American Journal of Biostatistics 1 (1): 9-16, 2010
ISSN 1948-9889

Modeling Heterogeneity in Phase II Clinical Trials

[1]Christopher N. Barnes and [1,2]Shesh N. Rai
[1]Department of Bioinformatics and Biostatistics, University of Louisville, United States
[2]Biostatistics Shared Facility, James Graham Brown Cancer Center,
Louisville Kentucky, United States

Abstract: Problem statement: The common assumption in non-randomized Phase II clinical trials is a homogeneous population with homogeneous response. This assumption is at odds with many trials today; a heterogeneous response due to the existence of subgroups. **Approach:** In order to examine the effects of heterogeneity on the trial outcome, a systematic platform is developed to quantify the range and classes of possible response heterogeneity using a mixed model approach. Five recent methods developed to handle heterogeneity, stratified analysis, beta-binomial models, Bayesian hierarchical models and regression models are compared and contrasted using a set of performance criteria to provide clinicians with scenarios where each method is applicable. **Results:** All methods require a priori information on the subgroup composition, a limiting factor in most trial conduct. The Bayesian methods require the least amount of assumptions, provide a methodology to share information across subgroups and allow partial subgroup outcomes, but require substantial computational resources and time. The stratified methods provide a simple improvement over the standard phase II Simon design, but lack the methodology to allow for partial subgroup stopping. **Conclusion:** The heterogeneity model provides a useful tool to model data under a heterogeneity assumption. The proper handling of heterogeneous populations under a Phase II design is a contentious debate; ignoring this fundamental assumption may lead to incorrect trial outcomes. New methods need to be developed which can include the heterogeneity structure in the trial design and allow for partial hypothesis testing.

Key words: Bayesian, over-dispersion, subgroup analysis

INTRODUCTION

Phase II clinical trials are generally single arm trials designed to estimate a response rate, π, for an experimental treatment. The most common type of trial design is the Simon 2-stage design (Ye and Shyr, 2007) with the primary assumption of a homogeneous population. In a Simon design, the number of responses, R, is assumed to follow a binomial distribution with variance $Var[R] = \pi (1-\pi)$. When the response does not comply with the assumption of homogeneity, such that $Var[R] >> \pi(1-\pi)$, the response is considered to be heterogeneous. Using methods that rely on the homogeneity assumption when heterogeneity is true can lead to biased inferences (Russek-Cohen and Simon, 1997), incorrect early stopping of the trial (Thall *et al*., 2003) or a subsequent failure of the Phase III trial resulting in a substantial loss of resources (Tuma, 2008). In many clinical trials, the possibility of response heterogeneity is handled in a less than optimal manner by applying methods that ignore the true data structure to force an assumption of response homogeneity. Complexity in design and analysis, lack of information on new methods and novelty of the heterogeneity methods seem to be the overriding motivations for the current practice of ignoring patient heterogeneity (Tuma, 2008; Wathen *et al*., 2008).

Standard practice in clinical trials to handle heterogeneity has been to conduct multiple trials (Wathen *et al*., 2008; London and Chang, 2005) or average the response profile to a response rate (Tuma, 2008; Wathen *et al*., 2008; Ayanlowo and Redden, 2008). The advantage of both methods is that they are simple and standard software exists to analyze the results; though there are several disadvantages. The first method, multiple trials, inflates the sample size by conducting multiple trials; a strain on trial resources. Due to possible low patient accrual in one or more trials, trials may not be completed; losing valuable information on the treatment efficacy over the entire population. Conducting multiple trials ignores a fundamental assumption of the motivation for a single

Corresponding Author: Christopher N. Barnes, Department of Bioinformatics and Biostatistics, University of Louisville, United States

trial; all patients share a common disease state. It can be assumed that the response rate in one subgroup will be correlated with the response rate other subgroups.

The second method, averaging, ignores the distribution of the response profile, the population subgroup proportions; possibly causing lack of association between the value of the test statistic and the trial outcome. Additionally, Phase II trials incorporate stopping boundaries to allow for the early termination of trial due to futility by conducting the trial in stages. By averaging the response profile, stopping boundaries are global, one boundary for the entire trial. This ignores the possibility of treatment futility in some subgroups and not others or the difference in futility bounds that may exist.

We develop a model to quantify heterogeneity and then apply this model to five methods that are currently available for handling patient heterogeneity, under a single trial design, to provide clinicians with a set of criteria to decide which method is applicable to a problem

MATERIALS AND METHODS

Heterogeneity model: Response heterogeneity in a population can be modeled by deconstructing the response rate into subgroups to form a response profile, $\pi = (\pi_1, \pi_2,\ldots,\pi_g)$, 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 the recent methodology for heterogeneous responses.

Let $\pi_T = (\pi_{T1}, \pi_{T2},\ldots,\pi_{Tg})$ be the vector of subgroup responses for i = 1,2,…,g subgroups where π_{Ti} is the response rate in subgroup i for treatment T = {S,E}. T = S denotes the known standard/historical treatment response and T=E denotes the hypothesized experimental treatment response. In addition, let the baseline historical response rate for the historical response profile be denoted by:

$$\pi_S^* = \arg\min_g (\pi_{Si})$$

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:

$$\delta^* = \arg\min_g (\delta_{Ei})$$

where, δ_{Ei} are the treatment effects for each subgroup, such that:

$$\pi_{Ti} = \pi_S^* + \eta_i + (\delta^* + \tau_i) I(T = E) \quad (1)$$

Where:

$0 \leq \pi_{Ti} \leq 1$ = A subgroup mixture model for heterogeneity
$I(\cdot)$ = A membership indicator. The historical response heterogeneity
η_i = A fixed prognostic effect while the treatment heterogeneity
τ_i = A predictive random effect

Using Eq. 1, 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 $i \neq i'$:

$$\begin{aligned} &\pi_{Si} \neq \pi_{Si'} \text{ and } \pi_{Ei} \neq \pi_{Ei'} \\ &\text{where } \eta_i \neq \eta_{i'} \text{ and } \tau_i = \tau_{i'} = 0 \text{ such that } \delta_i = \delta_{i'} \end{aligned} \quad (2)$$

defines the HRH class and:

$$\begin{aligned} &\pi_{Si} = \pi_{Si'} \text{ and } \pi_{Ei} \neq \pi_{Ei'} \\ &\text{where } \eta_i = \eta_{i'} = 0 \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'} \end{aligned} \quad (3)$$

defines the ARH class. In both classes, 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:

$$\begin{aligned} &\pi_{Si} \neq \pi_{Si'} \text{ and } \pi_{Ei} \neq \pi_{Ei'} \\ &\text{where } \eta_i \neq \eta_{i'}, \text{ and } \tau_i \neq \tau_{i'} \text{ such that } \delta_i \neq \delta_{i'} \end{aligned} \quad (4)$$

In Eq. 3, 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, $\pi_{Si} \neq \pi_{Si'}$. The treatment

effects are homogeneous across the subgroups, $\delta_i = \delta_{i'}$. In contrast to HRH, the heterogeneity in Eq. 3, is quantified through heterogeneous treatment effects, $\delta_i \neq \delta_{i'}$, where the estimated historical response rates are homogeneous, $\pi_{Si} = \pi_{Si'}$. 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 (2).

The general form of response heterogeneity, GRH, is a composite of both of the previous classes of response heterogeneity. The general form (4) 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, $\pi_{S1} = \pi_{S2} = \pi_{SM}$ and $\pi_{S3} = \pi_{SF}$ where $\pi_{SM} \neq \pi_{SF}$. 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:

$$\pi_{S1} = \pi_{S2} \neq \pi_{S3} \text{ and } \begin{cases} \pi_{E1} \neq \pi_{E2} \neq \pi_{E3} \\ \pi_{E1} \neq \pi_{E2} = \pi_{E3} \\ \pi_{E1} = \pi_{E2} \neq \pi_{E3} \end{cases}$$

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, absent or present, 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.

Heterogeneity methods: Five methods have been developed to handle response heterogeneity in Phase II clinical trials. 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:

$$K_m = \frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{g}\left(R_{ij} - N_{ij}\pi_{Si}\right)\right)}{\sqrt{\sum_{j=1}^{m}\left(\sum_{i=1}^{g} N_{ij}\pi_{Si}(1-\pi_{Si})\right)}} \quad (5)$$

Sample size computation and critical value determination are completed using an iterative simulation algorithm with set percentages of Type I and II errors spent in each stage (London and Chang, 2005). Prognostic and/or predictive heterogeneity is modeled through the choice of simulation parameters using model (1). 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 II errors for the trial. 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 Eq. 5 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 $\sum_{j=1}^{m}\sum_{i=1}^{g} n_{ij}\pi_{Si}$ and the denominator of (5) are constants given $(n_{i1},\ldots,n_{im},\pi_{Si})$. 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:

$$P(R_m = r_m) = \sum_{\substack{r_{1m}+,\ldots,+r_{gm}=r_m; \\ 0\le r_{im}\le n_{im}}} \prod_{i=1}^{g}\binom{n_{im}}{r_{im}}\pi_{0i}^{r_{im}}(1-\pi_{0i})^{n_{im}-r_{im}} \quad (6)$$

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 (6) 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 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 (1989) designs.

The third method, the beta-binomial distribution has been previously proposed as a model that can account for heterogeneity in binary outcome models (Dragalin and Fedorov, 2006). 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, $\pi_i \sim \text{beta}(a,b)$. Then $R_{i1} \mid \pi_i$, has a binomial distribution. The marginal of R_1 is a beta-binomial with probability function:

$$P(R_1 = r_1) = \binom{n_1}{r_1} \frac{\text{beta}(r_1 + a, n_1 - r_1 - b)}{\text{beta}(a,b)}$$

The mean and variance are:

$$E[R_1] = n_1\pi \text{ and } \text{Var}[R_1] = n_1\pi(1-\pi)\left\{1 + \frac{\rho}{1+\rho}(n_1 - 1)\right\}$$

where, $\pi = \frac{a}{a+b}$. 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,b), the sample size and test statistics can be calculated based on the type of difference to be detected (Hendriks *et al*., 2005). 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.

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 of 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:

$$\begin{aligned} f(R,\pi,\psi) &= l(R \mid \pi)p(\pi \mid \psi)p(\psi) \\ &= \left\{\prod_{i=1}^{g} \underbrace{l(R_i \mid \pi_i)}_{\substack{\text{data} \\ \text{likelihood}}} \underbrace{p(\pi_i \mid \psi)}_{\text{prior}}\right\} \underbrace{p(\psi)}_{\text{hyperprior}} \end{aligned}$$

with trial decision making using the posterior distribution:

$$P(\pi \mid R) = \frac{\int f(R,\pi,\psi)d\psi}{\int f(R,\pi,\psi)d\theta d\psi}$$

Due to the intractability and high dimension of the posterior, MCMC methods are used to compute the posterior probabilities for each stage of the trial (Gilks *et al*., 1996). The fourth heterogeneity method, Bayesian normal-binomial hierarchical model used in Thall *et al.* (2003) is based on the logit model (Collet, 2003) and is constructed such that:

$$\begin{aligned} &\theta_i = \text{logit}(\pi_i) \overset{iid}{\sim} N(\mu, \sigma^2) \\ &\text{with } \psi = (\mu, \sigma^2),\ \mu \sim N(\upsilon_1, \varphi_1{}^2) \text{ and } \sigma^2 \sim N(\upsilon_2, \varphi_2{}^2) \end{aligned} \tag{7}$$

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:

$$P(\pi_{Ei} > \pi_{Si} \mid \text{data}) < l \tag{8}$$

and

$$P(\pi_{Ei} > \pi_{Si} \mid \text{data}) \geq u \tag{9}$$

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.

The fifth method, Bayesian normal-binomial regression model or BANCOVA model, was proposed by Wathen *et al*. (2008). To compare the model with the earlier heterogeneity notation, the model was reparameterized. The model:

$$\text{logit}\left(\pi_{T\,g}(\theta)\right) = \xi + \sum_{i=1}^{g}\left\{\eta_i + \tau_i I(T = E)\right\} I(G = g) \qquad (10)$$

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 (1) and the model (10) which the models mean response rate on the logit scale. Model (10) has no assumption on the structure of the variance as in model (7), where $\theta_i = \text{logit}\left(\pi_i\right) \overset{iid}{\sim} N(\mu, \sigma^2)$ 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, Thall *et al*. (2003) and Wathen *et al*. (2008) 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 (Wathen *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 (8) and (9) 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.

RESULTS

Five methods, including the standard Simon design, were compared using a set of performance criteria, type of trial design, classes of applicable heterogeneity, types of stopping boundaries applicable, allowance of partial efficacy/futility, effect under lack of heterogeneity, sample sizes computation, robustness under parameter misspecification and computational time. A summary of the comparison criteria and results are in Table 1.

The class of heterogeneity that is accounted for in each method varies and should be the starting point in deciding the appropriateness of a method for a given problem. The conditional stratified method can accommodate all three classes of heterogeneity, while the unconditional method relies heavily on accurate estimates of the true population proportion of each subgroup in order to handle ARH or GRH. The beta-binomial distribution is able to account for all three heterogeneity types. The Bayesian methods do not need estimates of the subgroup proportions, but are designed to only accommodate certain classes of heterogeneity. The hierarchical method is designed to accommodate ARH, while the BANCOVA method is designed to accommodate HRH, ARH and GRH.

Table 1: Comparison of methods under different criteria to handle patient heterogeneity; Simon 2 stage design, Unconditional stratified (UC), Conditional stratified (C), Beta-binomial, Bayesian normal-Binomial Hierarchical Model (BHM), and Bayesian binomial-normal regression model (BANCOVA)

	Analysis method					
Criteria	Simon	UC stratified	C stratified	Beta binomial	Normal-B BHM	Normal-B BANCOVA
Type of trial	k-stage	k-stage	k-stage	k-stage group	Group	Group
Heterogeneity type	None	HRH	HRH	HRH		HRH
		ARH	ARH	ARH	ARH	ARH
		GRH	GRH	GRH		GRH
Require subgroup knowledge	No	Yes	Yes	No	Yes	Yes
Stopping boundary	Global	Global	Global	Global, subgroup	Subgroup	Subgroup
Partial efficacy	No	No	No	Yes	Yes	Yes
Sample size specified	Yes	Yes	Yes	Yes	Upper bound	Upper bound
Robust	No	No	Yes	Yes	Yes	Yes
Lack of heterogeneity	No effect	No effect	No effect	No effect	Minimal effect	Moderate effect
Computational time	Minimal	Minimal	Moderate	Minimal	Extensive	Extensive

Two types of stopping boundaries exist, global and subgroup. The Bayesian methods allow for subgroup specific stopping boundaries while the stratified and beta-binomial methods use global boundaries. In addition, the BANCOVA model allows for unique subgroup stopping boundaries further refining the boundaries based on the prognostic data from individual subgroups.

The type of trial may also play a part in the selection of an appropriate model for a particular problem. The stratified methods are k-stage trials. The usual number of stages for this type of trial is 2-3 stages with unequal sample sizes in each stage derived from the operating characteristics of the method (Chow *et al*., 2007). The Bayesian methods are group sequential with a length of usually greater than 3 stages with an equal sample size in each stage (Todd, 2005). To reduce the time necessary to complete a Bayesian group sequential trial, the trial is usually a modified group sequential; instead of waiting until all patients has been evaluated, after a set number of patients, the unevaluated patients are assumed to be positive responses and Eq. 8 is computed. If $P(\pi_{Ei}>\pi_{Si}|data)<1$ where the data includes the unevaluated assumed positive patients, this gives an early determination of futility and can allow the early stopping of the subgroup without waiting until all patients have been evaluated speeding up the trial conduct time. The beta-binomial can be used in either a k-stage or group sequential context.

For the stratified and beta-binomial models, sample size computation is performed before the trial commences. A minimum sample size is derived from the standard binomial sample size calculation and then iteratively increased until the power and size requirements are met using the test statistic for the stratified methods or a formula is used in the case of the beta-binomial model. For the Bayesian methods, a target range is specified with a minimum and maximum sample size (Thall *et al*., 2003; Wathen *et al*., 2008). The trial is conducted by splitting the maximum sample size into sequential groups with decisions made at the end of each group sequence up to the maximum sample size. If the maximum sample size is reached and there is not sufficient evidence to reject the null hypothesis in a subgroup, the experimental treatment is considered inferior in that subgroup; though no evidence is presented that this choice of sample size selection minimizes the false positive and false negative rates.

The robustness of all of the methods relies on the parameters estimates of the methods. The stratified methods are contingent on correctly specifying the proportion of each of the subgroups in the population through the sampling weights. If this estimate is biased, additional patient resources will need to be accrued after trial commencement to meet the original power and size constraints. The conditional stratified method allows for a greater flexibility due to the multiplicity of possible solutions. The beta-binomial model relies on having an accurate estimate for the parameters of the beta distribution on which the heterogeneity is constructed. Inaccurate estimates will mitigate the performance of the model. The Bayesian methods rely on estimates for the hyper-priors; though the use of the proposed algorithms mitigates the bias in the hyperpriors. The advantage of the Bayesian methods over the Frequentist methods is that they do not rely on estimates for the proportion of each subgroup. As such, they are more robust to model misspecification.

While the purpose of this study is to advocate control for possible heterogeneity in the population, there will be cases where the heterogeneity is appropriately accounted for in the analysis but it not actually present. The strength of any heterogeneity method under a heterogeneous population must also maintain strength under population homogeneity. The three Frequentist methods are robust under lack of heterogeneity; the test statistics degenerate into the standard binomial form test statistic for a homogeneous population. The Bayesian methods lose a small amount of power under lack of heterogeneity (Wathen *et al.* 2008).

The last criterion in method performance is computational time. As with other statistical methods, the sensitivity and flexibility of a method is contrasted with the computational time necessary to attain the desired characteristics. The unconditional stratified and beta-binomial methods use the least computational time. The conditional stratified method has an increase of time due to the multiplicity of the solutions. The Bayesian methods require substantial computational time due to the intractability of the posterior distribution. Thall *et al*. (2003) and Wathen *et al*. (2008) suggest the use of distributed processing systems to speed up the necessary time (Thall *et al*., 2003). This increase in trial resources should be balanced when considering a Bayesian method. This increase in computational cost and complexity may be a motivating factor in why the majority of clinical trials today are Frequentist in nature (Lee and Feng, 2005).

DISCUSSION

To our knowledge, broadly speaking, five methods currently exist for handling response heterogeneity. Each method was developed to address a specific type of heterogeneity by optimizing trial resources through

the use of a single trial, an advantage of using any of the five methods over conducting multiple trials. All the methods require one fundamental assumption, the known existence of subgroups before the trial.

The stratified methods of London and Chang were developed to handle a combination of prognostic and/or predictive heterogeneity for unbalanced subgroups using a single test statistic; rejecting or accepting the hypothesis of mean treatment efficacy over the entire population, a global hypothesis. The beta-binomial method was developed to allow for unidentified heterogeneity in correlated responses. The Bayesian hierarchical method of Thall *et al*. (2003) was developed to account for predictive heterogeneity in unbalanced subgroups using identical subgroup hypotheses. The BANCOVA method of Wathen *et al*. (2008) and Thall *et al*. (2003) was developed to account for both prognostic and predictive heterogeneity under subgroup specific hypotheses. In both of the Bayesian methods, the overriding motivation is to allow partial treatment efficacy across the subgroups, an aspect lacking in the stratified methods.

The heterogeneity model provides a critical component for the comparison of methods. The primary factor in deciding which method is applicable is determining which class of heterogeneity the data is assumed to follow. The conditional stratified and BANCOVA models are the most robust to heterogeneity. The beta-binomial method does well under all three heterogeneity classes but suffers an identifiability problem with the source of heterogeneity; individual components of the heterogeneity are not explicitly modeled resulting in a tradeoff clinically, a loss of information on the source of the heterogeneity.

A drawback of the unconditional stratified method is the reliance of the test statistic an accurate estimates for the sampling distribution of the subgroups. If patient accrual does not match the sampling estimates, a chronological bias is introduced into the test statistic(s) and the resulting test outcome is not valid (London and Chang, 2005; Srivastava *et al*., 2007). The conditional stratified method is more robust to accrual divergences removing the estimation bias by solving for multiple solutions. The Bayesian methods do not suffer from the issue of accurate sampling estimation, but suffer from the identification of subgroups issue which is an inherent problem in all of the contrasted methods.

Subgroup specific stopping boundaries allow for individual subgroup stopping boundaries similar to conducting multiple trials while a global stopping boundary only allows all subgroup trial termination similar to conducting an averaged response trial. An optimal heterogeneity method would incorporate the structure of the response profile, e.g., the subgroups, into the hypothesis testing. The stratified methods only include global boundaries while the Bayesian methods include subgroup boundaries; homogeneous boundaries for the hierarchical model and possibly unique boundaries for the BANCOVA model.

The third critical comparison between methods is sample size. The stratified methods and beta-binomial method determine a fixed sample size before trial conduct while the Bayesian methods rely on a maximum estimate for sample size. If expected accrual can accommodate this maximum sample size, say 100 patients, then the Bayesian methods are applicable. If expected accrual is determined to be much smaller, say 50, then the Bayesian methods may be precluded as a suitable method.

CONCLUSION

Each method has a list of strengths and possible weaknesses under different classes of heterogeneity. No method currently exists that optimizes the complete set of comparison criteria in this study. The stratified methods require smaller sample sizes, are only moderately computationally complex and are robust under no heterogeneity. The disadvantage of the unconditional form over the conditional form is the need to accurately estimate population proportions through sampling weights. A disadvantage of both methods is the lack of subgroup specific stopping boundaries. While the Bayesian methods, Bayesian hierarchical model and BANCOVA, require a larger sample size under a non-informative prior and more computational time, they allow the use of subgroup specific stopping boundaries refining patient efficacy characteristics. The beta-binomial distribution model provides a model for a middle of the road alternative to the other methods. It works under all three classes of heterogeneity, is computationally moderate and the necessary sample size is comparable to the stratified methods, but lacks the etiology of heterogeneity information of the other methods.

The limiting factor in the application of all the four main methods, stratified and Bayesian, is the a priori knowledge of the existence of subgroups. Each of the four methods is dependent on knowledge of the distribution of subgroups. If no knowledge is known about the existence of subgroups, none of the methods will be able to provide adequate inferences. In contrast, the beta-binomial does not need information on subgroups, but lacks the ability to differentiate the source of the heterogeneity, an important clinical aspect of the trial.

Methods need to be developed that can be applied to a problem without any knowledge of the existence of heterogeneity that maintain the desirable attributes of each of the compared methods, subgroup etiology, sharing of resources across subgroups, while maintaining the desirable attributes of a Simon design, high probability of early termination in the first stage and small sample sizes, if heterogeneity does not exist.

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: 428-438. DOI: 0.1016/j.cct.2007.10.003

Chow, S., J. Shao and H. Wang, 2007 Sample Size Calculations in Clinical Research. 2nd Edn., Taylor and Francis Group, ISBN: 9781584889823, pp: 117-143.

Collet, D., 2006. Modeling Binary Data. 2nd Edn., Taylor and Francis, ISBN: 1-584-88324-3, pp: 369.

Dragalin, V. and V. Fedorov, 2006. Design of multi-centre trials with binary response. Stat. Med., 25: 2701-2719. DOI: 10.1002/sim.2406

Gilks, W., R. Richardson and D. Spiegelhalter, 1996. Markov Chain Monte Carlo in Practice, Taylor and Francis Group, ISBN: 0-412-05551-1, pp: 206-209.

Hendriks, J.C.M.,S. Teerenstra, J.P.E. Punt-van der Zalm, A.M.M. Wetzels, J.R. Westphal and G.F. Borm, 2005. Sample size calculations for a split-cluster, beta-binomial design in the assessment of toxicity. Stat. Med., 24: 3757-3772. DOI: 10.1002/sim.2412

Hunt, D.L. and S.N. Rai, 2005. A new threshold dose-response model including random effects for data from developmental toxicity studies. J. Applied Toxicol., 25: 435-439. DOI: 10.1002/jat.1092.

Lee, J.J. and L. Feng, 2005. Randomized phase II designs in cancer clinical trials: Current status and future directions. J. Clin. Oncol., 23: 4450-4457. DOI: 10.1200/JCO.2005.03.197

Lee, P.M., 2005. Bayesian Statistics: An Introduction. 3rd Edn., Arnold, ISBN: 0-340-81405-5, pp: 73.

London, W.B. and M.N. Chang, 2005. One-and two-stage designs for stratified phase II clinical trials. Stat. Med., 24: 2597-2611. DOI: 10.1002/sim.2139

Russek-Cohen, E. and R.M. Simon, 1997. Evaluating treatments when a gender by treatment interaction may exist. Stat. Med., 16: 455-464. PMID: 9044532

Simon, R., 1989. Optimal two-stage designs for phase II clinical trials. Control Clin. Trials, 10: 1-10. PMID: 2702835

Srivastava, D.K., S.N. Rai and J. Pan, 2007. Robustness of an odds-ratio test in a stratified group sequential trial with a binary outcome measure. Biome. J., 49: 351-364. DOI: 10.1002/bimj.200610265

Stadler, W.M., 2007. The randomized discontinuation trial: A phase II design to assess growth-inhibitory agents. Mol. Cancer Ther., 6: 1180-1185. DOI: 10.1158/1535-7163.MCT-06-0249

Thall, P.F., J.K. Wathen, B.N. Bekele, R.E. Champlin, L.H. Baker and R.S. Benjamin, 2003. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. Stat. Med., 22: 763-80. DOI: 10.1002/sim.1399

Todd, S., 2007. A 25-year review of sequential methodology in clinical studies. Stat. Med., 26: 237-252. DOI: 10.1002/sim.2763

Tuma, R.S., 2008. Examining heterogeneity in phase II trial designs may improve success in phase III. J. Natl. Cancer Inst., 100: 164-166. DOI: 10.1093/jnci/djn006

Wathen, J.K. *et al*., 2008. Accounting for patient heterogeneity in phase II clinical trials. Stat. Med., 27: 2802-2815. DOI: 10.1002/sim.3109

Ye, F. and Y. Shyr, 2007. Balanced two-stage designs for phase II clinical trials. Clin. Trials, 4: 514-524. DOI: 10.1177/1740774507084102

Young-Xu, Y. and A. Chan, 2008. Pooling overdispersed binomial data to estimate event rate. BMC Med. Res. Methodol., 8: 58. DOI: 10.1186/1471-2288-8-58