satisfying an averaging constraint,

$$\begin{aligned} \tilde{\alpha} &= E\left[E\left[x + y > r \,|\, \pi = \pi_0 \right] | \, S \right] \quad \forall i \neq i' \, s_{[i]} = s_{[i']} \\ &\text{and} \\ \tilde{\beta} &= E\left[E\left[x + y \leq r \,|\, \pi = \pi_1 \right] | \, S \right] \quad \forall i \neq i' \, s_{[i]} = s_{[i']} \end{aligned} \tag{7}$$

where $S = s[i]$ is a specific combination of response and weight profile satisfying an averaging constraint, e.g. a partition of the complete space of possible weight*response profiles.

To illustrate the complexity in this problem, we will examine how to construct an error rate through simulation. Table 1 displays possible weight*response profiles that satisfy equation (5.1) given a 40:60 scheme and $\pi_0 = 0.35$. Error rates are means eg, expected values. For example, under a binary model, given a response rate, sample size, and a critical value, $\{\pi, n, r\}$ respectively, we can compute the type I error or size of the test as follows through simulation

$$\alpha^* = 1/b \sum_{i=1}^{b} I\left(\alpha^*_i < \alpha \right); \alpha^*_i = \tag{8}$$

where b is the number of simulations. The type II error is similarly constructed using π_1 in place of π_0. We can rewrite equation (8) in a second form using an indicator variable, I,

$$\alpha^* = \left(1/s \sum_{i=1}^{s} \alpha^*_i < \alpha \right) = E\left[E\left[\alpha \,|\, s, \pi_T \right] \right] \tag{9}$$

where α is the target type I error. If one chooses to partition the above simulation into, say, $s = 4$ subsimulations, the errors could still be constructed by taking the mean of the subsimulation errors.

$$\tilde{\alpha} = \frac{1}{s} \sum_{i=1}^{s} \left(\alpha^*_{[i]} < \alpha \right) = E\left[E\left[x + y > r \,|\, \pi = \pi_0 \right] | S \right] \tag{10}$$

which is equal to

$$\tilde{\alpha} = \frac{1}{s} \sum_{i=1}^{s} I\left(\alpha^*_{[i]} < \alpha \right) \tag{11}$$

Table 1 Multiple weight*response profiles satisfying response rate constraint

W_1	W_2	π_{01}	π_{02}
0.40	0.60	0.09	0.52
0.40	0.60	0.53	0.23
0.40	0.60	0.44	0.29
0.40	0.60	0.20	0.45

where [*i*] is the *ith* partition of a total of *s* partitions. Under heterogeneity, there is not an exact analogy, eg, equation (10) = equation (11) is not guaranteed since there is a conditioning present on the original expectation. In (10), the composition of the conditioning is exactly the same across all subsimulations, a homogeneous condition. Under an unequal subgroup assumption, the conditioning is the response*weight profile which results in a heterogeneous conditioning. For example, given the first line of Table 1, (0.09, 0.52), a type I and type II error exist. Type I and Type II errors also exist for each of the remaining lines following (8).

To compute the overall errors given a weight profile, one must first find the errors for each weight*response profile or partition from (8). Then using (10), the mean of the partitions, under heterogeneity, can be interpreted as the errors given a specific weighting scheme,

$$E[E[\alpha \mid \pi_1, \pi_2, w_1, w_2] \mid w_1, w_2] \tag{12}$$

from a clinical standpoint, (12) does not make much sense. When a single arm Phase II trial is run under latent heterogeneity, only an average null response is known. The trial is used to estimate if the treatment will increase the response rate a clinically meaningful difference above the null response rate. By using the partition mean definition, the trial parameters are not guaranteed to control the errors for a specific weight*response profile, only on average across all possible weight profiles.

A more appropriate estimate for the errors would be to use the form in (9). In this case, the errors given latent heterogeneity are

$$\tilde{\alpha} = \frac{1}{s}\sum_{i=1}^{s} I\left(\alpha^*_{[i]} > \alpha\right) \quad \tilde{\beta} = \frac{1}{s}\sum_{i=1}^{s} I\left(\beta^*_{[i]} > \beta\right) \tag{13}$$

if a trial is designed to control the errors in (13), then the trial is guaranteed to control the errors at a specific level for every weight*response profile as opposed to controlling the errors on average. This difference is clinically substantial.

Equation (13.2) provides evidence for the possible failure of Phase II trials when *ex vivo* evidence would suggest otherwise and equation (13.1) provides evidence for the failure of a Phase III from a subsequent successful Phase II. The distribution of type I and II error estimates describe the strength of the underpowering or oversizing of the trial. Error estimate distributions which have more mass centered on the nominal error are of less concern than location shifted distributions where the mass of the distribution is centered further from the nominal errors.

Simulation and results

For simplicity, the number of subgroups in the simulations was chosen to be $g = \{2, 4\}$. Given a combination of population and response profile, the type I and type II errors were computed using B_1 = 10000 Monte Carlo iterations. Due to the multiplicity of combinations of response*weight profiles with a common mean response and to allow π_{Ti} where $\pi_{Ti} > \pi_{Ti'}$ for $i \neq i'$ to be uniformly distributed across the *g* subgroups, $(B_2 \mid g = 2) = 40\,000$ and $(B_2 \mid g = 4) = 100\,000$ Monte Carlo iterations were conducted; for example, $(B_2 \mid g = 2) = 4$ and $\pi_S = 0.25$ using a simple average can result in

$$\left(p_1, p_2, \pi_{S1}, \pi_{S2}\right) = \left\{\begin{matrix}(.1,.9,.30,.20) & (.1,.9,.40,.10) \\ (.1,.9,.20,.30) & (.1,.9,.10,.40)\end{matrix}\right\}.$$

A sample of population proportion profiles or weights was chosen to cover a cases of heterogeneity imbalance in the range $\hat{I} = (0, 0.98)$ for the two subgroup simulations and $\hat{I} = (0, 0.48)$ for the 4 subgroup simulations and was simulated as follows:

1. Under HRH or ARH, given the population profile for an imbalance *I*, the first $(g - 1)$ historical response rates, π_{0i}, were randomly generated from a uniform distribution, $\pi_{01}, \pi_{02}, \ldots, \pi_{0(g-1)} \sim U\left(0, \bar{\pi}_0 + \delta^*\right)$ where $\left(\bar{\pi}_S, \delta^*\right)$ are specified, for example $\left(\bar{\pi}_S, \delta^*\right) = \left(0.25, 0.15\right)$. The parameters for the uniform distribution are problem specific and are subject to the constraints $0 \leq \pi_{0i} \leq 1$ for all *i*. The *gth* null response rate was generated to satisfy (5) depending on the type of averaging. The alternative response rate was constructed in a similar fashion for the HRH and ARH classes. Under GRH, the odds ratio of each subgroup was constrained to equal the odds ratio for the Simon design such that,

$$OR_0 = \frac{\pi_1 / \left(1 - \pi_1\right)}{\pi_0 / \left(1 - \pi_0\right)} = OR_i = \frac{\pi_{1i} / \left(1 - \pi_{1i}\right)}{\pi_{0i} / \left(1 - \pi_{0i}\right)}$$
$$= \ldots = OR_g = \frac{\pi_{1g} / \left(1 - \pi_{1g}\right)}{\pi_{0g} / \left(1 - \pi_{0g}\right)}$$

solving for π_{1i} given π_{0i}. Then, $\delta_i = \pi_{1i} - \pi_{0i}$.

2. If accrual is allowed to diverge from the population profile, an accrual profile is constructed for each subgroup to replace the population profile, a = (a_1, a_2, …, a_{g-1}) ~ tN (p_i, 1) and $a_g \sim 1-\sum_{i=1}^{g-1} a_i$ where truncation occurs for the first (g–1) subgroups at ($p_i \pm \partial p$).
3. Given a population or accrual profile and a response profile, simulate multinomial random variables $n_{11}, n_{21}, \ldots, n_{g1}$ with fixed sample size n_1 and cell probabilities $\pi_T = (\pi_{T1}, \pi_{T2}, \ldots, \pi_{Tg})$.
4. For values of $(N_{11}, N_{21}, \ldots, N_{g1}) = (n_{11}, n_{21}, \ldots, n_{g1})$, simulate binomial random variables x_{Ti} with sample size n_{i1} and response rate π_{Ti}. Then $x_T = \sum_{i=1}^{g} x_{Ti}$ is compared to the critical value r_1 derived from the Simon trial design using the target mean response rates and nominal errors. If $x_T \leq r_1$, then the trial is stopped for futility.
5. If $x_T > r_1$, repeat steps (3–4) for the second stage, n_2 to determine y; otherwise $y_{Ti} = 0$. Compare $x_T + y_T$ to the critical value r from the Simon trial design. If $x_T + y_T > r$, then the null response rate is rejected.
6. Repeat steps (2–5) for $B_1 = 10\,000$ simulations and $T = (0, 1)$. Then, $\left(\sum_{b=1}^{B} I\left(x_0 + y_0 > r\right)/B\right) | \pi = \pi_0$ is the type I error of the test and $\left(\sum_{b=1}^{B} I\left(x_1 + y_1 \leq r\right)/B\right) | \pi = \pi_1$ is the type II error of the test.
7. Repeat steps (1–6) for B_2 combinations of response and population profiles. Construct the actual type I and type II errors using equation (13).

The first simulation compared the effect of varying levels of heterogeneity imbalance using simple averages for a 2 subgroup trial. The target type I and type II errors are (α, β) = (0.10, 0.20). Table 2 displays the errors with corresponding 95% quantile intervals for each class of heterogeneity. Under all three classes of heterogeneity and a heterogeneity imbalance of $I \leq 0.20$, the errors approximate the nominal errors. When the imbalance increases, $I > 0.20$ under HRH and GRH, the errors exceed the nominal errors with increasing divergence as the imbalance increases. Under ARH, the type I error approximates the nominal error with the type II following a similar, but less extreme divergence pattern as HRH and GRH. As the imbalance increases, the ranges of error estimates increase with the exception of the ARH type I estimates which maintain a constant quantile interval irrespective of the imbalance. The effect of heterogeneity is most pronounced on the type I error range under HRH and more pronounced on the pseudo type II error range under GRH. Under an unknown response profile for 2 subgroups,

Table 2 Size and power for each class of heterogeneity by heterogeneity imbalance with corresponding 95% quantile and Monte Carlo intervals for a 2 subgroup example using simple averaging and 40, 000 iterations

Class	*I*	Actual Error I	95% QI	Actual Error II	95% QI
HRH	0.02	0.10	(0.08, 0.11)	0.20	(0.18, 0.22)
	0.20	0.11	(0.04, 0.20)	0.21	(0.11, 0.32)
	0.40	0.13	(0.01, 0.34)	0.22	(0.06, 0.46)
	0.60	0.16	(0, 0.50)	0.25	(0.03, 0.61)
	0.80	0.20	(0, 0.65)	0.28	(0.02, 0.76)
	0.98	0.23	(0, 0.76)	0.31	(0.01, 0.86)
ARH	0.02	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.20	0.10	(0.09, 0.11)	0.20	(0.14, 0.28)
	0.40	0.10	(0.09, 0.11)	0.21	(0.09, 0.37)
	0.60	0.10	(0.09, 0.11)	0.22	(0.06, 0.47)
	0.80	0.10	(0.09, 0.11)	0.23	(0.04, 0.58)
	0.98	0.10	(0.09, 0.11)	0.24	(0.03, 0.67)
GRH	0.02	0.10	(0.08, 0.11)	0.23	(0.19, 0.30)
	0.20	0.11	(0.04, 0.20)	0.24	(0.14, 0.46)
	0.40	0.13	(0.01, 0.34)	0.26	(0.07, 0.66)
	0.60	0.16	(0, 0.50)	0.30	(0.03, 0.83)
	0.80	0.20	(0, 0.65)	0.33	(0.01, 0.94)
	0.98	0.23	(0, 0.76)	0.36	(0, 0.98)

the mean probability that trial is moderately to extremely oversized is 22%, $|\hat{\alpha} - \alpha| \geq 0.04$, and the mean probability that the trial is underpowered is 42%, $|\hat{\beta} - \beta| \geq 0.04$.

To further identify the effect of heterogeneity, Tables 3 and 4 display the probability distributions for the oversizing or underpowering of the trial. Under HRH and GRH, as the heterogeneity imbalance increases, the mass of the error estimate distributions location shifts increasingly further to the left resulting in larger divergences from the nominal errors. This results in strong negative effects of heterogeneity on the trial operating characteristics. For example, for $I = 0.20$ under HRH, the majority of oversized trials are in the range of (0.10, 0.12), a small divergence from the nominal errors. When $I = 0.40$ and $I = 0.80$, the majority of oversized trials are in the ranges of (0.2, 0.3) and (0.4, 1) respectively, substantial divergences from the nominal error and of high concern to the trial conduct; a similar pattern is seen with the pseudo type II errors. The exception is the oversized trials under ARH. Irrespective of the heterogeneity imbalance, the majority of oversized trials are only slightly oversized in the range of (0.10, 0.12). This would imply that even though the trials are oversized, the effect of the heterogeneity is minimal on the type I error.

Table 5 displays the results for 4 subgroups. Similar results are seen comparing the 2 and 4 subgroups examples assumptions, but the divergence between actual and nominal

Table 3 Distribution of actual type I error for each class of heterogeneity and heterogeneity imbalance for a 2 subgroup example. α_{MC} is the upper bound of the Monte Carlo error bound for the target type I error

Class	*I*	Distribution of Actual Type I Error						
		(α_{MC}–0.12)	(0.12–0.14)	(0.14–0.18)	(0.18–0.2)	(0.2–0.3)	(0.3–0.4)	>0.4
HRH	0.02	0.31	0.01	0	0	0	0	0
	0.20	0.09	0.10	0.17	0.08	0.04	0	0
	0.40	0.05	0.05	0.09	0.04	0.17	0.09	0
	0.60	0.03	0.03	0.06	0.03	0.12	0.12	0.12
	0.80	0.02	0.03	0.04	0.02	0.09	0.07	0.22
	0.98	0.01	0.03	0.04	0.01	0.08	0.06	0.27
ARH	0.02	0.26	0	0	0	0	0	0
	0.20	0.26	0	0	0	0	0	0
	0.40	0.26	0	0	0	0	0	0
	0.60	0.26	0	0	0	0	0	0
	0.80	0.26	0	0	0	0	0	0
	0.98	0.26	0	0	0	0	0	0
GRH	0.02	0.31	0.01	0	0	0	0	0
	0.20	0.09	0.10	0.17	0.08	0.03	0	0
	0.40	0.05	0.05	0.09	0.04	0.17	0.09	0
	0.60	0.03	0.03	0.06	0.03	0.12	0.10	0.12
	0.80	0.02	0.03	0.04	0.02	0.09	0.08	0.22
	0.98	0.01	0.03	0.04	0.01	0.08	0.06	0.27

errors occurs earlier, $I \approx 0.1$. The distributions of actual errors are more highly location shifted to the left in the 4 subgroup simulation compared to the 2 subgroup simulation resulting in a mean probability that the trial is oversized of 28% and a probability that the trial is underpowered of 47%.

The second scenario is the weighted average, Table 6. Under HRH and ARH, the actual errors maintain the target errors with the quantile confidence intervals only slightly larger than the Monte Carlo error bounds. The mass of the actual error distributions are in the range of (0.10, 0.12) and (0.20, 0.22) respectively, a minimal divergence between target and actual errors. Under weighted averages, the effect of heterogeneity is minimal, but not absent, on the operating characteristics of the Simon trial.

To allow for the uncertainty in either the true proportions or the accrual, two levels of error were introduced during

Table 4 Distribution of actual type II error for each class of heterogeneity and heterogeneity imbalance for a 2 subgroup example. β_{MC} is the upper bound of the Monte Carlo error bound for the target type I error

Class	*I*	Distribution of Actual Type II Error						
		(β_{MC}–0.22)	(0.22–0.24)	(0.24–0.28)	(0.28–0.3)	(0.3–0.4)	(0.4–0.5)	>0.5
HRH	0.02	0.36	0.02	0	0	0	0	0
	0.20	0.08	0.09	0.16	0.07	0.08	0	0
	0.40	0.04	0.05	0.08	0.04	0.17	0.12	0.12
	0.60	0.03	0.03	0.06	0.03	0.11	0.10	0.25
	0.80	0.03	0.03	0.04	0.02	0.08	0.08	0.31
	0.98	0.01	0.02	0.03	0.01	0.07	0.07	0.35
ARH	0.02	0.37	0.01	0	0	0	0	0
	0.20	0.18	0.15	0.12	0.01	0.01	0	0
	0.40	0.10	0.09	0.14	0.05	0.10	0.01	0.10
	0.60	0.06	0.07	0.12	0.05	0.13	0.05	0.06
	0.80	0.04	0.06	0.08	0.05	0.13	0.08	0.13
	0.98	0.03	0.05	0.07	0.02	0.13	0.09	0.18
GRH	0.02	0.29	0.21	0.19	0.06	0.03	0	0
	0.20	0.06	0.07	0.07	0.05	0.14	0.10	0.10
	0.40	0.04	0.02	0.07	0.01	0.11	0.08	0.24
	0.60	0.03	0.02	0.03	0.03	0.07	0.07	0.32
	0.80	0.02	0.03	0.01	0.02	0.08	0.05	0.35
	0.98	0.01	0.02	0.01	0.01	0.07	0.03	0.38

Table 5 Actual errors for each class of heterogeneity by heterogeneity imbalance with corresponding 95% quantile and Monte Carlo intervals for a 4 subgroup example using simple averaging and 100, 000 iterations

Class	*I*	Actual Error I	95% QI	Actual Error II	95% QI
HRH	0.02	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.10	0.13	(0.02, 0.35)	0.20	(0.06, 0.39)
	0.20	0.12	(0.01, 0.37)	0.24	(0.06, 0.53)
	0.30	0.24	(0, 0.88)	0.27	(0, 0.84)
	0.40	0.28	(0, 0.97)	0.30	(0, 0.96)
	0.48	0.31	(0, 0.99)	0.32	(0, 0.99)
ARH	0.02	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.10	0.10	(0.09, 0.11)	0.20	(0.10, 0.30)
	0.20	0.10	(0.09, 0.11)	0.21	(0.10, 0.38)
	0.30	0.10	(0.09, 0.11)	0.22	(0.02, 0.57)
	0.40	0.10	(0.09, 0.11)	0.24	(0.01, 0.71)
	0.48	0.10	(0.09, 0.11)	0.26	(0, 0.81)
GRH	0.02	0.10	(0.09, 0.11)	0.24	(0.20, 0.31)
	0.10	0.13	(0.02, 0.35)	0.25	(0.09, 0.61)
	0.20	0.13	(0.01, 0.37)	0.28	(0.05, 0.74)
	0.30	0.23	(0, 0.88)	0.32	(0, 0.99)
	0.40	0.28	(0, 0.97)	0.35	(0, 1)
	0.48	0.31	(0, 0.99)	0.36	(0, 1)

patient accrual, $\partial a = \{0.05, 0.1\}$. The accrual heterogeneity imbalance was allowed to vary between 0 and 5% and between 0 and 10% of the population heterogeneity imbalance. The accrual difference can be attributable to accrual divergence or error in proportion estimation. Table 7 shows the results for $g = 2$ subgroups with an accrual divergence parameter of 5%.

Table 6 Actual errors for each class of heterogeneity by heterogeneity imbalance with corresponding 95% quantile for a 2 subgroup example using weighted averaging and 40, 000 iterations

Class	*I*	Actual Error I	95% QI	Actual Error II	95% QI
HRH	0.02	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.20	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.40	0.10	(0.08, 0.12)	0.20	(0.18, 0.22)
	0.60	0.10	(0.08, 0.12)	0.20	(0.18, 0.22)
	0.80	0.10	(0.08, 0.12)	0.20	(0.18, 0.22)
	0.98	0.10	(0.09, 0.12)	0.20	(0.18, 0.22)
ARH	0.02	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.20	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.40	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.60	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.80	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
	0.98	0.10	(0.09, 0.11)	0.20	(0.18, 0.22)
GRH	0.02	0.10	(0.09, 0.12)	0.23	(0.18, 0.24)
	0.20	0.10	(0.09, 0.12)	0.22	(0.18, 0.25)
	0.40	0.10	(0.08, 0.12)	0.21	(0.18, 0.25)
	0.60	0.10	(0.08, 0.12)	0.21	(0.18, 0.24)
	0.80	0.10	(0.08, 0.12)	0.20	(0.18, 0.24)
	0.98	0.10	(0.09, 0.12)	0.20	(0.18, 0.24)

Table 7 Simon Optimal design with s = 2 subgroups population using weighted average with accrual differences, $\partial a = 0.05$, for 500, 000 simulations

	∂a	*I*	Actual Error I	95% CI	Actual Error II	95% CI
HRH	0.10	0.02	0.10	(0.07, 0.14)	0.20	(0.16, 0.24)
		0.20	0.10	(0.07, 0.13)	0.20	(0.16, 0.24)
		0.40	0.10	(0.08, 0.14)	0.20	(0.16, 0.24)
		0.60	0.10	(0.07, 0.12)	0.20	(0.16, 0.24)
		0.80	0.10	(0.08, 0.13)	0.20	(0.16, 0.24)
ARH	0.10	0.02	0.10	(0.09, 0.11)	0.20	(0.17, 0.23)
		0.20	0.10	(0.09, 0.11)	0.20	(0.17, 0.23)
		0.40	0.10	(0.09, 0.11)	0.20	(0.17, 0.23)
		0.60	0.10	(0.09, 0.11)	0.20	(0.17, 0.23)
		0.80	0.10	(0.09, 0.11)	0.20	(0.17, 0.23)
GRH	0.10	0.02	0.10	(0.08, 0.13)	0.20	(0.16, 0.26)
		0.2	0.10	(0.07, 0.12)	0.21	(0.17, 0.26)
		0.4	0.10	(0.07, 0.12)	0.21	(0.17, 0.38)
		0.6	0.10	(0.07, 0.13)	0.22	(0.17, 0.39)
		0.8	0.10	(0.07, 0.14)	0.23	(0.18, 0.34)

The actual errors approximated the nominal errors in almost every case with the exception being under GRH pseudo type II errors. The reason for this divergence is unknown at this time. The distributions of the pseudo errors are more dispersed than the weighted average method due to the variation in accrual. The strength of the errors is increased when comparing the error estimate distributions between weighted averages and weighted averages with accrual divergence.

Discussion

There has been a substantial increase in the identification of disease subtypes over the past 5 years. For example with the increase in usage of genomic markers, diseases, once thought of as having a homogeneous response across a population, are showing response stratification as the specificity of the disease process increases through modern diagnostic techniques. The existence of these subgroups can lead to an increase in the variance of the response adding a new confounder to the conduct of Simon phase II trials. While the Simon designs are very powerful designs under the homogeneity assumption, the increase in variation or heterogeneity has a strong effect on the operating characteristics of the design.

The simulations have shown that under heterogeneity and an averaging practice to satisfy the input for a Simon design, the probabilities that the design is underpowered or oversized are larger than expected. Under simple averaging, as the level of heterogeneity imbalance increases, the actual errors diverge from the target errors with the mass of the error estimate distributions location shifted to the left, indicating

a larger divergence between the target and actual errors and substantial impact on the trial outcomes.

Using a weighted average minimizes the divergence between error types, but caution should still be advised. The quantile 95% confidence intervals for the actual errors are greater than the Monte Carlo error bounds and do not always maintain the nominal error. So even under a weighted scheme, a trial might fail due to heterogeneity. This divergence is attributed to the fact that the true response rate under the null hypothesis is not equivalent to the target response rate used in the design A clinician will not know immediately after trial conduct that what the true power or true size are to know how close they were to a successful outcome. Additionally, as the number of subgroups increase, the divergence between errors types increases. The application of the weighted average method is not always feasible. In practice, clinicians may not have a very accurate estimate of the true population proportions or may not have an accurate estimate for the accrual at the time of the study. The addition of an accrual confounder further increases the divergence between nominal and pseudo error rates under weighted averaging. Even small divergences are of concern given the relatively small sample sizes in Phase II trials.

The importance of these results may be correlated with the overall failure percentages of Phase II and Phase III trials. The percentage of Phase II trials that fail today exceeds 30% in fields such as oncology.[8,12] The etiology of the failure of the trials is partially unknown. The treatment in question may not have the response hypothesized or the trial may, in fact, be underpowered; though the clinician is unaware of this fact due to ignoring the heterogeneity though averaging. For example, in the simulations, approximately 32% of the 2 subgroup trials with only a slight heterogeneity imbalance, $I = 0.2$, would be underpowered, with a minimum type II error of 0.24. The Phase II trial could fail due to this underpowering or the trial could still have been a success, but unknowingly oversized with probability ~28% with a type I error of a minimum of 0.14. The subsequent Phase III might fail due to this oversizing. A clinician needs to address the issue of whether or not a 28%–32% probability of reaching the wrong outcome is acceptable if an averaging method is used.

The authors present this work to provide evidence that under the assumption of heterogeneity, the use of a Simon design presents risks that are beyond acceptable for most clinicians and may provide some information as to the reason behind the high number of Phase II trials that fail when *ex vivo* evidence suggests otherwise or the subsequent failure of Phase III trials after successful Phase IIs. Methods need to be constructed that can handle heterogeneity and still retain the simplicity and ease of use of the popular Simon designs.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 1989;10(1):1–10.
2. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Stat Med*. 2005;24(17):2597–2611.
3. Thall PF, Wathen JK. Bayesian designs to account for patient heterogeneity in phase II clinical trials. *Curr Opin Oncol*. 2008;20(4):407–411.
4. Behrendt CE, Gehan EA. Treatment – subgroup interaction: An example from a published, phase II clinical trial. *Contemp Clin Trials*. 2009;30(3):279–281.
5. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Stat Med*. 2008;27(15):2802–2815.
6. Ayanlowo AO, Redden DT. A two stage conditional power adaptive design adjusting for treatment by covariate interaction. *Contemp Clin Trials*. 2008;29(3):428–438.
7. Gadbury GL, Iyer HK, Allison DB. Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat*. 2001;11(4):313–333.
8. Tuma RS. Examining heterogeneity in phase II trial designs may improve success in phase III. *J Natl Cancer Inst.* 2008;100(3):164–166.
9. Ye F, Shyr Y. Balanced two-stage designs for phase II clinical trials. *Clin Trials*. 2007;4(5):514–524.
10. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials.* 3rd ed. New York: Springer Verlag; 1998.
11. Barnes CN, Rai SN. Modeling heterogeneity in Phase II clinical trials. *American Journal of Biostatistics*. 2010;1(1):9–16.
12. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *J Clin Oncol*. 2005;23(19):4450–4457.