5-2010

A phase II two stage clinical trial design to handle latent heterogeneity for a binary response.

Christopher Noel Barnes
University of Louisville

A PHASE II TWO STAGE CLINICAL TRIAL DESIGN TO HANDLE LATENT HETEROGENEITY FOR A BINARY RESPONSE

By

Christopher Noel Barnes

B.S., University of Louisville 2004

M.S., University of Louisville, 2007

A Dissertation

Submitted to the Faculty of the

School of Public Health and Information Sciences

Of the University of Louisville

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

Department of Bioinformatics and Biostatistics

University of Louisville

Louisville, Kentucky

May 2010

A PHASE II TWO STAGE CLINICAL TRIAL DESIGN TO HANDLE LATENT HETEROGENEITY FOR A BINARY RESPONSE

By

Christopher Noel Barnes

B.S., University of Louisville 2004

M.S., University of Louisville, 2007

A Dissertation Approved on

April 15, 2010

by the following Thesis Committee:

_____

Dissertation Director, Shesh N. Rai

_____

Jason Chesney

_____

Dongfeng Wu

_____

Guy Brock

_____

Seong Kim

DEDICATION

This dissertation is dedicated to my parents

Dr. George R. Barnes

and

Dr. Inessa Levi

who have inspired me to follow in their footsteps.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Shesh Rai, for his invaluable guidance and patience with me over the past couple of years. I would also like to thank Drs. Ziad Kanaan, Susan Galandiuk and Kelly McMasters for the opportunities they have provided me to hone my analytical skills and provide me with real world experience while working collaboratively on their projects.

ABSTRACT

A PHASE II TWO STAGE CLINICAL TRIAL DESIGN TO HANDLE LATENT HETEROGENEITY FOR A BINARY RESPONSE

Christopher N. Barnes

April 15, 2010

Phase II clinical trial are generally single arm trial where a homogeneity assumption is placed on the response. In practice, this assumption may be violated resulting in a heterogeneous response. This heterogeneous or overdispersed response can be decomposed into distinct subgroups based on the etiology of the heterogeneity. A general classification model is developed to quantify the heterogeneity. The most common Phase II trial design used in practice is the Simon 2-stage design which relies on the assumption of response homogeneity. This design is shown to be flawed under the assumption of heterogeneity with errors exceeding the target trial errors. To correct for the error inflation, a modification is made to the Simon design if heterogeneity is detected after the first stage trial conduct. The trial sample size is increased using an empirical estimate for the variance inflation factor and the trial is then completed with design parameters constructed through the posterior predictive Beta-binomial distribution given the first stage results. The new design, denoted the 2–stage Heterogeneity Adaptive (2HA) design, is applied to a two subgroup problem under latent heterogeneity. Latent heterogeneity represents the most general form of heterogeneity, no information is known prior to trial conduct. The results, through simulation, show that the target errors can be

maintained with this modification to the Simon design under a wide range of heterogeneity.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The primary assumption for most Phase II single arm binary trials 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 (Simon 1989). Many single arm Phase II trials do not adhere to this assumption in practice. When the variance of the response, denoted x which corresponds to the number of patients with a positive response, exceeds the binomial variance,

$$V(x) > n\pi(1-\pi), \tag{1}$$

the response is deemed a heterogeneous response (Williams 1982; Yamamoto and Yanagimoto 1994; Collet 2003). The common structure of this heterogeneous response is a response profile of disjoint subgroups, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_g)$, for $i = 1, 2, \ldots, g$ subgroups where π_i is the response probability for the *ith* subgroup and there exists at least two distinct subgroup response rates, $\pi_i \neq \pi_{i'}$ for some $i \neq i'$. In contrast, the response in a homogeneous population follows a single response rate, where $\pi = \pi_i = \pi_{i'}$ for all $i \neq i'$. Subgroup membership is defined by a single or multiple set of markers (London and Chang 2005; Thall and Wathen 2008; Behrendt and Gehan 2009). The markers can be composed of clinicopathologic features such as age, gender, diagnostic

or prognostic markers such as baseline insulin levels or single/multiple genomic markers such as the BRCA1 gene in Breast cancer.

A common practice in clinical trials when heterogeneity is assumed and the markers are known is to use a simple or weighted mean of the response profile of the subgroups to compute a single response rate which adheres to the homogeneity assumption (Green 1982; Gadbury and Iyer 2000; Emerson, Kittelson et al. 2007; Emerson, Kittelson et al. 2007; Ayanlowo and Redden 2008; Thall and Wathen 2008; Tuma 2008; Wathen, Thall et al. 2008). The weights are derived from either the known population proportions of each subgroup or estimated from a random sample of patients.

This leads to one of two averaging constraints on the response profile. Let π_T for $T = \{0,1\}$, the null response rate and alternative response rate respectively, be the response rate, then the average constraints are defined as

$$1/g \sum_{i=1}^{g} \pi_{Ti} = \pi_T , \tag{2}$$

or

$$\sum_{i=1}^{g} w_i \pi_{Ti} = \pi_T \tag{3}$$

for simple average and g subgroups and for a weighted average, where $w_i = n_i / n$ is the weight for subgroup i, n_i is the number of patients in subgroup i for a total of $\sum_{i=1}^{g} n_i = n$ patients in the sample, respectively.

Using methods that rely on the homogeneity assumption when heterogeneity is true will lead to biased inferences (Russek-Cohen and Simon 1997), incorrect early stopping of the trial (Thall, Wathen et al. 2003; Thall and Wathen 2008; Wathen, Thall et al. 2008) or a subsequent failure of the Phase III trial resulting in a substantial loss of resources (Rosner, Stadler et al. 2002; Stadler 2007; Tuma 2008). This is primarily due to the departure of the trial data distribution from the model distribution from which the trial parameters are constructed, the binomial distribution. It will be shown that this approach, when applied to the most common Phase II trial design, the Simon 2-stage trial, will result in unbounded errors, false positive or false negative trial conclusions, dependent on a combination of the magnitude of difference between the subgroup responses and the difference in subgroup weights.

A second method when heterogeneity is present is to conduct multiple trials, one for each subgroup. This will result in a heavy strain on trial resources especially for early development Phase II trials. Due to possible low patient accrual in one or more trials, trials may not be completed; losing valuable information on the treatment effect over the entire population. Conducting multiple trials ignores a fundamental assumption of the motivation for a single trial; all patients share a common disease state. It is assumed that the response rate in one subgroup will be partially correlated with the response rate in the other subgroups. Secondly, the subgroups must be known in advance of the trial conduct to conduct multiple trials which is not always a practical situation.

In the last few years, multiple methods have been developed to account for response heterogeneity by quantifying the structure of the subgroups in the test statistic

(London and Chang 2005; Thall and Wathen 2008). Two examples are briefly mentioned. The simplest form, the unconditional stratified test, assumes a stratified response based on known subgroups and modifies the Binomial test statistic into the form of a stratified log-rank test (London and Chang 2005). The resulting test has a global hypothesis, either the compound/treatment provides efficacy evidence to move onto further targeted Phase II testing or Phase III testing or it does not.

Bayesian methods have also been developed which rely on hierarchical models or ANCOVA models to model the structure of the subgroups (Thall, Wathen et al. 2003; Wathen, Thall et al. 2008). The Bayesian methods employ the desirable characteristic of local hypothesis tests, rejection of the efficacy hypothesis on a subgroup level allowing some subgroups to succeed while others may fail. Secondly, the Bayesian methods minimize the overall sample size as compared to running multiple trials by sharing response information across the subgroups when making decisions on individual subgroups. Drawbacks are that Bayesian methods will use considerably more computational resources and do not rely on fixed sample size estimates. The limiting drawback to implementing these designs in actual trial conduct and the remaining methods described in the literature is that all the methods rely on the assumption that the composition of the subgroups is known prior to trial conduct. The methods provide no methodology for when the subgroups are latent prior to trial conduct.

Recently, there has been a shift in focus to randomized Phase II designs to help mitigate heterogeneity in the response (Lee and Feng 2005). Randomized trials can provide a mechanism to estimate the source of the heterogeneity and the type of

heterogeneity. A major drawback to the randomized designs is the substantial increase in trial resources, usually a doubling of trial resources to reject a global hypothesis. The use of a randomized design is not always practical at such an early stage of estimating treatment efficacy due to patient accrual issues and will not be considered in this paper.

In practice, the composition of the subgroups is not known or only partially hypothesized. Latent subgroups are a more common problem in clinical trials and may provide an etiology for the high failure rate of Phase II trials. Phase II trials are not conducted unless there is substantial *ex vivo* evidence of compound/treatment efficacy. In practice, many Phase II trials still fail when this evidence is present; presenting the issue of whether the trial failure rests on inadequate efficacy of the compound/treatment, inadequacy of the trial design, or inaccurate estimates of the hypothesized response. We focus on the second issue, inadequacy of the trial design as a possible solution to the high failure rate of Phase II trials.

Before developing a new trial design, the structure to heterogeneity must be quantified. We have developed a classification model to quantify response heterogeneity, through the subgroups, into three classes, historical response heterogeneity (HRH), assumed response heterogeneity (ARH) and general response heterogeneity (GRH). These classes can help to detect when a trial may fail due to heterogeneity.

HRH is composed of known subgroups. In simplest terms, the subgroups are known either from responses to similar treatments, known biological motivations or can be estimated from the response in the control group of a randomized trial design denoted as the null response. Under HRH, the null response, *e.g.* response under no treatment, is

heterogeneous and the treatment effect is homogenous resulting in a heterogeneous response structured by the heterogeneity of the null hypothesis response.

In contrast, ARH assumes a homogeneous null response and a heterogeneous treatment effect. Under ARH, no known or latent subgroups exist on prior treatment, but a Treatment x Marker effect is identified causing the treatment and thus response under the treatment, denoted the alternative response, to vary by this Treatment x Marker subgroup composition. In both the previous classes, the alternative responses are unique. Each disjoint subgroup can be identified from a unique alternative treatment response rate.

A generalization of the first two classes is general response heterogeneity. GRH is composed of possibly both heterogeneous historical response and heterogeneous treatment effects. GRH does not always result in uniquely identifiable subgroups through the alternative response, but results in unique subgroups through the source of the heterogeneity. Multiple different combinations of null response and treatment effect can result in the same alternative response.

Under the context of a single stage design, in order to determine the composition of subgroups, a pre-clinical analysis would have to be conducted on a set of patients which would entail exposing the patients to the compound/treatment to determine response. A second set of patients would be used in the resulting trial. This is not an optimal use of trial resources. The first set of patients, in effect, can be construed as a separate trial in which the data is thrown away; not providing response information for use in the actual trial. A more suitable solution would be to conduct the "pre-clinical"

analysis during the trial; hence, minimizing time and patient resources. No information would be lost, all patients that undergo treatment would used in estimating response. The two stage designs of Simon provide a natural break for this analysis, between stages. While the two stage process is a suitable solution to this problem and comprises the majority of all conducted Phase II trials, the use of the binomial distribution as the model distribution is not appropriate.

We develop a two stage design which begins as the popular Simon 2-stage design and is adapted to accommodate heterogeneity if heterogeneity is identified between the conduct of the two stages. If no source of heterogeneity is identifiable, the trial continues on under the Simon design; otherwise an adaption is made and the trial is evaluated using new adaptive trial parameters.

The paper is organized as follows. Chapter two introduces the basic two stage design of Simon, the heterogeneity model and trial error construction. Chapter three provides a literature review of the current methods to handle heterogeneity with non-latent heterogeneity. Chapter four introduces the main components of the new trial design, subgroup identification, heterogeneity tests, the trial's model distribution, and finally, the trial algorithm. Chapter five investigates the operating characteristics of the Simon design and new trial design under heterogeneity with concluding remarks and future direction in Chapter six.

CHAPTER 2

SIMON DESIGN, HETEROGENEITY MODEL AND ERRORS

2.1 Simon Phase II Designs

The basic two stage binary Phase II trial design was first implemented by Gehan (Gehan 1961). Shultz modified the Gehan design to require a minimum of at least one response in the first stage with equal size sample sizes in both stages (Schultz, Nichol et al. 1973). The Gehan design can allow no response in the first stage. Simon later popularized the Shultz design by allowing unequal size sample sizes in the stages and constructing a search algorithm to determine the optimal and minimax designs which meet a set of sample size optimization criteria (Simon 1989).

For simplicity, the term treatment denotes a compound, treatment or regimen. Let x be the realized data in stage one with (r_1, n_1) as the critical value and sample size for stage one, and y be the realized data in stage two with (r, n) as the critical value and sample size for stage one and two combined. The trial parameters, (r_1, n_1, r, n), are constructed to estimate if the trial response rate under treatment, π, is greater than or equal to a clinically relevant target response rate, denoted the alternative response, $\pi_1 = \pi_0 + \delta$ where π_0 is the null response under no treatment and δ is the treatment effect, or formally, $H_0 : \pi < \pi_1$ vs. $H_1 : \pi \geq \pi_1$, the null and alternative hypothesis respectively.

If the sum of responses for the treatment in the first stage is not larger than the stage one critical value, $x \leq r_1$, the trial is stopped for futility; otherwise, the trial proceeds to stage two enrolling an additional $n - n_1$ patients. Once all of the patients have been evaluated, the sum of responses over both stages is compared to a second critical value. If the sum of responses is not larger than the stage one + stage two combined critical value, $x + y \leq r$, then the treatment is estimated to not have the desired effect; otherwise, the novel treatment is estimated to be promising with a response rate of $\pi \geq \pi_1$.

The construction of the parameters of the trial, (r_1, n_1, r, n), is dependent on the target errors of the trial known as the type I error or size of the trial, α, and type II error or 1-power of the trial, β. As such, the power of the trial is $1 - \beta$. The critical values and sample sizes for each stage are chosen from a set of possible designs constrained to satisfy the type I and type II errors per

$$P(\textit{reject } H_0 \mid \pi = \pi_T) = Bin(r_1 \mid n_1, \pi_T) + \sum_{x=r_1+1}^{\min(n_1, r)} bin(x \mid n_1, \pi_T)\, Bin(r - x \mid n_1, \pi_T) \qquad (4)$$

where *bin* is the binomial probability mass distribution and *Bin* is the binomial cumulative distribution for treatments $T = \{0,1\}$, the null and alternative hypothesis or null and alternative response rate, respectively.

In practice to determine the parameters, n_1, n, r_1, r, a sample size for stage 1 is first chosen such that $P(\text{reject } H_0 \mid \pi = \pi_0, \mathrm{N}_1 = \mathrm{n}_1) \in (.50, .80)$, also known as the probability of

early termination (PET). Using an iterative algorithm, given (n_1, r_1), a total sample size is selected to satisfy (4) under the pre-specified target type I and type II errors. This process is repeated to find a set number, say 50, solutions that satisfy the error constraints. Two of the solutions are then selected as the minimax and optimal designs. The optimal design is the design that minimizes the expected sample size,

$$EN(H_0) = n_1 + (1 - PET)(n - n_1), \tag{5}$$

under the null hypothesis over all possible designs and the minimax design is the design that minimizes *EN* over all designs with the minimum total sample size, n.

Under a Simon design with no heterogeneity, no type I error is spent in the first stage. This is due to the single bound of the critical value. The bound is for futility only. Only a percentage of power is spent in the first stage. This is evidenced in the form of (4) where the second component on the right hand side is a weighted sum weighted by the "power" spent in the first stage. Most Phase II designs follow this approach, only a futility bound in the earlier stages, since the primary goal of a Phase II trial is to estimate if the treatment is promising for further testing, not to establish is the treatment is efficacious.

2.2 A Model for heterogeneity

Response heterogeneity in a population can be modeled by deconstructing the response rate into subgroups to form a response profile, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_g)$, composed of g subgroups where π_i is the response rate for the ith subgroup and there exists $\pi_i \neq \pi_{i'}$

for some $i \neq i'$; in contrast, $\pi_i = \pi_{i'}$ for all $i \neq i'$ in a homogeneous population. The resulting subgroup model provides the basic platform to compare methodology for heterogeneous responses.

Let $\boldsymbol{\pi}_{\mathbf{T}} = \left(\pi_{T1}, \pi_{T2}, \ldots, \pi_{Tg}\right)$ be the vector of subgroup responses for $i = 1, 2, \ldots, g$ subgroups where π_{Ti} is the response rate in subgroup i for treatment $T = \{0,1\}$. $T = 0$ denotes the known standard/historical treatment response, e.g. the null response, and $T = 1$ denotes the hypothesized experimental treatment response, e.g. the alternative response.

In addition, let the baseline historical response rate for the historical response profile be denoted by $\pi_0^* = \arg\min_g \left(\pi_{0i}\right)$. Furthermore, let η_i be the prognostic response heterogeneity between subgroup i and the baseline historical response, τ_i be the predictive heterogeneity in treatment effect over the baseline treatment effect, $\delta^* = \arg\min_g \left(\delta_{1i}\right)$ where δ_{1i} are the treatment effects for each subgroup, such that,

$$\pi_{Ti} = \pi_0^* + \eta_i + (\delta^* + \tau_i) I(T = 1), \tag{6}$$

where $0 \leq \pi_{Ti} \leq 1$, defines a subgroup mixture model for heterogeneity (Barnes and Rai 2010). $I(\cdot)$ is a membership indicator.

The historical response heterogeneity, η_i, is a fixed prognostic effect while the treatment heterogeneity, τ_i, is a predictive random effect. Using equation (6), the

classification of response heterogeneity rests on the structure of the historical response profile and the treatment effect profile. To quantify the range of response heterogeneity, three classes, historical response heterogeneity (HRH), assumed response heterogeneity (ARH), and general response heterogeneity (GRH), are constructed. For all $i \neq i'$,

$$\pi_{0i} \neq \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i \neq \eta_{i'} \text{ and } \tau_i = \tau_{i'} = 0 \text{ such that } \delta_i = \delta_{i'}, \tag{7}$$

defines the HRH class and

$$\pi_{0i} = \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i = \eta_{i'} = 0 \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'} \tag{8}$$

defines the ARH class. In both classes, the experimental treatment response rates are unique.

The third class, GRH, relaxes the unique response constraint. A mixture of prognostic and predictive heterogeneity can result in non-unique experimental responses. 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

$$\pi_{0i} \neq \pi_{0i'} \text{ and } \pi_{1i} \neq \pi_{1i'} \text{ where } \eta_i \neq \eta_{i'}, \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'}. \tag{9}$$

In equation (7), a known covariate exists for which a prior historical response profile can be constructed. The prior distribution of historical response rates, given the historical covariate, is hypothesized to be consistent in the current trial. Heterogeneity in the experimental response profile is attributed to the different known historical response rates, $\pi_{0i} \neq \pi_{0i'}$. The treatment effects are homogeneous across the subgroups, $\delta_i = \delta_{i'}$. In contrast to HRH, the heterogeneity in equation (8), is quantified through

heterogeneous treatment effects, $\delta_i \neq \delta_{i'}$, where the estimated historical response rates are homogeneous, $\pi_{0i} = \pi_{0i'}$. The heterogeneity is measured by the inequality of the treatment effects between subgroups due to a covariate-treatment interaction as opposed to the inequality of historical rates as in (7).

The general form of response heterogeneity, GRH, is a composite of both of the previous classes of response heterogeneity. The general form (9) occurs when both the historical response rates and treatment effects are hypothesized to be heterogeneous. For example, under a three subgroup model, historically gender, (M, F), leads to different historical response rates, $\pi_{01} = \pi_{02} = \pi_{0M}$, and $\pi_{03} = \pi_{0F}$ where $\pi_{0M} \neq \pi_{0F}$. A biomarker present in males is hypothesized to lead to a further differentiation of response rates, male biomarker present and male biomarker absent, resulting in the following three possible response models,

$$\pi_{01} = \pi_{02} \neq \pi_{03} \text{ and } \begin{cases} \pi_{11} \neq \pi_{12} \neq \pi_{13} \\ \pi_{11} \neq \pi_{12} = \pi_{13} \\ \pi_{11} = \pi_{12} \neq \pi_{13} \end{cases} \tag{10}$$

The prognostic heterogeneity differs between gender, $\eta_1 = \eta_2 \neq \eta_3$, with a predictive heterogeneity only affecting the males, $\tau_1 \neq \tau_2$ and $\tau_3 = 0$. The first possible experimental response model results in three unique response rates. While the remaining two models result in two unique response rates with the effect of the male biomarker, present or absent, providing the same experimental response rate as for females. When no information is known about the structure of the heterogeneity, it is appropriate to

assume a general class structure. For this reason, the focus, in evaluating a new trial design under latent heterogeneity, will rest on the GRH class of heterogeneity.

2.3 Heterogeneity model example

To illustrate the different classes of heterogeneity, the following hypothetical example is provided. A trial is conducted to determine the response rate of drug A to treat early to moderate stage colon cancer, stage I-III. The researchers wish to test $H_0 : \pi = \pi_0 = .3$ against $H_1 : \pi > .3$ with a target treatment effect of $\delta = .2$ resulting in the alternative response of $\pi_1 = .5$. Table 1(a-d) provides four possible scenarios under a $g = 3$ subgroup trial for different groups of researchers testing the same drug. For simplicity, the sample sizes of the subgroups are assumed to be equal.

The first scenario, table 1a, is an example of HRH. Research group I knows that historically Drug A leads to a response profile based on cancer stage for a similar disease, breast cancer. This prognostic difference is assumed to be consistent in the current trial due to the similarity of pathways being targeted between the two cancers. The historical response profile for the standard treatment is $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \pi_{03}) = (.4,.3,.2)$ with $\bar{\pi}_0 = .3$ constructed from a baseline historical response rate of $\pi_0^* = .2$ and a historical heterogeneity effect of $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3) = (.2,.1,0)$. The objective is to test for a common treatment effect, $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3) = (.2,.2,.2)$ such that $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3) = (0,0,0)$, in a historically heterogeneous response resulting in the experimental response profile $\boldsymbol{\pi}_1 = (\pi_{11}, \pi_{12}, \pi_{13}) = (.6,.5,.4)$ with $\bar{\pi}_1 = .5$.

The second scenario is an example of ARH, table 1b. Group II contends that there is no historical precedent for the usage of drug A on colon cancer, but hypothesize a predictive difference based on a combination of two biomarkers resulting in three clinically relevant subgroups, both biomarkers present, both absent and one present. The response profile for the standard treatment is homogeneous, $\boldsymbol{\pi}_0 = (.3,.3,.3)$ with $\boldsymbol{\eta} = \mathbf{0}$, and it is the inequality of the treatment effect that is the source of the heterogeneity, $\boldsymbol{\delta} = (.3,.2,.1)$ such that $\boldsymbol{\tau} = (.25,.15,.05)$, leading to an experimental response profile of $\boldsymbol{\pi}_1 = (.6,.5,.4)$ with $\bar{\pi}_1 = .5$.

The third example is an example of GRH, table 1c. Group III suspects that there is both a prognostic effect based on cancer staging and a predictive effect based on the biomarkers. There is both a heterogeneous historical treatment effect, $\boldsymbol{\eta} = (.33,.03,0)$ such that $\boldsymbol{\pi}_0 = (.51,.21,.18)$, and heterogeneous treatment effect with $\boldsymbol{\tau} = (0,.20,.03)$ such that $\boldsymbol{\delta} = (.09,.29,.12)$. The experimental response profile is then $\boldsymbol{\pi}_1 = (.6,.5,.4)$ with $\bar{\pi}_1 = .5$.

The fourth group hypothesizes a more complex interaction between cancer stage and biomarker status as a combination of HRH and ARH only affecting a subsample of the subgroups, table 1d. Historically, the researchers feel evidence only provides a two subgroup prognostic difference in the efficacy of the drug, stage I *vs.* Stage II-III with the status of the biomarker only affecting the second group, Stage II-III. This results in $\boldsymbol{\pi}_0 = (.35,.275,.275)$ with $\boldsymbol{\eta} = (.075,0,0)$. The interaction between

biomarker status and the second prognostic subgroup leads to $\boldsymbol{\tau} = (0, .15, 0)$ for an overall experimental response profile of $\boldsymbol{\pi}_1 = (.55, .55, .40)$ with treatment effect profile $\boldsymbol{\delta} = (.20, .275, .125)$.

Table 1 Numerical example of three classes of response heterogeneity.

a: HRH							b: ARH						
π_S^*	η_i	π_{Si}	δ^*	τ_i	δ_i	π_{Ei}	π_S^*	η_i	π_{Si}	δ^*	τ_i	δ_i	π_{Ei}
.20	.20	.40	.20	0	.20	.60	.30	0	.30	.05	.25	.30	.60
.20	.10	.30	.20	0	.20	.50	.30	0	.30	.05	.15	.20	.50
.20	0	.20	.20	0	.20	.40	.30	0	.30	.05	.05	.10	.40

c: GRH I							d: GRH II						
π_S^*	η_i	π_{Si}	δ^*	τ_i	δ_i	π_{Ei}	π_S^*	η_i	π_{Si}	δ^*	τ_i	δ_i	π_{Ei}
.18	.33	.51	.09	0	.09	.60	.25	.075	.35	.20	0	.20	.55
.18	.03	.21	.09	.20	.29	.50	.25	0	.275	.125	.15	.275	.55
.18	0	.18	.09	.03	.12	.40	.25	0	.275	.125	0	.125	.40

2.4 Heterogeneity Imbalance

A second component to heterogeneity, heterogeneity imbalance, is a measure of the mean difference between subgroup population proportions or between accrual weights. Let $\mathbf{w} = (w_1, w_2, \ldots, w_g)$ be the vector of weights for $i = 1, 2, \ldots, g$ subgroups, then a measure of the information provided by $\mathbf{w}$ is the absolute difference in magnitude between the subgroup weights, denoted the heterogeneity imbalance,

$$I = \begin{cases} |w_i - w_{i'}| & g = 2 \\ \left(\sum_{i=1}^{C_{g,2}} |w_i - w_{i'}| \right) \Big/ C_{g,2} & g \geq 3 \end{cases} \tag{11}$$

where $C_{g,2}$ is the combination of g pairwise elements.

The simplest case is balanced population proportions where $I = 0$. To distinguish between population heterogeneity and accrual heterogeneity, *Ia* will be used to denote accrual heterogeneity. Heterogeneity imbalance will be used as a method to classify the range of heterogeneity and as a component to increase the sample size in the latter sections of the paper.

2.5 Clinical trial errors

Trial parameters are constructed such that the trial errors are maximized with respect to the target errors. Under a frequentist design, the target errors are the Type I and Type II errors. The errors are composed of four joint probabilities which specify the complete trial outcome space (Lee and Zelen 2000). The joint probabilities quantify the probability of the trial outcome, acceptance or rejection of the alternative hypothesis, and the population truth, the population response rate is greater than or equal to the target response rate or less than the target response rate,

$$\begin{aligned} P(R-) &= P\left(\text{Reject } \mathrm{H}_1, \pi < \pi_1\right); \quad P(A-) = P\left(\text{Accept } \mathrm{H}_1, \pi < \pi_1\right); \\ P(R+) &= P\left(\text{Reject } \mathrm{H}_1, \pi \geq \pi_1\right); \quad P(A+) = P\left(\text{Accept } \mathrm{H}_1, \pi \geq \pi_1\right), \end{aligned} \tag{12}$$

subject to, $\sum_{i=(A,R)} \sum_{j=(+,-)} P(ij) = 1$.

The first joint probability, $P(R-)$, is the probability of rejecting the alternative hypothesis and the null hypothesis is true in the population. While the fourth joint

probability, $P(A+)$, is the probability of accepting the alternative hypothesis and the null hypothesis is not true in the population. The frequentist errors, are constructed,

$$\begin{aligned} Type\ I = \alpha &= P\left(\text{Accept } H_1 \mid \pi < \pi_1\right) = \frac{P(A-)}{P(A-)+P(R-)}; \\ Type\ II = \beta &= P\left(\text{Reject } H_1 \mid \pi \geq \pi_1\right) = \frac{P(R+)}{P(R+)+P(A+)}. \end{aligned} \tag{13}$$

2.6 Trial errors under the subgroup assumption

Under a subgroup assumption the construction of the errors is not as straightforward as in section 2.5 due to the averaging constraints which allow for a multiplicity of weight*response profiles,

$$\mathbf{w\pi} = \left(w_1\pi_1, w_2, \pi_2, \ldots, w_g\pi_g\right); \quad \sum_{i=1}^{g} w_i\pi_i = \pi, \tag{14}$$

that sum to a single fixed response rate (Barnes and Rai 2010). The usual assumption, in homogeneous Phase II trials, is that only a single response exists and given this response and a set of critical values and sample sizes, the errors can be constructed. Under a subgroup model, the assumption of the single response still exists, through the mean response rate, but there exist two levels of additional variation which can result in the single mean response rate. The first level is the weight profile. The second level is the actual response profile. Multiple different combinations of weights and response profiles can lead to a single response rate.

Under a specific single fixed response rate and within each weight profile, there are weight multiple response profiles that exist which satisfy the main response

constraint, $\pi_T = \sum_{i=1}^{g} w_i \pi_{Ti}; \quad T = 0,1$. Table 2 displays multiple possible weight*response profiles that satisfy equation (3), a weighted average, given a 40:60 scheme and $\bar{\pi}_0 = .30$.

Table 2: Multiple weight*response profiles satisfying response rate constraint

w_1	w_2	π_{01}	π_{02}	$\bar{\pi} = \sum w_i \pi_i$
.40	.60	.73	.01	.30
.40	.60	.55	.13	.30
.40	.60	.31	.29	.30
.40	.60	.24	.34	.30

To illustrate the added complexity the problem when heterogeneity exists under a mean of the weight*response profile, we will examine how to construct an error rate through simulation. Error rates are means, *e.g.* expected values. For example, under a binary model, given a response rate, sample size and a critical value, (π, n, r) respectively, we can compute the type I error as follows through simulation

$$\alpha = E\left[x > r \mid \pi = \pi_0\right] = \frac{\sum_{i=1}^{b} I\left(x > r \mid \pi = \pi_0\right)}{b} \tag{15}$$

where b is the number of simulations, x is the sum of responses with critical value r and indicator variable $I(\cdot)$.

If one chooses to partition the above simulation into, S sub-simulations or partitions denoted [s], the errors could still be constructed by taking the mean of the sub-simulation errors since each subgroup simulation is exchangeable,

$$\alpha = E\left[E\left[x > r \mid \pi = \pi_0 \right] \mid S \right] \quad \forall i \neq i' \; s_{[i]} = s_{[i']} \tag{16}$$

which is equivalent to

$$\alpha = E\left[x > r \mid \pi = \pi_0 \right] = \frac{\sum_{j=1}^{s} \left(\frac{\sum_{i=1}^{b} I\left(x > r \mid \pi = \pi_0 \right)}{b} \right)}{s} \tag{17}$$

Under latent heterogeneity, the form of the type I error in (17) is not correct since the partition is not exchangeable. In (16), the composition of the conditioning is exactly the same across all sub simulations, *e.g.* exchangeability, a homogeneous condition.

Under a heterogeneity subgroup assumption, and say for explanation, only four possible weight*response profiles existed to satisfy the averaging constraint, the conditioning is not exchangeable. Each weight*response profile results in a separate set of errors, a heterogeneous conditioning. For example, given the first line of table 2, $(.73,.01)$, a type I and type II error exist. Separate Type I and Type II errors also exist for each of the remaining weight*response profiles.

Under a subgroup assumption, S is not exchangeable. We assume that the weights are fixed. Each partition $S = s$ results in a unique partitioning of the complete space for a fixed weight profile. Under this assumption, (16) becomes

$$\alpha = E\left[E\left[x > r \mid \pi = \bar{\pi}_0 \right] \mid S \right] \quad \forall i \neq i' \quad s_{[i]} \neq s_{[i']} \tag{18}$$

where the complete space S is composed of all possible partitions satisfying the weighted average constraint

$$S = \left\{ s_{[i]} = \left(\pi_{[i]1}, \pi_{[i]2} \right) \ni \sum_{j=1}^{2} w_{[i]j} \pi_{[i]j} = \bar{\pi} \right\} \tag{19}$$

Taking the expectation, under non-exchangeable subgroups, will result in an overall double expectation that is generally bounded by the target errors. This is not appropriate under a clinical trial context.

The trial design must guarantee that the error is bounded by the target error for every non-exchangeable subgroup; the double expectation only guaranteed this on average. A clinical trial will always be conducted with a specific response profile, whether known or not known, and the trial design errors must be guaranteed to be bounded by the target error.

A more appropriate estimate for the errors under heterogeneity where non-exchangeable subgroups exist

$$\tilde{\alpha} = \frac{\sum_{i=1}^{s} I\left(\alpha_{[i]} > \alpha \right)}{s} \quad \tilde{\beta} = \frac{\sum_{i=1}^{s} I\left(\beta_{[i]} > \beta \right)}{s} \tag{20}$$

where $\alpha_{(i)}$ and $\beta_{(i)}$ are the type I and type II errors for each partition, *e.g.* a specific weight*response profile satisfying the weighted averaging constraint. The trial errors are then the mean number of times a partition error crosses the target error boundary over all possible partitions of the complete fixed weight profile space. If a trial is designed to

control the errors in (20), 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.

CHAPTER 3

LITERATURE REVIEW: EXISTING METHODS FOR HETEROGENEITY

Five methods have been developed to handle response heterogeneity in single arm Phase II clinical trials. The methods cover both frequentist and Bayesian designs. A commonality between most methods is the reliance on a known composition structure to the subgroups; not including the Beta-binomial methodology.

3.1 Unconditional and Conditional Stratified Methods

The methods proposed by London and Chang, unconditional stratified and conditional stratified methods, account for subgroups with a binary response, similar to a stratified log-rank test for time-to-event data, under a k-stage design (London and Chang 2005).

Given a known covariate with g subgroups for stages $j = 1, 2, \ldots, m, \ldots, k$, let $R_m = \sum_{j=1}^{m} \sum_{i=1}^{g} R_{ij}$ be the sum of responses across all subgroups up to an intermediate stage m where R_{ij} is the sum of responses for the *ith* subgroup in the *jth* stage. The total sample size across k stages is denoted $N = \sum_{j=1}^{k} \sum_{i=1}^{g} N_{ij}$. Furthermore, let the sampling weights be proportional to the true population profile, then the general form of the test statistic for the unconditional stratified method is

$$K_m = \frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{g}\left(R_{ij} - N_{ij}\pi_{0i}\right)\right)}{\sqrt{\sum_{j=1}^{m}\left(\sum_{i=1}^{g} N_{ij}\pi_{0i}(1-\pi_{0i})\right)}} \quad . \tag{21}$$

Sample size computation and critical value determination are completed using an iterative simulation algorithm with set percentages of type I and type II errors spent in each stage similar in development to the Simon design; see (London and Chang 2005). A set of stopping boundaries, $((l_1,u_1),(l_2,u_2),\ldots,(u_k))$, where (l_1,u_1) are the futility and efficacy boundaries for stage 1 respectively, are constructed to maintain the target type I and type II errors for the trial. This is in contrast to the Simon design where only a futility boundary exists. The final result is a sample size and test statistic(s) based on the estimates for the true population proportions of each subgroup, the sampling weights.

Since the true population proportions of the subgroups are not usually known in practice, a second form the test statistic was proposed, the conditional stratified method. The sample size and outcome of the trial are conditioned on the sampling weights, as opposed to the true proportions, of each subgroup. Conditioning equation (21) on $\left(\frac{N_{i1}}{N_1}\right)=\left(\frac{n_{i1}}{n_1}\right),\ldots,\left(\frac{N_{im}}{N_m}\right)=\left(\frac{n_{im}}{n_m}\right)$, it can be seen that both $\sum_{j=1}^{m}\sum_{i=1}^{g} n_{ij}\pi_{0i}$ and the denominator of (21) are constants given $\left(n_{i1},\ldots,n_{im},\pi_{0i}\right)$. The sum of responses up to the immediate stage m is asymptotically equivalent to K_m and the rejection region of the null hypothesis can be expressed as $R_m > r_m$ where r_m is the critical value of the test statistic for the *mth* stage. The general form of the test statistic for the *mth* stage of the conditional method is

$$P(R_m = r_m) = \sum_{\substack{r_{1m}+\ldots+r_{gm}=r_m: \\ 0\le r_{im}\le n_{im}}} \prod_{i=1}^{g} \binom{n_{im}}{r_{im}} \pi_{0i}^{r_{im}} (1-\pi_{0i})^{n_{im}-r_{im}} \quad . \tag{22}$$

The final test statistic for k stages is the sum of independent random variables,

$R_1 + R_2 + \ldots + R_m + \ldots + R_k$.

In contrast to the unconditional method, many solutions exist to (22) by varying each of the subgroup sampling weights through $\left(\frac{N_{im}}{N_m}\right) = \left(\frac{n_{im}}{n_m}\right)$ under the type I and type II error constraints. This allows for a wide range of possible accrual scenarios and results in a similar output as the initial output, before making the selection of the minimax and optimal solutions, of the Simon designs (Simon 1989).

3.2 Beta-Binomial Method

The third method, the beta-binomial distribution has been previously proposed as a model that can account for heterogeneity in binary outcome models (Makuch, Stephens et al. 1989; Yamamoto and Yanagimoto 1994; Hendriks, Teerenstra et al. 2005; Hunt and Rai 2005; Dragalin and Fedorov 2006; Young-Xu and Chan 2008). For simplicity, we assume only one stage. To allow for an increase in variation of the response over the binomial, a subgroup composition is assumed for the responses where response rates are allowed to vary, $\pi_i \sim beta(a_0, b_0)$. Then $R_{i1} \mid \pi_i$, has a binomial distribution. The marginal of R_1 is a beta-binomial with probability function,

$$P(R_1 = r_1) = \binom{n_1}{r_1} \frac{beta(r_1 + a_0, n_1 - r_1 - b_0)}{beta(a_0, b_0)} \tag{23}$$

The mean and variance are

$$E[R_1] = n_1\pi \text{ and } Var[R_1] = n_1\pi(1-\pi)\{1+\phi(n_1-1)\}; \phi = \frac{\rho}{1+\rho} \tag{24}$$

where $\pi = \frac{a_0}{a_0 + b_0}$. The parameter ρ is the correlation between the response rates and quantifies the excess heterogeneity in the response profile above the binomial distribution. If $\rho = 0$, then the variance of R_1 degenerates into the binomial variance. After estimation of the parameters (a_0, b_0), the sample size and test statistics can be calculated based on the type of difference to be detected (Hendriks, Teerenstra et al. 2005; Chow, Shao et al. 2007). It should be noted that the estimation of the parameters does not require subgroup source knowledge, prognostic or predictive, about the heterogeneity; only the estimated amount of variation.

3.3 Bayesian Hierarchical Methods

To implement Phase II designs from the frequentist perspective, a fixed response rate, whether a single rate or response profile, is specified. Alternatively, a Bayesian design incorporates a level of uncertainty in the fixed rate by assuming that the response is random through the use of prior and hyper-prior distributions. A primary design principle this approach is that the parameters of the response are not independent, but correlated similar to the beta-binomial distribution (Lee 2009). One such model is the

Bayesian hierarchical model (BHM) which assumes a hyper-parameter distribution for the priors, ψ, to model the heterogeneity and correlation of the parameters. The joint distribution of all parameters is constructed by combining the data likelihood, prior and hyper-prior distributions,

$$f(\mathbf{R},\boldsymbol{\pi},\psi) = l(R\,|\,\boldsymbol{\pi})\,p(\boldsymbol{\pi}\,|\,\psi)\,p(\psi) = \left\{\prod_{i=1}^{g} \underbrace{l(R_i\,|\,\pi_i)}_{\substack{data\\likelihood}}\ \underbrace{p(\pi_i\,|\,\psi)}_{prior}\right\} \underbrace{p(\psi)}_{hyperprior} \tag{25}$$

with trial decision making using the posterior distribution,

$$P(\boldsymbol{\pi}\,|\,\mathbf{R}) = \frac{\int f(\mathbf{R},\boldsymbol{\pi},\psi)\,d\psi}{\int f(\mathbf{R},\boldsymbol{\pi},\psi)\,d\theta\,d\psi} \tag{26}$$

Due to the intractability and high dimension of the posterior, Monte Carlo Markov Chain MCMC methods are used to compute the posterior probabilities for each stage of the trial (Gilks, Richardson et al. 1996).

The fourth heterogeneity method, Bayesian normal-binomial hierarchical model used in Thall *et. al.*, is based on the logit model (Collet 2003; Hunt and Rai 2003) and is constructed such that

$$\theta_i = logit(\pi_i) \overset{iid}{\sim} N(\mu,\sigma^2) \text{ with } \psi = (\mu,\sigma^2),\ \mu \sim N(\upsilon_1,\varphi_1^{\,2}) \text{ and } \sigma^2 \sim N(\upsilon_2,\varphi_2^{\,2}). \tag{27}$$

The subgroups are assumed to be exchangeable implying no a priori prognostic difference in response rates. The heterogeneity is assumed to be predictive.

One advantage in using the Bayesian approach is the existence of within subgroup stopping boundaries allowing for partial subgroup efficacy/futility as opposed to a global boundary, *e.g.* Simon or London and Chang methods. As such, a set of identical within subgroup stopping boundaries, due the exchangeability of the subgroups, are constructed for each stage of the trial. Once all the patients in subgroup i are evaluated, futility and efficacy stopping boundaries are applied for this subgroup,

$$P(\pi_{1i} > \pi_{0i} \mid data) < l \tag{28}$$

and

$$P(\pi_{1i} > \pi_{0i} \mid data) \geq u, \tag{29}$$

using the data from *all subgroups* to determine if a particular subgroup portion of the trial should be stopped or continue accrual until the next decision point using an appropriately small value for l and a large value for u. The values for the boundaries are usually chosen to give good operating characteristics when compared to a frequentist design. Each subgroup has an identical stopping boundary similar to running multiple simultaneous trials with the conditioning allowing the sharing of information across subgroups and minimization of resources by using the data from all subgroups to determine individual subgroup outcomes.

3.4 Bayesian ANCOVA Method

The fifth method, Bayesian normal-binomial regression model or BANCOVA model, was proposed by Wathen and Thall (2008). To compare the model with the earlier heterogeneity notation of (6), the model was reparameterized. The model,

$$logit\left(\pi_{Tg}(\theta)\right) = \xi + \sum_{i=1}^{g}\left\{\eta_i + \tau_i I(T=1)\right\} I(G=g), \tag{30}$$

is constructed with $\eta_1 = 0$ for interpretational convenience. It should be noted that the ranges of the parameters are not consistent between the heterogeneity model (6) and the model (30) which the models mean response rate on the logit scale. Model (30) has no assumption on the structure of the variance as in model (27), where $\theta_i = logit(\pi_i) \overset{iid}{\sim} N(\mu, \sigma^2)$ is assumed, modeling the mean response as opposed to both the mean and variance of the response.

The prognostic effect of subgroup g compared with the baseline subgroup, *e.g.* subgroup 1, is η_g and the predictive effect for subgroup g is τ_g. To construct the hyper-parameters for each of the priors, Wathen and Thall developed an algorithm assuming small variances for historical priors and large variances for experimental priors by equating the moments of a beta distribution to a normal distribution. For the complete hyperparameter algorithm and the logic for their assumptions, see (Wathen, Thall et al. 2008).

Once the priors have been computed, the posteriors are constructed using MCMC methods. Subgroup-specific stopping boundaries are then constructed similar to (28) and

(29) where the subgroup specific stopping boundaries (l_i, u_i) are subgroup dependent on the prognostic effect as opposed to the BHM model where the boundaries are identical.

CHAPTER 4

AN ADAPTIVE PHASE II DESIGN TO ACCOMMODATE HETEROGENEITY

We present a method to account for latent heterogeneity under the Simon design context which adapts the second stage sample size and critical value of a Simon design based on the outcome of the first stage under the presence of a heterogeneity statistic. The adaptive design denoted the 2-stage heterogeneity adaptive design (2HA) preserves the operating characteristics of the Simon trial under no heterogeneity, e.g. no change to the design, and preserves the first stage operating characteristics, moderate probability of early termination. For simplicity and due to the relatively small sample sizes in Phase II trials, detecting only two groups is attempted.

The basic algorithm is as follows, compute the Simon design parameters given a weighted average response rate, which asymptotically *e.g.* $n \to \infty$, mirrors the population response and conduct the first stage of the trial. After the first stage and the first stage criterion was met, $x > r_1$, determine if subgroups exists through a classification algorithm. If the trial fails to meet the first stage critical value, the trial has failed. The etiology of the failure is unknown; either the trial design failed due to latent heterogeneity or the probability that the response meets or exceeds the clinically relevant response is minimal. Under this scenario, no change to the second stage will result in a successful trial and will

not be considered as relevant to the design at this point. The design focuses only on trials that have met the first stage criterion.

Once the subgroups are identified, the subgroups are tested for the presence of heterogeneity. If no heterogeneity is detected, enroll the remaining patients per the Simon sample size parameters and complete the trial using the Simon critical value. If heterogeneity is detected using a liberal test, the overall sample size will be increased, with the additional sample size for the first stage included in the new second stage sample size, using an empirically derived inflation factor. A new critical value will be constructed given the new second stage sample size and that the trial has succeeded into the second stage.

The Beta-Binomial posterior predictive distribution is used as the model distribution to determine new parameters under heterogeneity. Enroll the additional patients and test the global hypothesis with the new critical value. This new design will control the errors bounded at the target errors given knowledge on the average response, π_0 and π_1, the number subgroups expected, $g = 2$.

4.1 Subgroup identification

This design relies on the ability of a classifier to find the true subgroups in the sample. Multiple methods exist for finding subgroups in supervised and unsupervised manners. A supervised classifier is one in which the true identity of the object being classified is known. For example, supervised classification can be conducted on age or gender. In both instances, the true state is known. Unsupervised classifiers are ones in

which the true state is not known and is estimated through patterns in the data. An example may be a set of unknown biomarkers. The true state of the unknown biomarkers is generally not known. Patients are grouped into subgroups based on the expression patterns of these biomarkers.

The study of supervised classifiers is a broad subject with many classifiers that fall under this category. Some examples are recursive algorithms such as random forests, machine learning algorithms such as support vector machines or statistical classifiers such as linear discriminate analysis or principal component analysis.

The most popular method for unsupervised learning is clustering algorithms. Clustering is the assignment of samples into subsets based on a distance or dissimilarity measure which measures the distance between samples based on the data (Datta 2006). Multiple types of clustering exist such as agglomerative methods, and k-means. See Romesburg for an exhaustive summary of the multiple methods that exist (Romesburg 2004).

For the purposes of this paper, the classifier is assumed to have 100% accuracy. This is to remove any variation that might be caused by the actual classifier. In the case of unsupervised classifiers, there may be a level of error associated with either the classification method based on a small sample size as is the case in Phase II trial first stages or error associated with the measurement platform such us the case with high through-put microarray platforms.

The classifiers for these reasons will be assumed to be built using supervised variables, say age and gender. In this case, the classifier will always have 100% classification accuracy. The utilization of non- perfect classifiers in the algorithm is a subject for future work.

4.2 Testing for heterogeneity

Multiple methods exist for testing the assumption of heterogeneity or overdispersion in binomial data under a grouped data assumption. The preferred method is to test for lack of fit of the data to the binomial model with parameter π (Collet 2003). We focus on global goodness-of-fit methods where the test statistic evaluates the unspecific hypothesis, model fits versus model does not fit.

The two most common test statistics are the Pearson and Deviance test statistics,

$$X^2 = \sum_{i=1}^{g}\sum_{j=1}^{2}\frac{\left(x_{ij} - n_{ij}\pi_1\right)}{n_{ij}\pi_1} \sim X^2_{(g)} \tag{31}$$

and

$$D = 2\sum_{i=1}^{g}\sum_{j=1}^{2} x_{ij}\log\left(\frac{x_{ij}}{n_{ij}\pi_1}\right) \sim X^2_{(g)} \tag{32}$$

respectively (Kuss 2002). As the number of groups increases, $g \to \infty$, the two test statistics should be approximately equal, $X^2 \approx D$. Under the context of this problem, two groups and only one sample of each group, there is lack of data, known as sparcity, which results in $X^2 \neq D$. The sparcity is due to the fact that only one example of the

data exists during a single trial, e.g. $n=1$. It has been shown that for $n \leq 5$, the Pearson test is too conservative and the Deviance test is erratically anticonservative (Kuss 2002). This undermines the use of either statistic as a robust method of determining heterogeneity.

A third method is to use a modified Pearson test statistic where the Pearson statistic family is generalized by adding an additive constant to X^2 first described by Farrington,

$$X_F^2 = \sum_{i=1}^{g} \frac{\left(x_i - n_{1i}\pi_1\right)^2}{n_{1i}\pi_1\left(1-\pi_1\right)} + \sum_{i=1}^{g} \frac{-\left(1-2\pi_1\right)}{n_{1i}\pi_1\left(1-\pi_1\right)}\left(x_i - n_{1i}\pi_1\right). \tag{33}$$

The standardized test statistic is then compared to a standard normal distribution (Farrington 1996). This method has been shown to be more stable than either the deviance or standard Pearson statistics under sparcity (Kuss 2002).

Due to the sparcity of the data, heterogeneity is determined using a liberal p-value threshold, $p \leq .30$. The motivation for using a liberal p-value threshold is that it is advantageous to err on the heterogeneity side. If heterogeneity truly does exist and the test determines no heterogeneity, the Simon trial parameters are not a good fit to the data. The reverse, the test determines heterogeneity when heterogeneity does not exist, will result in the use of the Beta-Binomial which will still provide an adequate fit to the data.

In simulation, the Farrington test had a power to detect heterogeneity above 80%. The type I error is inflated, ~30%, which is allowable since the model distribution

will still fit and will result in a only modest increase in sample size. This is shown in the results section.

4.3 Variance inflation factor

To increase the sample size in the second stage to account for the response heterogeneity requires estimation of a variance inflation factor (VIF). The standard interpretation of the VIF is as an unknown scale parameter which relates the variance of a Binomial random variable to the variance of an overdispersed Binomial random variable, a Beta-Binomial random variable, section 4.3.1. This interpretation, under a two stage trial, will not result in a robust estimate since it relies on estimation of the VIF through a Pearson or Pearson type statistic.

A second interpretation for the VIF is the inflation factor necessary to increase the sample size to account for heterogeneity, section 4.3.2. Empirical results are used to construct this definition. This method will result in a robust method that leads to a sample size that will control the trial errors at the target errors.

4.3.1 Estimation of theoretical VIF

Given the following model,

$$x_i \mid \pi_i \sim Bin(n_i, \pi_i); \quad E[\pi_i] = \pi; \quad V(\pi_i) = \phi\pi(1-\pi), \tag{34}$$

for $i = 1, 2, \ldots, g$, we can compute the variance of the observed responses, X_i,

$$
\begin{aligned}
V(X_i) &= E\left[n_i\pi_i(1-\pi_i)\right]+V(n_i\pi_i) \\
&= n_i\left(E[\pi_i]-E\left[\pi_i^2\right]\right)+n_i^2\phi\pi_i(1-\pi_i) \\
&= n_i\left(\pi_i-\phi\pi_i(1-\pi_i)-\pi_i^2\right)+n_i^2\phi\pi_i(1-\pi_i) \\
&= n_i\pi_i(1-\pi_i)\left[1+(n_i-1)\phi\right]
\end{aligned}
\tag{35}
$$

Under the special case $n_i = n$ for all i,

$$
V(X_i) = n\pi_i(1-\pi_i)\underbrace{\left[1+(n-1)\phi\right]}_{\sigma^2}, \tag{36}
$$

such that σ^2 is denoted the heterogeneity factor and ϕ is denoted the VIF. Since,

$$
E\left[X^2\right] \approx g\left[1+(n-1)\phi\right] = g\sigma^2, \tag{37}
$$

it follows that

$$
\sigma^2 = \frac{X^2}{g} \quad \text{and} \quad \hat{\phi} = \frac{\hat{\sigma}^2-1}{n-1} \tag{38}
$$

Equation (36) provides the standard interpretation of the VIF and estimation through equations (37) and (38). Under the special case of a two stage trial, where only a single sample is used, the stage one results, the estimation of the VIF through (38) will not be robust due to sparcity.

The use of the Farrington X^2 to replace the standard X^2 was allowed in section 4.2 for the heterogeneity test statistic since it is advantageous to err on the side of heterogeneity. The only error which must be controlled absolutely is the Type II error; allowing for the type I error to be exceeded. This is not an acceptable solution in the

estimation of the VIF. The VIF is directly correlated with the resulting sample size of the trial. In turn, the sample size is directly correlated with the trial errors. Allowing for inadequate control of the errors in the estimation of the VIF will result in adequate control of the trial errors.

4.3.2 Estimation of empirical VIF

A second logical approach to estimate the VIF is through the trial conduct. The increase in variation can be attributed to two components not present under a homogeneous population, the weight profile and the response profile. Under a homogeneous population, the response profile is a single response. The second interpretation of the VIF is as the minimum amount necessary to increase the sample size to account for heterogeneity.

Given the first stage results and the presence of heterogeneity, an estimate for the heterogeneity imbalance, $\hat{I}$ can be estimated through (11). The absolute magnitude of difference in the response profile can be estimated

$$\left|\hat{\pi}_{11} - \hat{\pi}_{12}\right| \tag{39}$$

where $\hat{\pi}_{1i} = \frac{x_i}{n_{1i}}$ is the estimate of the response in the *ith* subgroup with response x_i and sample size n_{1i} such that $\sum_{i=1}^{2} x_i = x$ and $\sum_{i=1}^{2} n_{1i} = n_1$. Equation (11) provides information on the weight profile while equation (39) provides information on the response profile beyond the information provided by a fixed homogeneous response rate. For simplicity, the number of subgroups is 2.

A natural estimate for the VIF for a single trial is the product of the estimate of the heterogeneity imbalance and the response rate imbalance,

$$\hat{\phi} = |w_1 - w_2||\hat{\pi}_{11} - \hat{\pi}_{12}| = \hat{I}|\hat{\pi}_1| \tag{40}$$

A new sample size to account for heterogeneity can be constructed

$$n^* = (1+\hat{\phi})n \tag{41}$$

where n is the original Simon total sample size. When the heterogeneity imbalance is non-existent, $\hat{I} = 0$. Small differences in the response profile lead to a small VIF. As either the difference in weight profile diverges or the difference in the response profile diverges, the sample size will increase.

4.4 Model Distribution

Once heterogeneity has been detected through subgroup identification, the trial no longer adheres to the assumption of a Binomial distribution. As such, the Binomial trial parameters and model are no longer valid. Two factors will determine the new model distribution, the structure of the trial and the structure of the data.

4.4.1 Predictive posterior Beta-Binomial

The structure of the trial is a two-stage process and this process should be inherently modeled in the model distribution for parameter construction. The data is structured such that

$$X_i \sim P(x_i \mid \pi_i) \sim Bin(n_i, \pi_i); \quad i = 1, 2, \ldots, g \tag{42}$$

and

$$X = \sum_{i=1}^{g} X_i \tag{43}$$

where X is the sum of responses in the first stage of the data. This two-stage process and the extra-binomial variation due to (42) and (43) can be explicitly modeled using the posterior predictive Beta-Binomial distribution.

The posterior predictive distribution (PP) quantifies the probability of a future observation of the data, y, out of m samples given some data has already been observed, x, out of n samples,

$$p(y \mid x) = \int p(y \mid x, \pi) p(\pi \mid y) \partial \pi . \tag{44}$$

In the case of a two-stage trial, the PP distribution quantifies the distribution of the stage two outcome given the stage one outcome. In addition, by treating the parameter π as random, as opposed to a fixed as in the binomial distribution, the variance of X is larger than a strictly single parameter binomial model for X.

For completeness, the composition of the Beta-binomial is repeated, removing subgroup notation for simplicity,

$$\begin{gathered} p(\pi) \sim Beta(a_0, b_0); \quad p(x \mid \pi) \sim Bin(n, \pi); \\ p(\pi \mid x) \sim Beta(\pi; a_0 + x, b_0 + n - x), \end{gathered} \tag{45}$$

where $Beta(x;a,b) = \int_0^x (\varphi)^{a-1}(1-\varphi)^{b-1} d\varphi$. The first distribution in (45) is the prior distribution of the response parameter π. The second distribution is the data likelihood which satisfies the data structure in (42). Combining the data prior with the data likelihood results in the posterior distribution $p(\pi|x)$. Heuristically, one can interpret the posterior distribution, $p(\pi|x)$, as an assumption distribution, $p(\pi)$, updated with actual data from the trial, $p(x|\pi)$.

Then, it can be easily seen that the PP distribution is Beta-binomial through the conjugateness of the Beta distribution and Binomial distribution,

$$\begin{aligned} p(y|x) &= \int p(y,\pi|x)\partial\pi \\ &= \int p(y|\pi)p(\pi|x)\partial\pi \\ &\approx Bin(n-n_1,\pi)\,Beta(\pi;a_0+x,b_0+n_1-x) \\ &= Beta(a_0,b_0)-1Beta(\pi;a_0+y,b_0+(n-n_1)-y) \end{aligned} \tag{46}$$

where $Beta(a_0,b_0) = \Gamma(a_0)\Gamma(b_0)/\Gamma(a_0+b_0)$. The extra-binomial variation or heterogeneity is modeled through the data prior (See (24)).

4.4.2 Prior specification

The standard practice in prior specification is to specify a non-informative prior. A non-informative prior will minimize the impact of a subsequently misspecified prior on the overall posterior distribution (Lee 2009). Non-informative priors are priors with large variance. The standard method to parameterize a non-informative Beta prior is to base the prior on a small sample size such as $n = \{1, 1/2, 2\}$ (Thall, Wathen et al. 2003). For

example, basing a beta prior with $\bar{\pi} = .30$ and $n = 1$ will result in the parameterization of $a_0 = .30$ and $b_0 = .70$ since

$$\bar{\pi} = a_0 / (a_0 + b_0) = .30 \quad V(\pi) = a_0 b_0 / (a_0 + b_0)^2 (a_0 + b_0 + 1) \tag{47}$$

As the sample size that the prior is based upon increases, the variance of the random response rate shrinks towards zero as seen in table 4.

Table 3 Mean and variance for different prior specifications by sample size

n	(a_0, b_0)	$\bar{\pi}$	$V(\pi)$
1	(.3,.7)	.30	.105
10	(3,7)	.30	.033
100	(30,70)	.30	<.01

The prior for our model will be based on the null response given a sample size of $n = 1$ which will result in the parameterization given in (47).

4.2.3 Beta-Binomial predictive posterior error construction

The structure of the errors in section 2.5 can be used to develop the errors for trial design using the Beta-Binomial PP distribution (Barnes and Rai 2010). The joint probability of outcome and truth in a *k*-stage trial is composed of (*k*+*1*) subspaces, *k* stage outcomes and the population truth. For a two stage design, the joint probabilities are specified as follows,

$$P(\text{outcome, truth}) = P(\text{outcome stage 1, outcome stage 2, truth}). \tag{48}$$

The joint probabilities are specified using the conditional probabilities and marginal probabilities of the (k+1) components through Bayes theorem.

$$\begin{aligned} P(A-) &= P(\pi < \pi_1, \text{Accept } H_1) \\ &= P(\pi < \pi_1, x > r_1, x+y > r) \\ &= P(\pi < \pi_1 \mid x > r_1, x+y > r)\, P(x+y > r \mid x > r_1)\, P(x > r_1) \\ &= \sum_{i=r_1+1}^{n_1} \sum_{j=r+1-i}^{n-n_1} P(\pi < \pi_1 \mid y = j, x = i)\, P(y = j \mid x = i)\, P(x = i) \end{aligned} \tag{49}$$

The joint probabilities that include an accept outcome are intuitive. This is not the case for the joint probabilities which include a reject outcome.

Under a two stage design, if the first stage criterion is not met, then the trial stops without proceeding stage two. To specify the joint probabilities under rejection, it is necessary to include the conditional probability of the second stage criterion not being met given the first stage criterion is not met; a situation which is impossible in actual trial conduct. This specification is not intuitive, but necessary to assure that the total outcome space is complete.

The probability of rejecting the alternative hypothesis and the null hypothesis is true is the sum of the product of the conditional probability that the null hypothesis is true given the first stage criteria is not met, the conditional probability that the second stage criteria is not met given that the first stage criteria is not met, and the marginal probability that the first stage criteria is not met and the product of the conditional probability that the null hypothesis is true given the trial is successful, the conditional

probability that the second stage criteria was met given the first stage criteria was met, and the marginal probability that the first stage criteria was met,

$$
\begin{aligned}
P(R-) &= P(\pi < \pi_1, \text{Reject } H_1) \\
&= P(\pi < \pi_1, x \le r_1, x+y \le r) + P(\pi < \pi_1, x > r_1 x+y \le r) \\
&= P(\pi < \pi_1 \mid x \le r_1) P(x+y \le r \mid x \le r_1) P(x \le r_1) + \\
&\quad P(\pi < \pi_1 \mid x+y > r, x > r_1) P(x+y > r \mid x > r_1) P(x > r_1) \\
&= \sum_{i=0}^{r_1} \sum_{j=0}^{r-i} P(\pi < \pi_1 \mid y = j, x = i) P(y = j \mid x = i) P(x = i) + \\
&\quad \sum_{i=r_1+1}^{n_1} \sum_{j=0}^{r-i} P(\pi < \pi_1 \mid y = j, x = i) P(y = j \mid x = i) P(x = i).
\end{aligned}
\tag{50}
$$

The remaining two joint probabilities are similarly found replacing $\pi < \pi_1$ with $\pi \ge \pi_1$. Once all four joint probabilities are constructed, the errors in section 2.5 (13) can be constructed. The chosen set of parameters is the set of parameters satisfying the error constraints and resulting in the optimal solution as with Simon's design.

4.5 Two stage Adaptive Heterogeneity trial algorithm

Combining the theory in the previous sections, an algorithm for the 2HA design is constructed which determines the trial outcome. For simplicity, the number of subgroups to be detected is two. The algorithm is as follows:

1. Compute Simon parameters given a null response rate, π_0, treatment effect, δ, and target errors, α, β resulting in parameters r_1, n_1, r, n
2. Conduct first stage of trial using Simon parameters resulting in the number of successes, observed value x

3. Determine if a classifier exists to partition the first stage sample into two groups. Test for heterogeneity using the Farrington test,

$$X_F^2 = \sum_{i=1}^{g} \frac{(x_i - n_{1i}\pi_1)^2}{n_{1i}\pi_1(1-\pi_1)} + \sum_{i=1}^{g} \frac{-(1-2\pi_1)}{n_{1i}\pi_1(1-\pi_1)}(x_i - n_{1i}\pi_1)$$

If a classifier does not exist, proceed to step 6.

4. Calculate a new maximum sample size, n^*, given the observed heterogeneity imbalance and observed absolute difference in response profile, $n^* = n\left(1+\hat{I}\left|\hat{\pi}_1\right|\right)$.

5. Calculate a new critical value, $r^*|(X = r_1, n^*)$, using the Beta-Binomial PP distribution given that the first stage criterion was met for n^* under the target errors.

6. Conduct the second stage under the appropriate sample size. If no heterogeneity exists (r,n); if heterogeneity exits, (r^*,n^*) resulting in observed value y

7. Compare $x+y$ to r under no heterogeneity and to r^* under heterogeneity. If $x+y > r$ or $x+y > r^*$, then the trial is estimated to be a success.

4.6 Estimation of response rate

It has been shown that under multiple stage designs, *e.g.* sequential tests, the maximum likelihood estimator (MLE) is generally biased (Li and Li 2000; Jung and Kim 2004). Since only extreme cases are observed in a 2-stage Phase II trial, *e.g.* crossing a futility boundary in stage 1 or crossing an efficacy boundary in stage 2, an optional sampling effect is introduced biasing the MLE (Whitehead 1986). The optional sampling effect causes the variance of the estimate to increase thus increasing the bias. The bias is

most pronounced in trials with only a futility boundary in stage 1, as compared to both a futility and efficacy boundaries, which is the case with both the Simon and 2HA designs.

Jung and Kim have shown that in a two stage trial the statistic (M,S), where M denotes the stage the trial terminates, $M=\{1,2\}$, and S denotes the total number of responses accumulated up to and including stage M, is a complete and sufficient statistic for π, the response rate of the trial.

Then, since $\hat{\pi}_1 = x_1 / n_1$ is an unbiased estimator of $\pi \mid M=1$ and the complete and sufficient statistic (M,S), the uniformly minimum variance unbiased estimator (UMVUE) of π can be constructed by the Rao-Blackwell theorem (Blackwell 1947),

$$\hat{\pi}_{UMVUE} = E\left[\hat{\pi}_1 \mid (m,s)\right] \tag{51}$$

which will not ,by definition, suffer from the bias of the MLE in a sequential test. The UMVUE for π, given a two stage trial is then

$$\hat{\pi}_{UMVUE} = \frac{\sum_{i=(r_1+1)\cup(s-n+n_1)}^{s\cap(n_1-1)} \binom{n_1-1}{i-1}\binom{n-n_1}{s-i}}{\sum_{i=(r_1+1)\cup(s-n+n_1)}^{s\cap(n_1-1)} \binom{n_1}{i}\binom{n-n_1}{s-i}} \tag{52}$$

where $(n_1, n-n_1)$ are the first and second stage sample sizes and r_1 is the first stage futility boundary (Jung and Kim 2004).

As an example, say that the following responses were observed at the end of the second stage of a trial $(m,s)=(2,7)$ with trial parameters $(r_1,n_1,r,n)=(3,13,12,33)$. The

support space, *e.g.* the summation space, is constructed for the UMVUE, $(r_1+1)\cup(s-n+n_1)=(3+1)\cup(7-30)=4$ and $s\cap(n_1-1)=7\cap(14-1)=7$ resulting in

$$\hat{\pi}_{UMVUE}\,|\,(m=2,s=7)=\frac{\sum_{i=4}^{7}\binom{13-1}{i-1}\binom{30}{7-i}}{\sum_{i=4}^{7}\binom{n_1}{i}\binom{30}{s-i}}=.322 \tag{53}$$

In contrast, the MLE is $\hat{\pi}_{MLE}=\frac{7}{33}=.212$ which is heavily downward biased.

Using these results and the composition of the source of the heterogeneity, estimable and identifiable subgroups, a general form of the UMVUE can be constructed. Given the sum of responses in stage one and stage two follows a Binomial distribution, $(X_i+Y_i)\,|\,\pi_i \sim Bin(n_i,\pi_i)$ for each subgroup, the UMVUE of the response rate π is the weighted sum of the UMVUEs for each individual subgroup,

$$\hat{\pi}_{UMVUE}=\sum_{g=1}^{2}\hat{w}_i\left(\frac{\sum_{i=(r_{1g}+1)\cup(s-n_g+n_{1g})}^{s_g\cap(n_{1g}-1)}\binom{n_{1g}-1}{i-1}\binom{n_g-n_{1g}}{s_g-i}}{\sum_{i=(r_{1g}+1)\cup(s_g-n_g+n_{1g})}^{s_g\cap(n_{1g}-1)}\binom{n_{1g}}{i}\binom{n_g-n_{1g}}{s_g-i}}\right) \tag{54}$$

where n_{1g} is the sample size for subgroup g in the first stage, n_g is the total sample size for subgroup g and $r_{1g}=\hat{w}_g r_1$ is the weighted first stage critical value for the *gth* subgroup. A weak assumption of independence is assumed and the assumption that the observed subgroups weights approximate the true population weights as $n\to\infty$.

CHAPTER 5

EFFECTS OF HETEROGENEITY ON PHASE II TRIALS: RESULTS

5.1 Effects of heterogeneity on Simon trial designs

For simplicity, the number of subgroups in the simulations was chosen to be $g = 2$. Given a combination of weight and response profile, the type I and type II errors were computed using $B_1 = 10{,}000$ Monte Carlo iterations. Due to the multiplicity of combinations of response/population 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, $(B_2 \mid g = 2) = 40{,}000$ Monte Carlo iterations were conducted; for example, $(B_2 \mid g = 2) = 4$ and $\pi_S = .25$ using a simple average can result in

$$(w_1, w_2, \pi_{01}, \pi_{02}) = \begin{Bmatrix} (.1,.9,.30,.20) & (.1,.9,.40,.10) \\ (.1,.9,.20,.30) & (.1,.9,.10,.40) \end{Bmatrix} \tag{55}$$

5.1.1 Simulation parameters

A sample of population proportion profiles were chosen to cover a heterogeneity imbalance of $\hat{I} = (0,.98)$ for the two subgroup simulations and was simulated as follows:

1. Under HRH or ARH, given the population profile for a 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 $\left(\bar{\pi}_S, \delta^*\right) = (.25, .15)$. The parameters for the uniform distribution are problem specific and are to subject to the constraints $0 \le \pi_{0i} \le 1$ for all i. The gth null response rate was generated to satisfy the averaging method. 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,

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

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, $\mathbf{a} = (a_1, a_2, \ldots, a_{g-1}) \sim tN(w_i, 1)$ and $a_g \sim 1 - \sum_{i=1}^{g-1} a_i$ where truncation occurs for the first $(g-1)$ subgroups at $(w_i \pm \partial a)$.

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 $\boldsymbol{\pi}_T = (\pi_{T1}, \pi_{T2}, \ldots, \pi_{Tg})$.

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 $y_{T_i} = 0$. 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, $\left(\sum_{b=1}^{B} I(x_0 + y_0 > r)/B\right) | \pi = \pi_0$ is the type I error of the test and $\left(\sum_{b=1}^{B} I(x_1 + y_1 \leq r)/B\right) | \pi = \pi_1$ 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 (20).

5.1.2 Results

The first simulation compared the effect of varying levels of heterogeneity imbalance using simple averages for a 2 subgroup trial (Barnes and Rai 2010). The data was simulated using R.9.2 (Team 2005). The target type I and type II errors are $(\alpha, \beta) = (.10, .20)$. Table 5 displays the errors with corresponding 95% quantile intervals for each class of heterogeneity. Under all three classes of heterogeneity and a heterogeneity imbalance of $I \leq .20$, the actual mean errors approximate the target errors. When the imbalance increases, $I > .20$ under HRH and GRH, the actual mean errors

exceed the target errors with increasing divergence as the imbalance increases. Under ARH, the type I mean error approximates the target 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 actual type II error range under GRH. Under an unknown response profile for 2 subgroups, the mean probability that trial is moderately to extremely oversized is 22%, $|\hat{\alpha}-\alpha| \geq .04$, and the mean probability that the trial is underpowered is 42%, $|\hat{\beta}-\beta| \geq .04$.

To further identify the effect of heterogeneity, tables 6 and 7 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 shift increasing farther 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 $I = .20$ under HRH, the majority of oversized trials are in the range of $(.10,.12]$, a small divergence from the nominal errors. When $I = .40$ and $I = .80$, the majority of oversized trials are in the ranges of $(.2,.3]$ and $(.4,1]$ respectively, substantial divergences from the target error and of high concern to the trial conduct; a similar pattern in seen with the actual 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 $(.10,.12]$. This would imply

that even though the trials are oversized, the effect of the heterogeneity is minimal on the type I error.

Table 4: Size and power for each class of heterogeneity by heterogeneity imbalance with corresponding 95% quantile and Monte Carlo intervals for a 2 subgroup example

Class	I	Actual Error I	95% QI	Actual Error II	95% QI
HRH	.02	.10	(.08, .11)	.20	(.18, .22)
	.20	.11	(.04, .20)	.21	(.11, .32)
	.40	.13	(.01, .34)	.22	(.06, .46)
	.60	.16	(0, .50)	.25	(.03, .61)
	.80	.20	(0, .65)	.28	(.02, .76)
	.98	.23	(0, .76)	.31	(.01, .86)
ARH	.02	.10	(.09, .11)	.20	(.18, .22)
	.20	.10	(.09, .11)	.20	(.14, .28)
	.40	.10	(.09, .11)	.21	(.09, .37)
	.60	.10	(.09, .11)	.22	(.06, .47)
	.80	.10	(.09, .11)	.23	(.04, .58)
	.98	.10	(.09, .11)	.24	(.03, .67)
GRH	.02	.10	(.08, .11)	.23	(.19, .30)
	.20	.11	(.04, .20)	.24	(.14, .46)
	.40	.13	(.01, .34)	.26	(.07, .66)
	.60	.16	(0, .50)	.30	(.03, .83)
	.80	.20	(0, .65)	.33	(.01, .94)
	.98	.23	(0, .76)	.36	(0, .98)

Table 5 Distribution of actual type I error for each class of heterogeneity and heterogeneity imbalance for a 2 subgroup example.

		Distribution of Actual Type I Error						
Class	I	$(\alpha_{MC} - .12]$	$(.12 - .14]$	$(.14 - .18]$	$(.18 - .2]$	$(.2 - .3]$	$(.3 - .4]$	$> .4$
HRH	.02	.31	.01	0	0	0	0	0
	.20	.09	.10	.17	.08	.04	0	0
	.40	.05	.05	.09	.04	.17	.09	0
	.60	.03	.03	.06	.03	.12	.12	.12
	.80	.02	.03	.04	.02	.09	.07	.22
	.98	.01	.03	.04	.01	.08	.06	.27
ARH	.02	.26	0	0	0	0	0	0
	.20	.26	0	0	0	0	0	0
	.40	.26	0	0	0	0	0	0
	.60	.26	0	0	0	0	0	0
	.80	.26	0	0	0	0	0	0
	.98	.26	0	0	0	0	0	0
GRH	.02	.31	.01	0	0	0	0	0
	.20	.09	.10	.17	.08	.03	0	0
	.40	.05	.05	.09	.04	.17	.09	0
	.60	.03	.03	.06	.03	.12	.10	.12
	.80	.02	.03	.04	.02	.09	.08	.22
	.98	.01	.03	.04	.01	.08	.06	.27

Table 6 Distribution of actual type II error for each class of heterogeneity and heterogeneity imbalance for a 2 subgroup example

		Distribution of Actual Type II Error						
Class	I	$(\beta_{MC} - .22]$	$(.22 - .24]$	$(.24 - .28]$	$(.28 - .3]$	$(.3 - .4]$	$(.4 - .5]$	$>.5$
HRH	.02	.36	.02	0	0	0	0	0
	.20	.08	.09	.16	.07	.08	0	0
	.40	.04	.05	.08	.04	.17	.12	.12
	.60	.03	.03	.06	.03	.11	.10	.25
	.80	.03	.03	.04	.02	.08	.08	.31
	.98	.01	.02	.03	.01	.07	.07	.35
ARH	.02	.37	.01	0	0	0	0	0
	.20	.18	.15	.12	.01	.01	0	0
	.40	.10	.09	.14	.05	.10	.01	.10
	.60	.06	.07	.12	.05	.13	.05	.06
	.80	.04	.06	.08	.05	.13	.08	.13
	.98	.03	.05	.07	.02	.13	.09	.18
GRH	.02	.29	.21	.19	.06	.03	0	0
	.20	.06	.07	.07	.05	.14	.10	.10
	.40	.04	.02	.07	.01	.11	.08	.24
	.60	.03	.02	.03	.03	.07	.07	.32
	.80	.02	.03	.01	.02	.08	.05	.35
	.98	.01	.02	.01	.01	.07	.03	.38

The second scenario is the weighted average, table 7. Under HRH and ARH, the actual mean 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 $(.10,.12]$ and $(.20,.22]$ respectively, a divergence between target and actual errors implying that some trials do not meet the error targets. Under weighted averages, the effect of heterogeneity is minimal, but not absent, on the operating characteristics of the Simon trial.

Table 7 Errors for each class of heterogeneity by heterogeneity imbalance with corresponding 95% quantile for a 2 subgroup example using weighted averaging.

Class	I	Actual Error I	95% QI	Actual Error II	95% QI
HRH	.02	.10	(.09, .11)	.20	(.18, .22)
	.20	.10	(.09, .11)	.20	(.18, .22)
	.40	.10	(.08, .12)	.20	(.18, .22)
	.60	.10	(.08, .12)	.20	(.18, .22)
	.80	.10	(.08, .12)	.20	(.18, .22)
	.98	.10	(.09, .12)	.20	(.18, .22)
ARH	.02	.10	(.09, .11)	.20	(.18, .22)
	.20	.10	(.09, .11)	.20	(.18, .22)
	.40	.10	(.09, .11)	.20	(.18, .22)
	.60	.10	(.09, .11)	.20	(.18, .22)
	.80	.10	(.09,.11)	.20	(.18, .22)
	.98	.10	(.09. 11)	.20	(.18, .22)
GRH	.02	.10	(.09, .12)	.20	(.18, .24)
	.20	.10	(.09, .12)	.20	(.18, .24)
	.40	.10	(.08, .12)	.21	(.18, .24)
	.60	.10	(.08, .12)	.22	(.18, .25)
	.80	.10	(.08, .12)	.22	(.18, .25)
	.98	.10	(.09, .12)	.23	(.18, .24)

To allow for the uncertainty in either the true proportions or the accrual, two levels of error were introduced during patient accrual, $\partial a = .05$. The accrual heterogeneity imbalance was allowed to vary between 0 and 5% of the population heterogeneity imbalance. The accrual difference can be attributable to accrual divergence or error in proportion estimation. Table 8 shows the results for $g = 2$ subgroups with an accrual divergence parameter of 5%. The actual mean errors approximated the target errors in almost every case with the exception being under GRH actual type II errors. The reason for this divergence is unknown at this time. The distributions of the errors are more dispersed than the weighted average method due to the variation in accrual which

leads to more specific combinations of a weight and response profile being underpowered or oversized.. The strength the errors is increased when comparing the error estimate distributions between weighted averages and weighted averages with accrual divergence.

Table 8 Simon Optimal design with s=2 subgroups population using weighted average with accrual differences, $\partial a = .05$.

	∂a	I	Actual Error I	95% CI	Actual Error II	95% CI
HRH	.10	.02	.10	(.07, .14)	.20	(.16, .24)
		.20	.10	(.07, .13)	.20	(.16, .24)
		.40	.10	(.08, .14)	.20	(.16, .24)
		.60	.10	(.07, .12)	.20	(.16, .24)
		.80	.10	(.08, .13)	.20	(.16, .24)
ARH	.10	.02	.10	(.09, .11)	.20	(.17, .23)
		.20	.10	(.09, .11)	.20	(.17, .23)
		.40	.10	(.09, .11)	.20	(.17, .23)
		.60	.10	(.09, .11)	.20	(.17, .23)
		.80	.10	(.09, .11)	.20	(.17, .23)
GRH	.10	.02	.10	(.07, .14)	.20	(.16, .26)
		.20	.10	(.07, .13)	.21	(.17, .26)
		.40	.10	(.07, .12)	.22	(.17, .38)
		.60	.10	(.07, .12)	.22	(.17, .29)
		.80	.10	(.08, .13)	.23	(.17, .34)

5.2 Effects of heterogeneity on Adaptive trial design

5.2.1 Simulation parameters

The data for the 2HA simulation was simulated in a similar manner as with the Simon simulation in section 5.1.1. Only two groups under GRH were simulated with the weighted average constraint. Under latent heterogeneity, GRH is the most appropriate class of heterogeneity. It is also assumed that if a response rate was hypothesized, that the rate was hypothesized from data that follows the population as a whole. For example, if a trial response rate of .30 is hypothesized, then the population in general follows a

mean response of .30. No subgroups are known prior to trial conduct, though this does not occlude the existence of said subgroup, and a Treatment x Marker interaction is not known. For each weight scheme of the weighting profiles, 10,000 simulations were conducted. The actual errors, the percentage not meeting the target errors, are reported.

5.2.2 Simulation Algorithm

The data was simulated using a Linux cluster by parallelizing the simulation using R.9.2 (Team 2005). The data and trial conduct is simulated as follows

1. Given a null response rate and alternative response rate with a specified weighting scheme, $\pi_0, \pi_1, (w_1, w_2)$, and target errors, (α, β) construct the Simon trial parameters and a null and alternative response profile by weighted averages satisfying the odds ratio criterion.

2. Given a weight profile and an alternative response profile, simulate multinomial random variables n_{11}, n_{21} with fixed sample size n_1 and cell probabilities $\boldsymbol{\pi}_T = (\pi_{T1}, \pi_{T2})$.

3. For values of $(N_{11}, N_{21}) = (n_{11}, n_{21})$, simulate binomial random variables x_i with sample size n_{i1} and response rate π_{1i}. Then $x = \sum_{i=1}^{2} x_{1i}$ is compared to the critical value r_1 derived from the Simon trial design. If $x \leq r_1$, then the trial is stopped for futility.

4. If $x > r_1$ then determine if heterogeneity is present using the Farrington X^2 statistic and a liberal p-value, $p \geq .30$.

5. Compute the heterogeneity imbalance and absolute magnitude in difference in responses. Then construct a new sample size, n^* using (41).

6. Repeat steps 2-5 for B=50,000 iterations.

7. Given the unique possible new sample sizes in the 50,000 iterations, construct new critical values, r^*, given r_1 out of n_1 responses in the first stage and for each of the new sample sizes using the predictive posterior Beta-Binomial distribution.

8. If $p \geq .30$, then $n_2 = n^* - n_1$ and $r = r^*$; otherwise, $n_2 = n - n_1$ and $r = r$.

9. Repeat steps (2-3) for the second stage, n_2 to determine y; otherwise $y = 0$. The power of the test, $1-\beta$ is constructed such that $1-\beta = \dfrac{\sum_{i=1}^{b} I(x+y \leq r)}{b}$.

10. Repeat steps 2-6 for the null response rate. Given the new sample size, the critical values determined using the alternative response are used.

11. If $p \geq .30$, then $n_2 = n^* - n_1$ and $r = r^*$; otherwise, $n_2 = n - n_1$ and $r = r$.

12. The size of the test or α is constructed such that $\alpha = \dfrac{\sum_{i=1}^{b} I(x+y > r)}{b}$.

13. Repeat steps 1-12 for 10,000 iterations of different response profiles that satisfy the weighted average given the weight profile. The trial errors given a weight profile are then constructed using (20) where $s = 10,000$.

14. Repeat all steps for each weight profile,

$$\mathbf{w} = \{(.1,.9),(.2,.8),(.3,.7),(.4,.6),(.5,.5)\}.$$

This algorithm results in the estimates of the trial operating characteristics given a specific weight profile and any possible response profile satisfying the weighted average constraint.

5.2.3 Results

Under latent heterogeneity, the appropriate form of heterogeneity is generalized response heterogeneity. Under GRH, no information is known *a priori* on the source of the heterogeneity. A two subgroup trial was simulated under GRH. Table 9 shows the results given 10,000 simulations of weight*response profiles satisfying the weighted average response rate constraint for the following parameters

$\pi_0 = .30; \pi_1 = .45; \alpha = .10; \beta = .20$.

The errors reported are the percentage of times the individual weight*response profile errors crossed the target error boundaries in the inappropriate direction, $\hat{\alpha}_b > \alpha$ and $\hat{\beta}_b > \beta$ where the weight*response type I error for the *bth* simulation is $\hat{\alpha}_b$. α_{2HA} and α_S denote the type I error for the Adaptive and Simon designs. As with the

Simon simulations, the expected value of the size and power are very close to the targets, but it is the range that is more clinically important or as a proxy, the percentage above the target error bounds.

Table 9 Simulated error estimates for various weight profiles with target errors of $(\alpha,\beta)=(.1,.2)$ and $(\pi_0,\pi_1)=(.30,.45)$.

w_1	w_2	I	α_{2HA}	β_{2HA}	α_S	β_S
0.1	0.9	.8	0	0.18	0.22	0.40
0.2	0.8	.6	0	0.17	0.15	0.40
0.3	0.7	.4	0	0.14	0.08	0.30
0.4	0.6	.2	0	0.05	0.04	0.28
0.5	0.5	0	0	0	0	0

The adaptive design maintains the target type I error under all levels of heterogeneity, *e.g.* weight profiles. The error increases as the heterogeneity imbalance increases, but is below the target type II error. This is not the case with the Simon design. From the simulations in section 5.1, the divergence from the target is marginal, within, for example, $\beta \pm .05$, but under the conduct of these type of trials, any divergence from the target is the wrong direction is clinically substantial.

The Simon design does meet the target errors with an equal weight profile. As the heterogeneity imbalance increases, the percentage of trials under a specific weight profile that exceed the target error bounds increases. In the case of extreme weighting of the subgroups, $I=.8$, 40% of the trials will not have a minimum of 80% power with 22% of the trials oversized which can result in a successful Phase II trials but be the cause of a failed Phase III trial.

A second consideration of Phase II trials is the expected sample size of the trial. Given a weight profile, summary statistics can be computed for the expected sample size under heterogeneity for the 2HA design which relate the mean, standard deviation and confidence intervals for the total sample size and expected sample size *EN* following (5) where the mean or confidence interval is substituted for n,

$$EN(H_0) = n_1 + (1 - PET)(\bar{n} - n_1) \tag{56}$$

Given $(\pi_0, \pi_1) = (.30, .45)$ and $(\alpha, \beta) = (.10, .20)$, the Simon trial design results in the following parameters and operating characteristics,

$(n_1, n, PET, EN) = (20, 55, .6070, 41.24)$ under the optimal design. Under the Adaptive design, the *PET* remains the same. Given 10,000 simulations of weight*response profiles for each weight profile, the sample size summary statistics are listed in table 11.

Table 10 Sample sizes under the 2HA design

I	$\bar{n}[sd]$	Range	$EN(H_0 \mid I)$	95% CI for *EN*
.8	70 [7]	(55, 105)	50.37	(41.62, 59.11)
.6	63 [6]	(55, 105)	45.98	(38.78, 53.18)
.4	59 [4]	(55,102)	43.94	(38.82,49.05)
.2	58 [3]	(55,98)	42.83	(39.50,46.18)
0	57 [3]	(55,100)	42.59	(39.49,45.69)

Under the extreme case of heterogeneity imbalance, $I = .8$, the 2HA design results in an average increase of 27% of the Simon sample size with a corresponding 49% increase in the expected sample size to control the trial errors at the target errors. As the heterogeneity imbalance decreases, the expected sample size decreases as does the mean sample size. In the case of no heterogeneity imbalance, the 2HA design only increases

the sample size by 2 with an increase in the expected sample size of ~1. This provides a justification for allowing the type I error to be inflated for the heterogeneity test. If heterogeneity is detected where none is truly present, the trial will only lead to a minimal increase in sample size.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Summary and conclusions

The primary purpose of a Phase II trial is to estimate if a treatment has a clinically meaningful effect on a group of patients for further more targeted testing. Generally, *ex vivo* evidence exists which provides evidence that the treatment does have efficacy on either tissue samples or cell lines. The purpose of the Phase II trial is to demonstrate this efficacy on a small subset of the affected patient population. If the Phase II trial is successful, then either more targeted Phase II trials are conducted or the treatment moves on to Phase III testing. Phase III testing provides the definitive answer as to the efficacy of a treatment.

Phase II trials have some unique properties not seen in the earlier or later phases of clinical trials. The trials are a single arm with no control arm. Though there is a move towards Phase II trials with a randomized design with control and treatment arms, the majority of Phase II trials are still single arm. The sample size in these trials is small, from only 10 patients up to 120 patients. As such, Phase II trials only estimate a response, and the larger Phase IIIs are used to define the response. As the name suggests, the type of trial is early in the drug development so historical information on the efficacy of the treatment or the patient population may be lacking.

All trials, Phase I-IV, rely on some homogeneity of the population to draw inferences. In a Phase III trial or a randomized Phase II trial, the control and treatment arms are expected to be similar, e.g. homogeneous in response. This is not always possible in a Phase II trial. With only one arm, the homogeneity does not rest between two samples, e.g. two trial arms, but rests on the homogeneity of the response in a single population. The single population being tested, through the response rate, must be homogeneous. Phase IIIs do not need this assumption; only that homogeneity exists between samples. For example, a Phase III or randomized Phase II can have multiple subgroups in the trial, each with a unique response rate. As long as the distribution of the subgroups is the same across the two arms, the homogeneity assumption is met. In a Phase II trial, no comparison can be made, so subgroups cannot exist. All patients must come from the same population with the same response. This assumption can be lacking in actual trial practice and poses a substantial hurdle to accurate inferences from said trials.

The most common Phase II trial design is the Simon Phase II designs as described in section 2.1. These designs rely on a homogeneity assumption in the response in order for the trial data to fit the model distribution of the trial parameters, the Binomial distribution. In practice, subgroups may exist causing a divergence in the trial data from the Binomial distribution. A large number of Phase II trials fail. It is paramount to determine why these trials have such a high failure rate. In examining the problem, multiple failure modes can be suggested, 1) the trial fails due to inaccurate estimation of the true response rate, 2) the trial fails due to difference in the patient

population cells and ex vivo cell lines, 3) the trial fails due to the use of a homogeneous trial design when heterogeneity exists. This paper has focused on third failure mode.

Response heterogeneity can be constructed from a set of subgroups in the population with subgroups having either unique response rates or unique etiologies of their response rates. This paper provided a model for heterogeneity to quantify the effects of heterogeneity on trial designs. The model was applied to the popular Simon design and the Simon design was shown to have inflated errors beyond the target errors rates across all three classes of heterogeneity. This is a standard practice in clinical trials, subgroups are known to exist but for simplicity and to minimize patient resources, a single trial is conducted using an averaged response with the Simon designs. Intuitively, the inflation of error would be expected when a simple average is used to combine a set of subgroup response rates into a homogeneous response rate. Under simple averaging, there is no weighting to the average, but in fact, the population may have a specific weighting profile. Hence, the simple average response rate may substantially diverge from the true weighted population response rate.

To correct for this, weighted averages were also applied to the Simon designs. In theory, the weighted average would correspond to the weighted average of the population and correspond to a homogeneous response rate mathematically. This would hold true except for one issue. The response is a weighted sum of binomials, but the weighted sum of binomials is not binomial. Hence, the use of a binomial model distribution may not be appropriate. The simulation results confirm this fact. While the expected value of the errors across simulations, given a weight profile, usually did

maintain the target errors, the range of actual errors crossed the target error bounds. Specific weight*response profiles would exceed the target error bounds. In trial conduct, any divergence is substantial since the trial is designed to never exceed the target error bounds.

From a clinical interpretation using the expected value under heterogeneity is not valid as a measure of the trials errors. Trial designs must maintain a target error for every specific weight*response profile or within an allowable error level. The Simon designs do not have this operating characteristic under heterogeneity. Hence, even under weighted averages, the Simon design will not maintain the target errors. This provides solid evidence that the Simon trial design may be *a* source of the high failure rates of Phase II trials. Under homogeneity, the Simon design is the most efficient design, but the design was not constructed to handle heterogeneity.

In order to develop a new design to handle heterogeneity, multiple current methods were evaluated. All of the methods that have been developed in the past few years suffer from a limiting factor, the composition of the subgroups must be known in advance. It can be argued that if subgroups are known to exist, the most conservative path is to conduct separate trials for each subgroup. This is a second common practice, in opposition to averaging, conduct multiple trials. This can lead to a substantial increase in trial resources which may be a motivating factor for using an averaged response.

In practice, subgroups are not generally known at such an early stage of a treatment's efficacy exploration. Failed Phase IIs provide solid evidence for future Phase II's with the same or similar treatment. As more Phase II's are conducted a better image

of the treatments efficacy is refined. This approach uses the accumulation of trials to accumulate and refine knowledge on the existence and composition of subgroups, but is costly when evaluating patient and financial resources. Ethically, if a more optimal solution to the allocation of patient resources exists, then this optimal method should be utilized. When no information is available at the beginning of trial conduct, the type of heterogeneity is known as latent heterogeneity and corresponds to the generalized response heterogeneity class of the heterogeneity model. This would correlate to the first conduct of a Phase II trial. We suggest a more optimal method than conducting multiple Phase II trials to refine efficacy estimates.

We have developed a new Phase II design that can handle latent response heterogeneity with a modest increase in the sample size. The new design works by incorporating the trial structure, a two stage process, and the data structure, subgroups, into the model distribution of the trial parameters, the Beta-Binomial posterior predictive distribution. The predictive posterior form is chosen to account for the two stage process of the trial. The trial starts as a standard Simon design with Simon design parameters. After the first stage has concluded, the trial data, a single grouped sample, is tested for the existence of subgroups using an unsupervised classification algorithm. Then a heterogeneity test is computed. If heterogeneity is detected, the trial's overall sample size is increased by an empirically determined variance inflation factor.

The variation inflation factor is the empirical estimate of the product of the magnitude of the response profile and the magnitude of the weight profile from the first stage conduct. This method to increase the sample size seems very intuitive. The

response heterogeneity can be decomposed into two sources which differ in heterogeneous responses as compared to a homogeneous response, weight and response profiles versus no weights and a single fixed response rate. Multiple other methods were tried, but no method was able to consistently maintain the errors across all possible weight profiles. The trial uses the posterior predictive Beta-Binomial distribution to construct a new critical value by an exact method to determining the trial errors.

This new design, the 2-stage heterogeneity adaptive design, is shown to never exceed the target trial errors. Even with an extreme heterogeneity imbalance, the target trial errors are maintained. Under small or no heterogeneity imbalance or a small difference in the response magnitude, there is only a marginal increase in the total or expected sample size. This fact gives credence to the earlier use of a very liberal heterogeneity test. With only a single grouped sample, no single heterogeneity test is reliably going to result in robust inferences or always control the type I error, the probability of determining heterogeneity when one exists. The chosen heterogeneity test, the Farrington test, does maintain an acceptable level of power, >80% which is of primary concern. If one does not detect heterogeneity when it truly exists, then the Simon design is not appropriate. The scalable sample size based on the variance inflation factor results in only minimal increases in sample size when heterogeneity is detected, but truly not existent.

A limiting factor of this design is the determination of subgroups. A full exposition of how to determine subgroups from an unsupervised approach is beyond the scope of this paper, but the limitations imposed by this issue are understood. A second

limiting factor is the stopping of the trial to determine the existence of subgroups. Many designs have been developed which maintain the operating characteristics of the Simon design but allow for continuous patient accrual with no between stage stopping. Many clinicians may feel that waiting a few weeks between stages is too long, but in the end, this extra time may result in a saved trial. As such, the determination of the classifier and the time is takes to determine the classifier are limitations to this design.

In conclusion, the 2-stage heterogeneity adaptive design maintains the target errors of a binary 2-stage single arm trial. The trial design preserves the desirable operating characteristic of the Simon design, moderate probability of early termination, without a substantial increase in trial resources. The increase in trial resources is determined by the first stage results. If subgroups are detected and the imbalance in these subgroups either in weights or response rates is high, then the increase in the overall sample size compared to the Simon trial will be substantial. In no way, it is claimed to be the minimum necessary increase in sample size, but through simulation was shown to always be adequate. This method works under the full range of possible heterogeneity which will at times result in larger than necessary sample sizes.

6.2 Direction for future work

This design presents multiple areas of limitation that need to be addressed. The primary limitation is the detection of subgroups. More work is needed to indentifying an unsupervised classification algorithm that can work under the small sample sizes in the first stage of a Phase II clinical trial. An unsupervised algorithm presents a very practical case since the source of heterogeneity in many diseases is determined to be genomic. If a

classifier is used that does not have perfect classification, then an additional source of variation is introduced into the problem. In this case the VIF becomes the product of both the weight and response profiles and the classification accuracy. Less accurate classifiers will result in an increase in the sample size. Secondly, a better test for heterogeneity that preserves the type I error would also be an optimal improvement.

The detection of only two groups is adequate for trials where the sample size is relatively small, but in larger Phase II trials, the detection of more than two groups should be possible. Understanding what the limitations are for a higher detection number and robust tests are necessary. The use of the Beta-Binomial posterior predictive distribution provides a necessary correction factor for heterogeneity in the second stage of the trial. Work needs to be conducted to determine if it would be more optimal to not start as a Simon Design, but conduct the entire trial using the posterior predictive distribution.

A final improvement would be to include local hypothesis tests. The most desirable attribute of the Bayesian ANOVA and hierarchical methods is the sharing of information across subgroups and allowing for subgroup specific stopping or acceptance.

REFERENCES

Ayanlowo, A. O. and D. T. Redden (2008). "A two stage conditional power adaptive design adjusting for treatment by covariate interaction." Contemp Clin Trials **29**(3): 428-438.

Barnes, C. and S. Rai (2010). "Modeling Heterogeneity in Phase II Clinical Trials." American Journal of Biostatistics **1**(1): 9-16.

Barnes, C. and S. N. Rai (2010). "The effects of heterogeneity on Simon Phase II clinical trial design operating characteristics." Open Access Journal of Clinical Trials **Accepted**.

Barnes, C. and S. N. Rai (2010). "An exact method for link parameter estimation in error benchmarking." Statistical methods in medical research **Submitted**.

Behrendt, C. E. and E. A. Gehan (2009). "Treatment–subgroup interaction: An example from a published, phase II clinical trial." Contemp Clin Trials **30**: 279-281.

Blackwell, D. (1947). "Condtional expectation and unbiased seqeuntial estimation." Annals of Mathematical Statistics **18**(1): 105-110.

Chow, S., J. Shao, et al. (2007). Sample Size Calculations in Clinical Research. Boca Raton, Chapman & Hall/ CRC group.

Collet, D. (2003). Modeling Binary Data. New York, CPR.

Datta, S. (2006). "Evaluation of clustering algorithms for gene expression data." BMC Bioinformatics **7 Suppl 4**: S17.

Dragalin, V. and V. Fedorov (2006). "Design of multi-centre trials with binary response." Stat Med **25**(16): 2701-2719.

Emerson, S. S., J. M. Kittelson, et al. (2007). "Bayesian evaluation of group sequential clinical trial designs." Stat Med **26**(7): 1431-1449.

Emerson, S. S., J. M. Kittelson, et al. (2007). "Frequentist evaluation of group sequential clinical trial designs." Stat Med **26**(28): 5047-5080.

Farrington, C. P. (1996). "On assessing goodness of t of generalized linear models to sparse data." Journal of the Royal Statistical Society, Series B **58**: 349-360.

Gadbury, G. L. and H. K. Iyer (2000). "Unit-treatment interaction and its practical consequences." Biometrics **56**(3): 882-885.

Gehan, E. A. (1961). "The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent." J Chronic Dis **13**: 346-353.

Gilks, W., R. Richardson, et al. (1996). Markov Chain Monte Carlo in Practice. London, Chapman & Hall.

Green, S. B. (1982). "Patient heterogeneity and the need for randomized clinical trials." Controlled Clinical Trials **3**(3): 189-198.

Hendriks, J. C., S. Teerenstra, et al. (2005). "Sample size calculations for a split-cluster, beta-binomial design in the assessment of toxicity." Stat Med **24**(24): 3757-3772.

Hunt, D. L. and S. N. Rai (2003). "A Threshold Dose-Response Model with Random Effects in Teratological Experiments." Communications in Statistics **32**: 1439-1457.

Hunt, D. L. and S. N. Rai (2005). "A new threshold dose-response model including random effects for data from developmental toxicity studies." J Appl Toxicol **25**(5): 435-439.

Jung, S. H. and K. M. Kim (2004). "On the estimation of the binomial probability in multistage clinical trials." Stat Med **23**(6): 881-896.

Kuss, O. (2002). "Global goodness-of-fit tests in logistic regression with sparse data." Stat Med **21**(24): 3789-3801.

Lee, J. J. and L. Feng (2005). "Randomized phase II designs in cancer clinical trials: current status and future directions." J Clin Oncol **23**(19): 4450-4457.

Lee, P. M. (2009). Bayesian Statistics: An Introduction. New Jersey, Wiley.

Lee, S. and M. Zelen (2000). "Clinical trials and sample size considerations: another perspective." Statistical Science **15**: 95-110.

Li, Z. and Y. Li (2000). "A homogeneity test in overviews with group sequentially monitored clinical trials." Biometrics **56**(1): 134-138.

London, W. B. and M. N. Chang (2005). "One- and two-stage designs for stratified phase II clinical trials." Stat Med **24**(17): 2597-2611.

Makuch, R., M. Stephens, et al. (1989). "Generalized Binomial Models to Examine the Historical Control Assumption in Active Control Equivalence Studies." The Statistician **38**(1): 61-70.

Romesburg, H. C. (2004). Cluster Analysis for Researchers, Kreiger Pub. Co.

Rosner, G. L., W. Stadler, et al. (2002). "Randomized discontinuation design: application to cytostatic antineoplastic agents." J Clin Oncol **20**(22): 4478-4484.

Russek-Cohen, E. and R. M. Simon (1997). "Evaluating treatments when a gender by treatment interaction may exist." Stat Med **16**(4): 455-464.

Schultz, J. R., F. R. Nichol, et al. (1973). "Multiple-stage procedures for drug screening." Biometrics **29**(2): 293-300.

Simon, R. (1989). "Optimal two-stage designs for phase II clinical trials." Control Clin Trials **10**(1): 1-10.

Stadler, W. M. (2007). "The randomized discontinuation trial: a phase II design to assess growth-inhibitory agents." Mol Cancer Ther **6**(4): 1180-1185.

Team, R. d. C. (2005). R: A language and environment for statistical computing. R. F. f. S. Computing. Vienna, Austria.

Thall, P. F. and J. K. Wathen (2008). "Bayesian designs to account for patient heterogeneity in phase II clinical trials." Curr Opin Oncol **20**(4): 407-411.

Thall, P. F., J. K. Wathen, et al. (2003). "Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes." Stat Med **22**(5): 763-780.

Tuma, R. S. (2008). "Examining heterogeneity in phase II trial designs may improve success in phase III." J Natl Cancer Inst **100**(3): 164-166.

Wathen, J. K., P. F. Thall, et al. (2008). "Accounting for patient heterogeneity in phase II clinical trials." Stat Med **27**(15): 2802-2815.

Whitehead, J. (1986). "On the bias of maximum likelihood estimation following a sequential test." Biometrika **73**(3): 573-581.

Williams, D. A. (1982). "Extra-Binomial Variation in Logistic Linear Models." Journal of the Royal Statistical Society. Series C (Applied Statistics) **31**(2): 144-148.

Yamamoto, E. and T. Yanagimoto (1994). "Statistical methods for the beta-binomial model in teratology." Environ Health Perspect **102 Suppl 1**: 25-31.

Young-Xu, Y. and A. Chan (2008). "Pooling overdispersed binomial data to estimate event rate." BMC Med Res Methodol **8**(1): 58.

CURRICULUM VITAE

Christopher N. Barnes, Ph.D.

Contact Information:

17 Brosnan Street #3
San Francisco, CA 94103
(502) 554-4940 (Mobile)
ChrisBarnesSF@gmail.com

EDUCATION

8/06 – 5/10	Ph.D., Biostatistics, Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky.
8/07	M.S., Biostatistics, Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky.
5/04	B.S., Mathematics, University of Louisville, Louisville, Kentucky.

PROFESSIONAL WORK EXPERIENCE

8/08-present

Fellow
Biostatistics Shared Facility, James Graham Brown Cancer Center, Louisville, Kentucky.

Provide statistical support to Cancer Center researchers and faculty, develop Phase II clinical trial protocols and SAPs, design and analysis of genomic biomarker studies, and develop theoretical framework for novel experiments/data analysis.

Mentor: Dr. Shesh N. Rai, Director, Biostatistics Shared Facility, James Graham Brown Cancer Center; Professor, Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville

8/06-present **Statistical Consultant**
Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky.

Develop statistical analysis plans, provide data analysis and interpretation, and write summary reports for interdisciplinary projects in the Health Sciences; Expertise in Bioinformatics applications.

Mentor: Dr. Susmita Datta, Bioinformatics Core Chair, Professor, Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville

1/05 -7/07 **Statistical Consultant**
Behavioral Oncology Program, James Graham Brown Cancer Center, Louisville, Kentucky.

Provide Statistical consulting for Behavioral Oncology Program research projects; develop statistical design and analysis plans/sample size calculations for research proposals.

Mentor: Dr. Jamie Studts, Director, Behavioral Oncology Program, James Graham Brown Cancer Center; Associate Professor, Dept of Medicine.

5/04 – 8/04 **Research Assistant**
Department of Mathematics, University of Louisville.

Perform beta testing of the statistical interface between SAS and ArcGIS; completed and presented a research project exploring the spatial relationships of Prostate Cancer.

Mentors: Dr. Patricia Cerrito, Professor, Dept. of Mathematics; Dr. Carol Hanchette, Associate Professor, Dept. of Geosciences.

COMPUTER LANGUAGES/PROGRAMS

SAS	R/S-PLUS	StatExact
WinBugs	Ingenuity	Microsoft Office

SPECIAL TRAINING

2008 NIEHS SNPs Workshop, Louisville, KY, Jan 10-11.

HONORS & AWARDS

2010	Graduate Dean Citation
2009	Travel Award JSM
2007	3rd Place Award, Poster Presentation, Research! Louisville
2006	2nd Place Award, Poster Presentation, Research! Louisville
2006	Golden Key Honor Society

PUBLICATIONS

Barnes, C.N., Rai, S.N. (2010). Modeling heterogeneity in Phase II clinical trials. *American Journal of Biostatistics 1(1):9-16.*

Barnes, C.N., Rai, S.N. (2010). The effects of heterogeneity on Simon Phase II clinical trial design operating characteristics. *Open Access Journal of Clinical Trials.* (Accepted)

Barnes, C.N., Rai, S.N. (2010). An exact method for link parameter estimation in error benchmarking. *Statistical Methods in Medical Research.* (Under Review)

Kanaan, Z., Eichenberger, M.R., **Barnes, C.N.**, Rai, S.N., Hicks, N., Mulhall, A., Gholson, R., Mottern, E., Rai, S.N., Galandiuk, E. (2010). MicroRNA Regulation of p53 and Wnt-Beta catenin pathways in inflammatory bowel disease-associated colorectal cancer. *Clinical Cancer Research* (Under Review)

Kanaan, Z., **Barnes, C.N.**, Rai, S.N., Winston, V.A., Mulhall, A., Gholson, R., Galandiuk, S. (2010). Extraintestinal manifestations of inflammatory bowel disease and the influence of smoking. *Diseases of the Colon & Rectum.* (Under Review).

Wintergerst, K.A., Hinkle, K., **Barnes, C.N.**, Omoruyi, A., Foster, M.B. (2010). The impact of health insurance coverage on pediatric diabetes management. *Pediatric Diabetes.* (Under Review)

Studts, J. L., Ghate, S. R., Gill, J. L., Studts, C. R., **Barnes, C. N.**, LaJoie, A. J., Andrykowski, M. A., LaRocca, R. V. (2006). Validity of self-reported smoking status among participants in a lung cancer screening trial. *Cancer Epidemiology, Biomarkers and Prevention 15(10): 1825 - 1828.*

MANUSCRIPTS IN PREPARATION

Barnes, C.N., Rai, S.N. (2010). A Phase II Two stage clinical trial design to handle latent heterogeneity.

Barnes, C.N., Rai, S.N. (2010). Hierarchical filters in high-throughput screening to improve the true positive call rate.

Barnes, C.N., Kanaan, Z., Galandiuk, S. Rai, S.N. (2010). Statistical Issues in qRT-PCR miRNA analysis.

RESEARCH PRESENTATIONS

Barnes, C.N. (2009). *Hierarchical filters in high-throughput screening to improve the true positive call rate.* Seminar given at Abbot Labs, Biomarker Discovery Group, Chicago, IL.

Barnes, C.N. (2009). *Evaluation of Bayesian designs under frequentist validity: Phase II 2-stage clinical trials.* Seminar given at Novartis Oncology, Florham Park, NJ.

Barnes, C.N. (2009). *Modeling heterogeneity and its effect on the design parameters in Simon phase II clinical trials.* Seminar given at the Joint Statistical Meetings, Washington D.C.

Barnes, C.N. (2009). *Understanding heterogeneity in Phase II trials.* Seminar given at University of Western Illinois Department of Mathematics Colloquia.

Barnes, C.N. (2009). *Impact of heterogeneity on the operating characteristics of Phase II clinical trials.* Seminar given at University of Louisville Department of Bioinformatics and Biostatistics Seminar Series.

Barnes, C.N., Rai, S.N. (2010). *Hierarchical filters in high-throughput screening to improve the true positive call rate.* Poster presented at the James Graham Brown Cancer Center Annual retreat 2009.

Barnes, C.N., Wintergerst, K., Hertweck, S.P., Todd, R., Foster, M., Dietrich, J.E. (2007). *A clinical decision rule for differentiating between polycystic ovarian syndrome and late onset congenital adrenal hyperplasia in adolescent females.* Poster presented at Research! Louisville 2007.

Barnes, C.N., Studts, C.R., Studts, J.L. (2006). *Participant adherence in a RCT of lung cancer screening: baseline to year 1.* Poster presented at Research! Louisville 2006.
Poster presented at the James Graham Brown Cancer Center Annual retreat 2006.

Studts, J. L., Ghate, S. R., Gill, J. L., Studts, C. R., **Barnes, C. N.** (2006). *Validity of self-reported smoking status among participants in a lung cancer screening trial.* Poster presented at Kentucky Lung Cancer Conference.

EDITORIAL POSITIONS

2009-present Reviewer- *Contemporary Clinical Trials*

TEACHING EXPERIENCE

Graduate Teaching Assistant, Advanced Clinical Trials, Spring 2009
Graduate Teaching Assistant, Advanced Survival Analysis, Spring 2010

MEMBERSHIP IN STUDENT AND PROFESSIONAL ORGANIZATIONS

American Statistical Association
Student Government Association, SPHIS:
President, 07-08
Faculty Forum Representative, 06-07
Program Chair, Biostatistics, 06-07
Technology Committee, 06-07

FUNDED GRANTS

Agency: Clinical & Translational Science Pilot Grant Program Advanced Award, University of Louisville
Title: *A novel method for the diagnosis and prognosis of inflammatory bowel disease associated cancer*
Amount: $82,797
PI: Susan Galandiuk M.D.
Role: Biostatistician

Agency: Melanoma Research Foundation
Title: *Develop a Prognostic Scoring System in Node-Negative Patients*
Amount: $180,000
PI: Kelly M. McMasters, M.D.
Role: Biostatistician

Agency: University of Louisville Collaborative Matching Grant
Title: *Develop Gene Expression Signatures to Predict Prognosis in Melanoma Patients with Tumor-Negative Sentinel Lymph Nodes*
Amount: $40,000
PI: Hongying Hao, Ph.D.
Role: Biostatistician

PENDING GRANTS

Agency: University of Louisville Innovative Translational Research Award
Title: *Develop a prognostic system incorporating gene signatures in melanoma patients with positive sentinel lymph nodes*
Amount: $250,000
PI: Kelly McMasters M.D.
Role: Biostatistician