PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 227–244
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.300

Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset‡,§

MAIN PAPER

Sue-Jane Wang[1,*,†], Robert T. O'Neill[1] and H. M. James Hung[2]

[1] *Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA*

[2] *Division of Biometrics I/OB, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA*

With the advances in human genomic/genetic studies, the clinical trial community gradually recognizes that phenotypically homogeneous patients may be heterogeneous at the genomic level. The genomic technology brings a possible avenue for developing a genomic (composite) biomarker to predict a genomically responsive patient subset that may have a (much) higher likelihood of benefiting from a treatment.

Randomized controlled trial is the mainstay to provide scientifically convincing evidence of a purported effect a new treatment may demonstrate. In conventional clinical trials, the primary clinical hypothesis pertains to the therapeutic effect in all patients who are eligible for the study defined by the primary efficacy endpoint. The aspect of one-size-fits-all surrounding the conventional design has been challenged, particularly when the diseases may be heterogeneous due to observable clinical characteristics and/or unobservable underlying the genomic characteristics.

Extension from the conventional single population design objective to an objective that encompasses two possible patient populations will allow more informative evaluation in the patients having different degrees of responsiveness to medication. Building in conventional clinical trials, an additional genomic objective can generate an appealing conceptual framework from the patient's perspective in addressing personalized medicine in well-controlled clinical trials. There are many perceived benefits of personalized medicine that are based on the notion of being genomically proactive in the identification of disease and prevention of disease or recurrence. In this paper, we show that an adaptive design approach can be constructed to study a clinical hypothesis of overall treatment effect

*Correspondence to: Sue-Jane Wang, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. FDA, HFD-700, WO 22, Mail Stop 6105, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA.
† E-mail: suejane.wang@fda.hhs.gov
‡ The views expressed in this article are not necessarily those of the US Food and Drug Administration.

Published in 2007 by John Wiley & Sons, Ltd.

and a hypothesis of treatment effect in a genomic subset more efficiently than the conventional non-adaptive approach. Published in 2007 by John Wiley & Sons, Ltd.

Keywords: ***alpha allocation; genomic composite biomarker; adaptive design; futility; prevalence; effect size ratio***

1. INTRODUCTION

Novel clinical trial designs are becoming popular as a part of pharmacogenomics clinical trials to evaluate treatment effects related to genomic profiles or genomic composite biomarkers [1,2]. Pharmacogenomics is the science of determining how the benefits and adverse effects of a drug vary among a target population of patients based on genomic features of the patient's germ line and diseased tissue [2]. The nature of such a study design entails two critical aspects. One aspect is to identify a patient subset possessing the genomic profile that is sensitive to the particular treatment under study. Another aspect is to maximize the overall treatment effect in the study patient population.

To illustrate such a design, consider a binary genomic (composite) biomarker that is used to classify patients into two mutually exclusive genomic subgroups: those who are classified as marker positive ($g+$) and those classified as marker negative ($g-$). Conventional clinical trial objectives require that the overall treatment effect in the entire randomized study population be established before exploring the effect in a genomic subgroup. The overall treatment effect is defined as the average effect for the entire randomized study population derived by weighting the treatment effect in each genomic subgroup by the sample size fraction of $g+$ and that of $g-$. However, the treatment benefit in the genomic marker positive subgroup is a joint function of the $g+$ sample size and the magnitude of the treatment effect size in $g+$. When the combination of the two weighted genomic subgroups does not show an overall treatment effect statistically, the conventional trial approach is not only inefficient but also may fail to assess the genomic biomarker predictivity of the clinical or therapeutic response. This is simply a variation on including at the planning stage an evaluation of the heterogeneity of treatment effect, but, without formal testing of heterogeneity necessarily.

The genomic composite biomarker may be a single biomarker, e.g. HER2 [3], or derived from a high-dimensional expression profiles, e.g. [4,5]. The diagnostic value of a genomic composite biomarker in controlled trials is described in the situations depicted in Table I. When a treatment effect exists only in the $g+$ patient subset, such as Scenario A of Table I and e.g. [1,2], the genomic biomarker is predictive of treatment effect, but not disease response. The conventional clinical trial design and analysis are unlikely to be able to detect a treatment effect limited to the $g+$ subgroup, especially when the prevalence of $g+$ is low. When there is no treatment effect in any subgroup, the

Table I. True response rates by genomic status and by treatment intervention.

	Scenario A		Scenario B		Scenario C	
Genomic Status*	Control (%)	Drug A (%)	Control (%)	Drug B (%)	Control (%)	Drug C (%)
$g-$	33	33	39	39	39	46
$g+$	33	50	61	61	61	75

*$g+$ or $g-$ is patient's genomic status determined from a diagnostic assay.

genomic biomarker may only be prognostic of the underlying disease mechanism or disease prevalence shown in Scenario B outcome. The genomic biomarker can also be prognostic of disease response and prognostic of a therapeutic effect, as shown in Scenario C, e.g. [5,6], resulting in a therapeutic effect that is quantitatively different in each genomic biomarker subgroup. In this case, one might view the genomic biomarker as predictive of differential treatment effect. For discussion, we will use the commonly known term prognostic (of therapeutic effect) for Scenario C.

The terms flexible design and adaptive design (AD) have been used interchangeably often in the statistical literature, e.g. [7,8]. Unlike the more general flexible design described in [7], which does not require complete pre-specification of the adaptation rules and is very flexible, the AD we propose in this paper requires a pre-specified adaptivity with a defined flexibility limited to a selected number of options. In this design, a prospectively planned adaptation may consider modification to study the hypothesis and/or reallocating the sample size midstream to achieve the planned objective. Although sample size reestimation can be incorporated as an additional adaptive feature, it is not considered in this paper.

We explore an AD approach to address a clinical composite hypothesis [1,2] that includes a hypothesis of overall treatment effect and a hypothesis of treatment effect in the $g+$ subset. The performances of the AD/analysis methods are compared with the non-AD approach and those in the AD approach of Freidlin and Simon (FS) [9]. In Section 2, the framework is set to test each of these individual objectives. In Section 3, the bivariate normal model incorporating the correlation between the two test statistics for each hypothesis derived from the subset and the overall trial population is described and later (Section 5) is used to highlight the enhanced treatment effect size that can be detected in the $g+$ subset. Other analysis methods are used for power comparison. Monte Carlo simulation studies compare the performances of the various design/method combinations. These results are summarized in Section 4. Section 5 considers some issues pertaining to statistical design and analysis in a targeted pharmacogenomics clinical trial. Concluding remarks are given in Section 6.

2. THE PROBLEM

We consider a parallel group randomized clinical trial to evaluate an experimental treatment relative to a control treatment. The study consists of two mutually exclusive genomic patient subgroups $g+$ and $g-$. Let μ_1 be the true mean response of the primary efficacy outcome for the experimental treatment and μ_2 the true mean response for the control treatment. Assume that the response variable in the treated and the control groups has an equal variance denoted by σ^2. The standardized treatment effect size Δ is defined by $\Delta = (\mu_1 - \mu_2)/\sigma$. Assume an equal sample size $N/2$ per treatment group. Let Z_N be the test statistic for testing whether a treatment effect exists in the entire study population, which is assumed to be approximately normally distributed with mean $\sqrt{N/4}\Delta$ and variance 1. Similarly, in the genomic $g+$ subset, let μ_{1g+} be the true mean response for the test treatment, μ_{2g+} the true mean response for the control treatment, and σ^2_{g+} the common variance. The standardized effect size of the $g+$ subset is $\Delta_{g+} = (\mu_{1g+} - \mu_{2g+})/\sigma_{g+}$. Assume an equal sample size $M/2$ $(M < N)$ in the test and the control study arms in the pre-specified $g+$ subset. Let Z_M be the test statistic assumed to be approximately normally distributed with mean $\sqrt{M/4}\Delta_{g+}$ and variance 1. When the genomic subset is prospectively planned *a priori*, a pertinent alternative composite hypothesis of interest for the study is

$$H_1 : \Delta > 0 \quad \text{or} \quad \Delta_{g+} > 0$$

The alternative hypothesis that the test treatment is superior to the control treatment in the overall patient population is $H_{1a} : \Delta > 0$. For the genomic subset, the corresponding alternative hypothesis is $H_{1g+} : \Delta_{g+} > 0$. Their respective null hypotheses are $H_{0a} : \Delta = 0$ and $H_{0g+} : \Delta_{g+} = 0$. Also necessary to consider is the intersection null hypothesis that neither all patients nor genomically selected

patients benefit from the new treatment

$$H_0 = H_{0a} \cap H_{0g+} = \{\Delta = 0 \text{ and } \Delta_{g+} = 0\}$$

To test the two null hypotheses, H_{0a} and H_{0g+}, we consider multiplicity adjustment methods that achieve a strong control of the experimentwise type I error probability needed for testing a composite hypothesis as described above. In addition, it is desirable that the study objective provides an opportunity to test the pre-specified genomic subset, regardless of the outcome of testing for all eligible patients.

3. STATISTICAL ANALYSIS METHODS

Under H_0, Z_N and Z_M are approximately jointly distributed as a bivariate normal with zero mean vector and covariance $\sqrt{M/N}$. That is,

$$\begin{pmatrix} Z_N \\ Z_M \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{M/N} \\ \sqrt{M/N} & 1 \end{pmatrix}\right)$$

Let α_a be the level of statistical significance assigned to the testing of $\Delta \leqslant 0$; let α_{g+} be the significance level for testing $\Delta_{g+} \leqslant 0$ in the genomic subset. Let α be the target level of type I error probability associated with testing H_0 which is the overall experimentwise error probability at issue.

3.1. General bivariate normal method

For testing $H_{0a} : \Delta = 0$ and $H_{0g+} : \Delta_{g+} = 0$, the overall experimentwise type I error probability, associated with the two correlated tests Z_N and Z_M with the respective critical values z_{α_a} and $z_{\alpha_{g+}}$, is

$$\begin{aligned} &p\{Z_N > z_{\alpha_a} \text{ or } Z_M > z_{\alpha_{g+}} | H_0\} \\ &\quad = 1 - p\{Z_N \leqslant z_{\alpha_a} \text{ and } Z_M \leqslant z_{\alpha_{g+}} | H_0\} \end{aligned} \tag{1}$$

The formulation in equation (1) accounts for the correlation between the test statistics for the subset and the overall population. Let $f = M/N$. Z_N can be written as a weighted combination of Z_M for the $g+$ subset and the Z-statistic Z_{N-M} for the $g-$ subset. Thus, we have

$$\begin{aligned} &p_{H_0}\{\sqrt{f}Z_M + \sqrt{(1-f)}Z_{N-M} \leqslant Z_{\alpha_a}, Z_M \leqslant Z_{\alpha_{g+}}\} \\ &\quad = p_{H_0}\left\{Z_{N-M} \leqslant \frac{Z_{\alpha_a} - \sqrt{f}Z_M}{(1-f)}, Z_M \leqslant Z_{\alpha_{g+}}\right\} \\ &\quad = E\left\{p_{H_0}\left\{Z_{N-M} \leqslant \frac{Z_{\alpha_a} - \sqrt{f}Z_M}{\sqrt{(1-f)}}\right\} I_{(Z_M \leqslant Z_{\alpha_{g+}})}\right\} \\ &\quad = \int_{-\infty}^{z_{\alpha_{g+}}} \phi\left(\frac{z_{\alpha_a} - \sqrt{f}Z}{\sqrt{(1-f)}}\right)\phi(z)\,\mathrm{d}z \end{aligned} \tag{2}$$

By setting equation (2) to $1 - \alpha$, α_{g+} can be solved for given α_a and f. The subset α_{g+} so calculated is shown in Figure 1 for three cases: (i) $\alpha_a = \alpha_{g+} = 0.0125$ in dash diamond, (ii) $\alpha_a = 0.02$ and $\alpha_{g+} = 0.005$ in solid square, and (iii) $\alpha_a = 0.005$ and $\alpha_{g+} = 0.02$ in dash triangle. Incorporation of the correlation can result in $\alpha_a + \alpha_{g+} > \alpha$, we consider $\alpha_a \leqslant \alpha$ and $\alpha_{g+} \leqslant \alpha$. For instance, suppose $\alpha_a = 0.020$ is considered for testing the overall trial population. If $f = 0.3$, then $\alpha_{g+} = 0.0071$. If one does not account for the correlation between the subset and the overall trial results, a one-sided α_{g+} is always 0.005. Since the experimentwise error rate is controlled at 0.025 level, the error rates for the overall trial and the subset are inversely related; thus, one has the choice whether to allocate a larger error rate for the overall effect or for the genomic subset effect. The higher the sample size fraction is, the higher is the α_{g+}, which is less than α. When $f = 1.0$, the subset does not exist, all subjects are entered as $g+$, and the usual test of an overall treatment effect at a one-sided 0.025 level applies.

The use of correlation for alpha allocation may be challenged when the trial is not prospectively planned with a sample size to detect a pre-specified (presumably larger) treatment effect in the genomic subset. In this case, the sample size fraction M/N is unknown at the trial design stage and is not determined until the end of the study. Thus, the correlation estimated from the sample size ratio observed or from the sample size ratio itself is a random variable. Practically, one can use the worst assumed correlation for the alpha allocation.

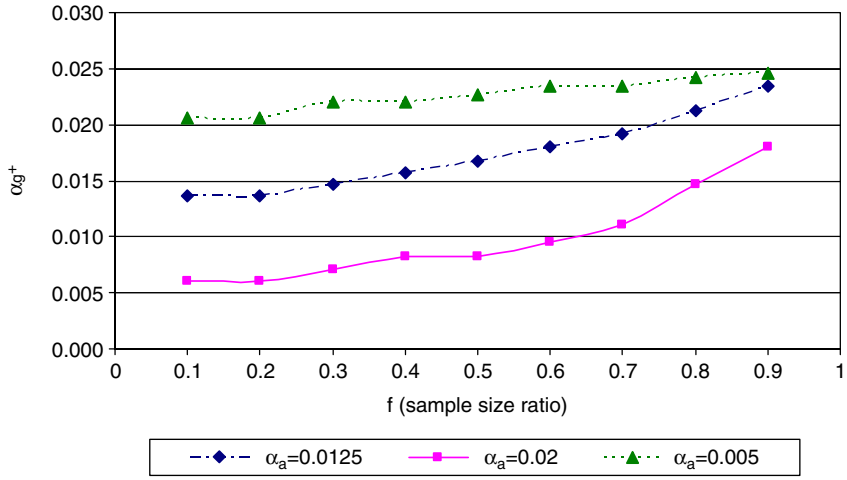

Figure 1. Subset alpha-level (α_{g+}) required to maintain the overall type I error at one-sided 0.025 level.

Two multiplicity adjustment methods that have a strong control of experimentwise type I error rate are considered. One method is allocating α to the two hypotheses, such as that proposed by Moye and Deswal [10] requiring the sum of alphas allocated be at, for instance, a one-sided $\alpha = 0.025$ level. But this approach does not account for the correlation and thus is conservative. Whether one allocates the individual alphas, α_a and α_{g+}, equally or unequally, H_{0g+} can always be tested irrespective of the test result on H_{0a}. Accounting for the correlation between Z_N and Z_M may yield some gains in statistical efficiency. The other method is applying Hochberg's step-up method [11], which can provide flexibility for testing H_{0g+} in the sense that the testing for $g+$ does not require the test of H_{0a} to achieve statistical significance first. If the correlation between Z_N and Z_M is incorporated under bivariate normal, a step-down version of Dunnett's test can be more powerful than the Hochberg method, which computes critical values assuming independence between Z_N and Z_M [12]. The following two design approaches are considered: a non-AD and an AD.

3.2. Fixed study design with no adaptation

In the conventional fixed design (FD) that does not include any interim look, the study is prospectively planned for testing H_{0a} and H_{0g+} only at the trial end. In principle, pre-specification of a $g+$ subgroup hypothesis suffices to properly control the experimentwise type I error probability. One could consider planning the study based on the effect size in the all-patient population, particularly when it is not anticipated that the treatment effect differs between the $g+$ and $g-$ genomic patient subsets. In this situation, the overall treatment effect would be estimated by the average effect weighted by the sizes of the subpopulations, that is $\hat{\Delta} = f\hat{\Delta}_{g+} + (1-f)\hat{\Delta}_{g-}$. However, if the effect in the genomic $g+$ subset is of specific interest that is considered either predictive or prognostic, the unknown effect size for the $g+$ genomic subset should be (much) larger than that in the all-patient population, i.e. $\Delta_{g+} \gg \Delta \geqslant 0$. In such cases, the sample size M for the $g+$ subset is dictated by the final sample size fraction at study completion. One can also pre-specify a minimum subset sample size, that is at least $M < N$ allowing for making a valid statistical inference. It can also be $f_p N$, where f_p is the prevalence of $g+$ obtained from the literature.

For the split-alpha method allocating α_a to testing $\Delta \leqslant 0$ and α_{g+} to testing $\Delta_{g+} \leqslant 0$ with $\alpha_a + \alpha_{g+} = \alpha$ under the FD, the power function of Δ in

all-patient population is

$$P(P_a \leqslant \alpha_a|\Delta) = \Phi(-z_{\alpha_a} + \sqrt{N}\Delta) \quad (3)$$

and the power function of α_{g+} in the $g+$ genomic patient subset is

$$P(P_{g+} \leqslant \alpha_{g+}|\Delta_{g+}) = \Phi(-z_{\alpha_{g+}} + \sqrt{M}\Delta_{g+}) \quad (4)$$

For the Hochberg method, the power functions for Z_N and Z_M are a mixture of (3) and (4) depending on the N, M, Δ Δ_{g+}, and α.

3.3. Adaptive study design

We consider a two-stage design in which the total sample size is planned according to the non-adaptive approach. A prospectively planned interim look is performed at the end of Stage I, time t $[0 < t < 1]$, on the primary efficacy endpoint and the total available safety information. A decision based on the data of the tN patients from Stage I is then made on whether to maintain the original study plan or to modify the plan to recruit only the patients classified as $g+$ in Stage II and to test only H_{0g+} at the study end.

The Stage I null hypothesis is denoted by $H_0^1 = H_0$. False-positive errors can occur in three types of null distributions that are summarized in Table II. The two-stage AD provides some flexibility at Stage I to make a decision based on the probability of observing a treatment effect in the $g-$ subgroup. If in the $g-$ subgroup the probability of showing no or futile treatment effect is high or if the toxicity is of serious enough concern, the patient recruitment into Stage II may be limited to only the $g+$ genomic subset, but the total sample size is kept at the originally planned N. Depending on the efficacy futility boundary b_t and the safety margin b_t^{safety} used at time t, the Stage II null hypothesis can be expressed as

$$H_0^2 = \begin{cases} H_{0g+} & \text{if } Z_{3t} \leqslant b_t \text{ or } Z_{3t}^{\text{safety}} \geqslant b_t^{\text{safety}} \\ H_0 & \text{otherwise} \end{cases}$$

where Z_{3t} and Z_{3t}^{safety} are the test statistics of the corresponding primary efficacy outcome and the primary safety outcome for the $g-$ subgroup at time t. Although the formal test Z_{3t}^{safety} for the primary adverse event can be incorporated, safety assessment is generally empirically based rather than formally tested in practice. At the end of Stage I, if the treatment effect estimate for Δ_{g-} is zero or negative, and the test for the $g-$ subset meets the boundary $Z_{3t} < b_t$, where $b_t \leqslant 0$, then stop the recruitment of $g-$ patients and allocate the remaining sample size $(1-t)(N-M)$ originally planned for the $g-$ patients subgroup to only patients whose genomic status is $g+$, so that the total sample size for $g+$ patients in the final analysis is $M' = tM + (1-t)N$.

Consequently, the conventional test for H_{0g+} : $\Delta_{g+} \leqslant 0$ at the two-stage study end is

$$U_2 = \sqrt{\frac{tM}{M'}}Z_{2t} + \sqrt{1 - \frac{tM}{M'}}W_{2t}$$

where Z_{2t} is the test at time t and W_{2t} is the Z test for $(1-t)N$ after time t. An alternative option is the CHW-weighted Z test [13]

$$T_2 = \sqrt{\omega}Z_{2t} + \sqrt{1-\omega}W_{2t}$$

where ω is a fixed weight, e.g. $\omega = t$.

Let $p_a^1, p_{g+}^1, p_a^2, p_{g+}^2$ be the nominal p-values for the all-patient trial (subscript a) and the genomic subset (subscript $g+$) after the first-stage interim analysis (superscript 1) and the second-stage final analysis (superscript 2), respectively. Although, in principle, early trial termination may be considered based on a significant treatment effect, this possibility is not considered in the proposed AD. The intent is to avoid a seemingly superior observed effect early in the trial that may be due to an observed random high estimate from the interim look, particularly when the treatment effect in the genomic $g+$ subset is estimated from

Table II. The null scenarios of effect size for individual hypotheses.

Scenario	All-patient population	$g+$ subset	$g-$ subset
1	0	0	0
2	$0 < \Delta < \Delta_{g-}$	0	$\Delta_{g-} > 0$
3	0*	$\Delta_{g+} > 0$	$\Delta_{g-} < 0$

*A special case of $\Delta_{g-} < \Delta < \Delta_{g+}$ in Scenario 3.

a small sample size in the early stage of the trial because the prevalence of $g+$ is low. The rationale for the Stage I futility assessment in the $g-$ genomic subgroup is to not expose more $g-$ patients than necessary when the interim data strongly suggest that the new treatment does not benefit this patient subset. Therefore, p_a^1 and p_{g+}^1 are essentially compared with an alpha equivalent to 0.0 for efficacy evaluation at the interim time. Consequently, at the study end, p_a^2 and/or p_{g+}^2 are compared with the respective appropriate alpha levels as if there were no interim analyses. The above alpha handling strategies can be discussed using the p-value combination framework, e.g. [14].

4. SIMULATION STUDY

The type I error probability and statistical power are compared between the two statistical methods under each study design type via Monte Carlo simulation. Without loss of generality, we set the sample size per treatment arm, $N/2 = 100$, 200, since the effect sizes can always be adjusted to fit other sample size scenarios. The targeted experimentwise type I error probability to control is set to a one-sided $\alpha = 0.025$ level. In the split-alpha method, equal split ($\alpha_1 = \alpha_2 = 0.0125$) and unequal split ($\alpha_1 = 0.02$, $\alpha_2 = 0.005$; $\alpha_1 = 0.005$, $\alpha_2 = 0.02$) are considered. The range of prevalence rate for the $g+$ subgroup is set to $f = 0.1$ (0.1) 0.9. The sample size in the genomic $g+$ subset is planned as $M = fN$. For the AD, the interim time considered is $t = 0.25$, 0.50, and 0.75. The futility criterion we choose is $b_t = \Phi^{-1}(0.02), \Phi^{-1}(0.15), \Phi^{-1}(0.50)$ corresponding to approximately -2, -1, 0 on the z-scale, respectively. The null hypotheses for $(\Delta, \Delta_{g+}, \Delta_{g-})$ considered include (0, 0, 0), (0, Δ_{g+}, Δ_{g-}), (Δ, 0, Δ_{g-}).

For the alternative hypotheses of specific interest are the scenarios where the presence of genomic biomarker acts as a predictive factor predictive of treatment effect but not disease response (Scenario A of Table I) or as a prognostic factor prognostic of disease response and prognostic of treatment effect (Scenario C of Table I). For a brief summary from our extensive simulation studies, we consider $(\Delta, \Delta_{g+}, \Delta_{g-})$ with the following configurations: (i) (+ , 0.4, 0) of Scenario A, (ii) (0, 0.4, −), and (iii) (0.2, 0.4, Δ_{g-}). The genomic biomarker is considered predictive of clinical benefit under configurations (i) and (ii). However, in configuration (iii), $\Delta = \frac{1}{2}\Delta_{g+}$, the genomic biomarker is considered prognostic for $f < 0.5$ where $\Delta_{g-} > 0$ (Scenario C), but, is considered predictive for $f > 0.5$ where $\Delta_{g-} < 0$ (a variation of Scenario A). Note that the configurations that do not address the predictive or prognostic utility of a genomic biomarker, e.g. (0.4, 0.4, 0.4), would be of no interest as the treatment effect is not different between $g+$ and $g-$ subsets, and can be addressed by the conventional trial design. For each design/method scenario, 1 000 000 data sets are simulated.

In the AD, the tests Z_N and Z_M are each compared with $Z_{0.0125}$ if the original plan is not changed. However, if only the genomic $g+$ subset is considered in the study end, the adaptive tests U_2 and T_2 are each compared with $Z_{0.0125}$, namely, there is no alpha reallocation. When the comparison includes the adaptive method of FS [9], the split-alpha method also includes $\alpha_a = 0.02$ and $\alpha_{g+} = 0.005$.

The inference for Δ using the adaptive split-alpha method or adaptive Hochberg method might not be straightforward due to a possible change in the final subset sample size $M'(> M)$; thus, the estimated prevalence rate of the $g+$ subgroup in the entire study due to such a change is no longer f. In the simulations, inference about Δ in the AD setting uses only patients from Stage I, a conservative estimate. This indicates that the actual power for testing H_{0a} will be larger than those obtained here if all patients in Stage II (either keeping the same fraction of $g+$ or recruiting all $g+$ patients due to an adaptive process) are included for power evaluation. In addition to the two individual power functions and the global power function for H_1 $(1 - \beta_{1or2})$, the conventional restricted power function that requires statistical significance be shown for $\Delta > 0$ and then for $\Delta_{g+} > 0$ $(1 - \beta_{1n2}$, labeled as 1n2) is also computed for comparison.

4.1. Simulation results

4.1.1. Type I error probability

4.1.1.1. Fixed study design. With the split-alpha method, when Δ in equation (3) and Δ_{g+} in equation (4) are set to 0, $\alpha_a + \alpha_{g+} = \alpha$. When Z_N and Z_M are positively correlated, which is the case here, Hochberg method strongly controls the experimentwise type I error rate.

4.1.1.2. Adaptive study design. When the futility analysis is performed at 25%, 50%, or 75% of the originally planned sample size using the futility cutoff of approximately -2, -1, and 0, the empirical experimental type I error probabilities under the complete null (0, 0, 0) are maintained at 0.025 level [only results for $t = 0.5$ and $b_t = \Phi^{-1}(0.15)$ are shown in Table III]. To address the question: whether the type I error probability is inflated for concluding $\Delta_{g+} > 0$, we investigated the error probabilities for the individual null hypotheses for (+, 0, 0.4) and (0, 0.4, −), where '+' is a positive value determined by $f\Delta_{g+} + (1-f)\Delta_{g-}$, and '−' is a negative value deter-

Table III. Empirical type I error probability for the adaptive designs compared.

				Hochberg*		Split-alpha†		Freidlin and Simon	
f	Δ	Δ_{g+}	Δ_{g-}	H_{0a}	H_{0g+}	H_{0a}	H_{0g+}	H_{0g+}	$1-\beta_{1n2}$
0.1	0	0	0	0.0136	0.0134	0.0125	0.0123	0.0124	0.0006
0.2	0	0	0	0.0141	0.0139	0.0125	0.0123	0.0124	0.0009
0.3	0	0	0	0.0145	0.0143	0.0125	0.0123	0.0124	0.0012
0.4	0	0	0	0.0150	0.0148	0.0125	0.0123	0.0124	0.0015
0.5	0	0	0	0.0154	0.0153	0.0124	0.0123	0.0124	0.0018
0.6	0	0	0	0.0159	0.0159	0.0124	0.0123	0.0124	0.0022
0.7	0	0	0	0.0164	0.0166	0.0123	0.0123	0.0124	0.0025
0.8	0	0	0	0.0170	0.0175	0.0122	0.0123	0.0124	0.0029
0.9	0	0	0	0.0181	0.0188	0.0121	0.0123	0.0124	0.0032
0.1	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0124
0.2	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0124
0.3	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0124
0.4	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0123
0.5	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0122
0.6	+	0	0.4	na	0.0248	na	0.0123	0.0124	0.0118
0.7	+	0	0.4	na	0.0247	na	0.0123	0.0124	0.011
0.8	+	0	0.4	na	0.0246	na	0.0123	0.0124	0.0093
0.9	+	0	0.4	na	0.0243	na	0.0123	0.0124	0.0066
0.1	0	0.4	−	0.0211	na	0.0125	na	na	0.0043
0.2	0	0.4	−	0.0238	na	0.0123	na	na	0.0082
0.3	0	0.4	−	0.0219	na	0.0114	na	na	0.0107
0.4	0	0.4	−	0.0185	na	0.0094	na	na	0.0119
0.5	0	0.4	−	0.0204	na	0.0097	na	na	0.0123
0.6	0	0.4	−	0.0243	na	0.0119	na	na	0.0124
0.7	0	0.4	−	0.0251	na	0.0124	na	na	0.0124
0.8	0	0.4	−	0.0250	na	0.0125	na	na	0.0125
0.9	0	0.4	−	0.0249	na	0.0125	na	na	0.0124

Note: $N/2 = 200$, $t = 0.5$, $\alpha = 0.0250$, $b_t = \Phi^{-1}(0.15)$; na: not applicable for calculating type I error rate; $1-\beta_{1n2}$ measures the probability of rejecting both H_{0a} and H_{0g+} based on Freidlin and Simon method (note that $1-\beta_{1n2}$ for adaptive Hochberg and adaptive split alpha methods are much smaller than 0.025 as expected and are all larger than that with the FS method); '+' is a positive value determined by $f\Delta_{g+} + (1-f)\Delta_{g-}$; and '−' is a negative value determined by $\Delta_{g-} = (\Delta - \Delta_{g+}f)/(1-f)$.

*The empirical experimentwise type I error rates using adaptive Hochberg procedures, in the order of f, are 0.0245, 0.0241, 0.0236, 0.0231, 0.0225, 0.0219, 0.0213, 0.0207, and 0.0203.

†Equal alpha split is used for Split-alpha method.

mined by $\Delta_{g-} = (\Delta - \Delta_{g+}f)/(1-f)$. For illustrative purpose, we present as a typical case the results for $N/2 = 200$, $t = 0.5$, and $b_t = \Phi^{-1}(0.15)$. From the simulation, the difference in the type I error probability between the unweighted test U_2 statistics and the weighted test T_2 statistics is in the fourth decimal point, that is, essentially there is no difference, and thus only the type I error probability of U_2 is reported. In sum, with the AD, both the split-alpha method and the Hochberg method can be seen to have a strong control of the experiment-wise type I error probability (Table III). Other scenarios also render the same conclusion.

4.1.2. *Power performance*

4.1.2.1. Fixed study design. Figure 2 shows the empirical power of concluding H_{1g+} as a function of sample size ratio f when the genomic biomarker is predictive of treatment effect (the results for $\Delta = 0, \Delta_{g+} = 0.4$ is very similar to Figure 2 and are thus not shown), and Figure 3 when the genomic biomarker is prognostic. In all cases, the Hochberg method appears slightly more powerful for predictive biomarker and more powerful for prognostic biomarker than the equal alpha-split method. When more (or less) alpha (α_{g+}) is allocated for testing H_{0g+}, the empirical power is higher (or lower) when compared with the Hochberg method. Under equal alpha split, the apparent power gain with the Hochberg method (e.g. 5% gain when $f = 0.1$) is when the genomic biomarker is a prognostic factor where f is less than 0.5. The power gain diminishes for $f > 0.5$, where the genomic biomarker becomes predictive (Figure 3).

For a predictive genomic biomarker, the empirical global power curve $1-\beta_{1or2}$ is slightly higher, but very similar to that of $1-\beta_{g+}$ plotted in Figure 2. When the genomic biomarker is prognostic, we note that $1-\beta_{1or2}$ is much higher than $1-\beta_{g+}$ for $f < 0.5$ plotted in Figure 3 (e.g. there is an absolute power increase of 27%, 16%, and 8% with the Hochberg method when $f = 0.1$, 0.2, and 0.3). The simulation results also indicated that the global power of the Hochberg method and that of the equal alpha-split method converge for all prevalence of $g+$. These results are not shown.

The power curves $1-\beta_{1n2}$ represent the widely used approach of testing H_{0g+} only if H_{0a} is rejected. This approach almost always has much lower power than $1-\beta_{1or2}$ in the FD approach (results not shown). We will show these power curves that may not be as extreme as the FD approach in the AD section below.

4.1.2.2. Adaptive study design. In all the AD scenarios considered, the empirical power is slightly higher (e.g. 1%) with the unweighted test

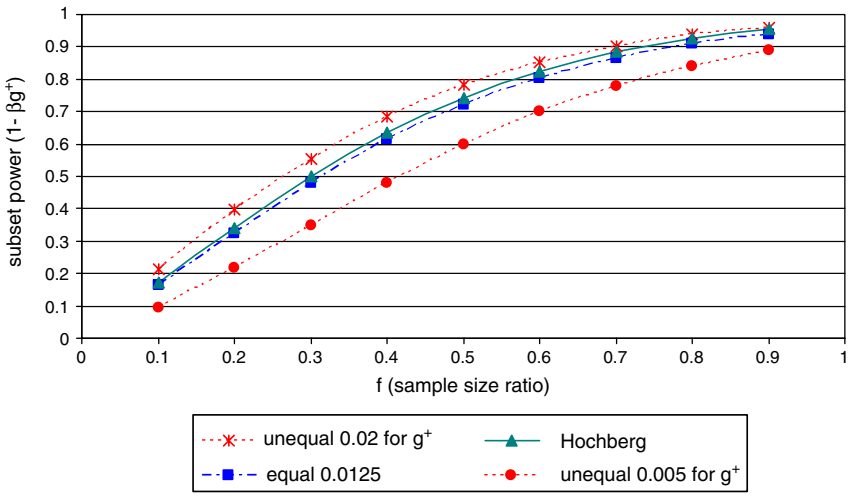

Figure 2. Power comparison for Δ_{g+} under non-adaptive design ($\Delta_{g+} = 0.4$, $\Delta_{g-} = 0$).

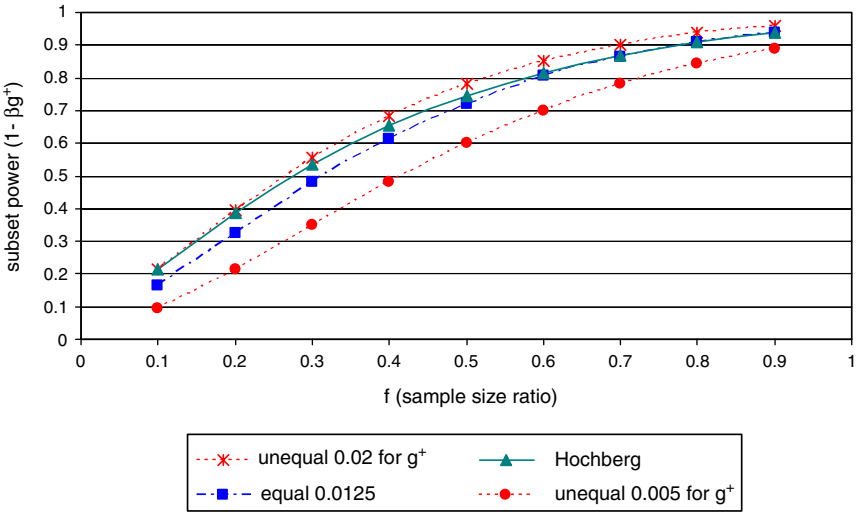

Figure 3. Power comparison for Δ_{g+} under non-adaptive design ($\Delta = 0.2$, $\Delta_{g+} = 0.4$).

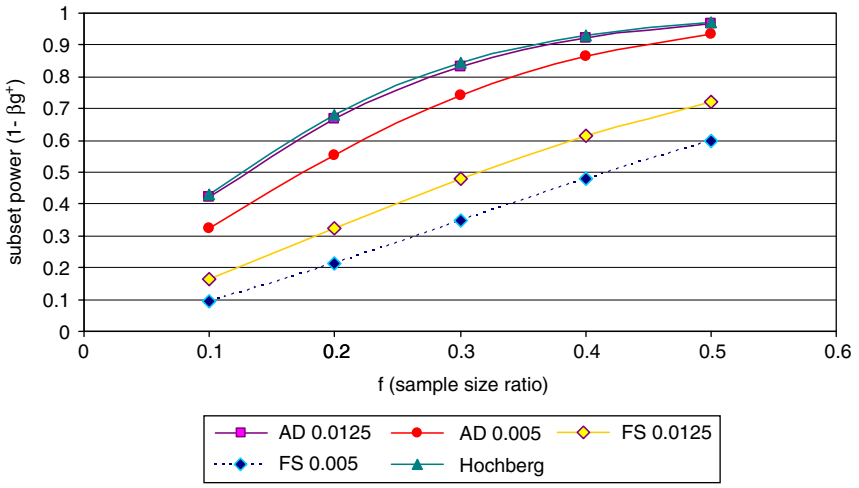

Figure 4. Power comparison for Δ_{g+} under adaptive design ($\Delta_{g+} = 0.4$, $\Delta_{g-} = 0$).

(U_2) than with the weighted test (T_2). Thus, in what follows, we choose U_2 for the AD approach to discuss the power performances.

Split-alpha method versus Hochberg method: For $\Delta_{g+} = 0.4$ and $\Delta_{g-} = 0$, Hochberg method gives similar power performance for $g+$ subset power $1 - \beta_{g+}$ as the equal alpha-split method (Figure 4). From Figure 5(a), Hochberg method is more powerful than the equal-split method (e.g. 9% $1 - \beta_{g+}$ power gain for $f = 0.1$) when the genomic biomarker is prognostic; most power gains occur when f is less than 0.4. The improvement in the global power $1 - \beta_{1or2}$ (in Figure 5(b)) over $1 - \beta_{g+}$ power (in Figure 5(a)) is not as large as that seen in the non-adaptive setting (e.g. absolute power increase of 9%, 8%, and 7% for $f = 0.1$,

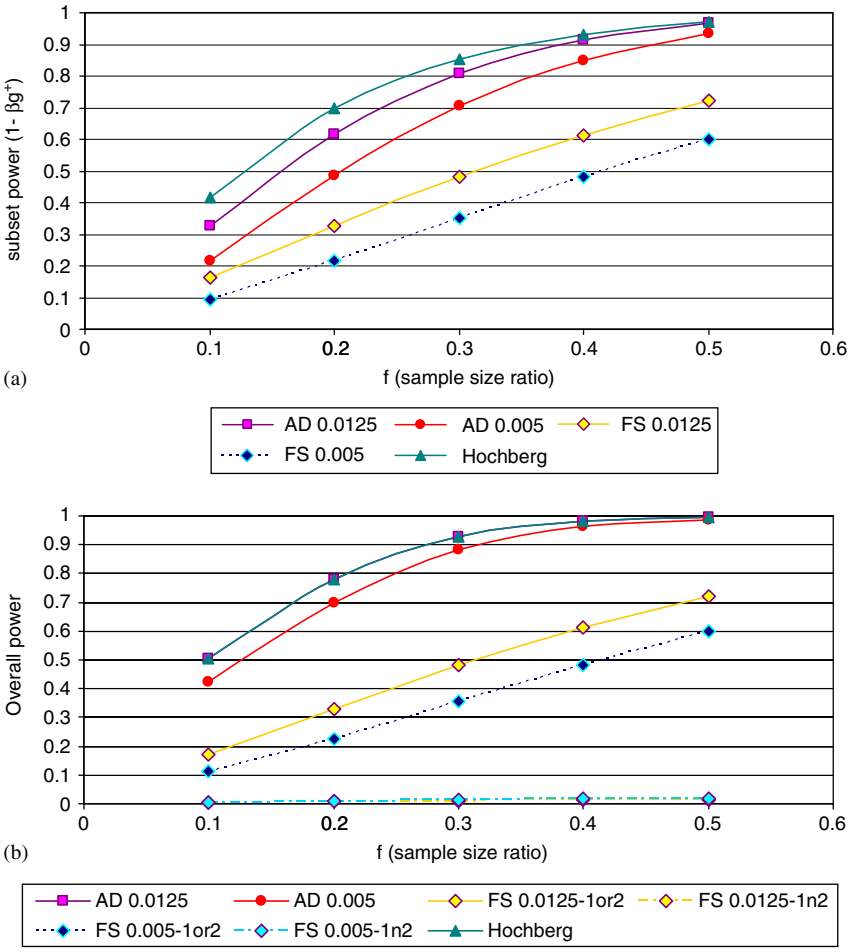

Figure 5. (a) Power comparison for Δ_{g+} under adaptive design ($\Delta = 0.2$, $\Delta_{g+} = 0.4$) and (b) overall power comparison under adaptive design ($\Delta = 0.2$, $\Delta_{g+} = 0.4$).

0.2, and 0.3). These results are not shown. As depicted in Figure 5(b), these two global power curves are superimposable.

Split-alpha method versus Freidlin–Simon method: The AD of FS [9] prospectively explores whether the treatment effect Δ_{g+} exists in the genomic $g+$ subset using gene expression profiling data in Stage I. The FS approach prospectively allocates α_a and α_{g+} using the split-alpha method at the design stage. The rationale for allocating a smaller α_{g+} is to prospectively test a pre-specified subhypothesis if, instead of spurious finding, a $g+$ genomic subset is identifiable in Stage I. The study sample size is planned based on the overall treatment effect in a two-arm comparison. In Stage II, two tests are performed. The null hypothesis of $\Delta = 0$ is tested at α_a, e.g. 0.02, based on N. However, the null hypothesis of $\Delta_{g+} = 0$ is tested at α_{g+} level, e.g. 0.005, using the $g+$ genomic patient subset in Stage II alone, i.e. only the $(1-t)M$ $g+$ patients. The sample size fraction f remains unchanged throughout.

Given that the composite hypothesis of interest is the same and pre-specification of alpha allocation to each individual hypothesis is required for multiplicity adjustment in the AD we consider and

Table IV. Impact on power for Δ_{g+} by interim time t and futility b_t criterion: Hochberg method.

H_1 scenario $(\Delta, \Delta_{g+}, \Delta_{g-})$	Interim time t	Futility criterion (b_t)		
		$\Phi^{-1}(0.02)$	$\Phi^{-1}(0.15)$	$\Phi^{-1}(0.50)$
(+, 0.4, 0.0)	0.25	0.5073	0.5636	0.7146
(+, 0.4, 0.0)	0.50	0.5057	0.5519	0.6763
(+, 0.4, 0.0)	0.75	0.5031	0.5311	0.6067
(0.0, 0.4, −)	0.25	0.5202	0.6454	0.8155
(0.0, 0.4, −)	0.50	0.5339	0.6569	0.7831
(0.0, 0.4, −)	0.75	0.5260	0.6071	0.6746
(0.2, 0.4, Δ_{g-})	0.25	0.5372	0.5623	0.6648
(0.2, 0.4, Δ_{g-})	0.50	0.5359	0.5502	0.6206
(0.2, 0.4, Δ_{g-})	0.75	0.5358	0.5423	0.5788

Note: The powers are obtained using the Hochberg method when $N/2 = 100$ and $f = 0.3$, thus, '+' = 0.12, '−' = −0.171, and $\Delta_{g-} = 0.114$.

the FS design approach, we compare the power performances focusing on the split-alpha method where the two methods use the same alpha allocation for testing H_{0a} and H_{0g+} under the AD setting. The AD with the alpha-split method clearly yields a higher power than the FS method. For instance, when $\alpha_a = \alpha_{g+} = 0.0125$, the largest power gains can be 34%, 45% 45%, 38% for $f = 0.1, 0.2, 0.3, 0.4$ in configuration (0, 0.4, −), followed by 27%, 36%, 34%, 32% in configuration (+, 0.4, 0), and, as shown in Figure 5(a), 16%, 29%, 33%, 30% in configuration (0.2, 0.4, Δ_{g-}), respectively.

This is expected since the sample size used in testing for the $g+$ subset is very likely to be much larger for the adaptive methods evaluated than for the FS method. For the same reason described above, the restricted power $1 - \beta_{1n2}$ for the FS approach if testing H_{0g+} is performed only after $\Delta = 0$ is rejected will be severely impacted due to the much smaller sample size used for testing H_{0g+} (Figure 5(b)). Note that the adaptation schemes differ between the considered AD and the FS approach, and thus these comparisons are to give some insights into the considered AD. The comparisons are by no means to imply that one method is more efficient than the other. In the FS design, the rationale for use of only the Stage II $g+$ patients to test H_{0g+} is to separate learning of a possible important (larger) treatment effect in the genomic subset in Stage I from confirming such an effect in Stage II, so as to not inflate the type I error probability of falsely concluding a treatment effect in either the overall patient population or the genomic subset or both; see the last two columns of Table III for the type I error probability.

Impact of interim time t or futility criterion b_t: The interim time t or the futility criterion b_t can have an impact on power as shown in Table IV. Using the Hochberg method to illustrate, when the genomic biomarker is predictive, i.e. $\Delta_{g+} = 0.4$ and $\Delta_{g-} \leqslant 0$, the power to detect a significant treatment effect Δ_{g+} greatly increases if the futility analysis is performed at information time $t = 0.25$ as compared with $t = 0.50$ or 0.75 when $f = 0.3$ for $b_t = \Phi^{-1}(0.15)$ and $b_t = \Phi^{-1}(0.5)$. There is essentially no power difference when $b_t = \Phi^{-1}(0.02)$. The power improvement can be larger using a relatively more aggressive futility criterion than using a conservative futility criterion. For instance, in the alternative state, (0.0, 0.4, −), when the futility analysis is performed at $t = 0.25$, the power for Δ_{g+} can be 30% higher if one considers discontinuing recruitment of $g-$ patients when the treatment effect estimate for $g-$ is negative versus when it is approximately two standard deviations below zero.

4.2. Power comparison between the non-adaptive and adaptive designs

The power comparison is made between the non-adaptive (i.e. fixed) (FD) and the AD) for the split-alpha method (Figure 6) and the Hochberg

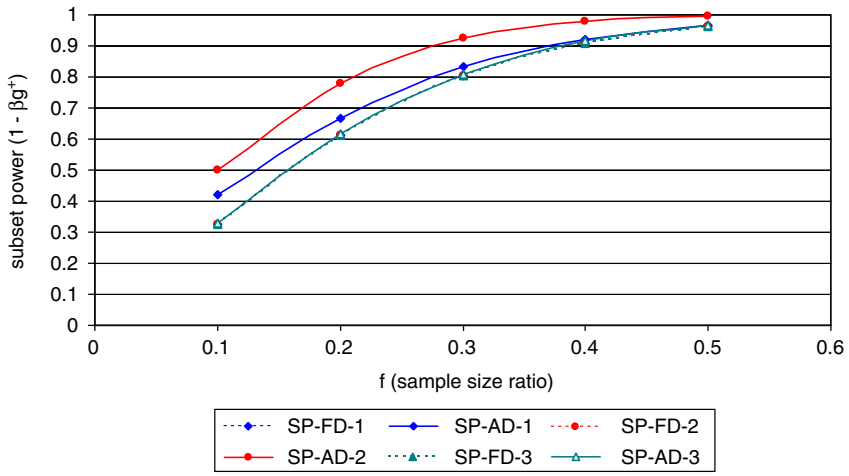

Figure 6. Power comparison for Δ_{g+} with equal-split method (1 = (Δ, 0.4, 0); 2 = (0, 0.4, $\Delta_{g-} < 0$); 3 = (0.2, 0.4, Δ_{g-})).

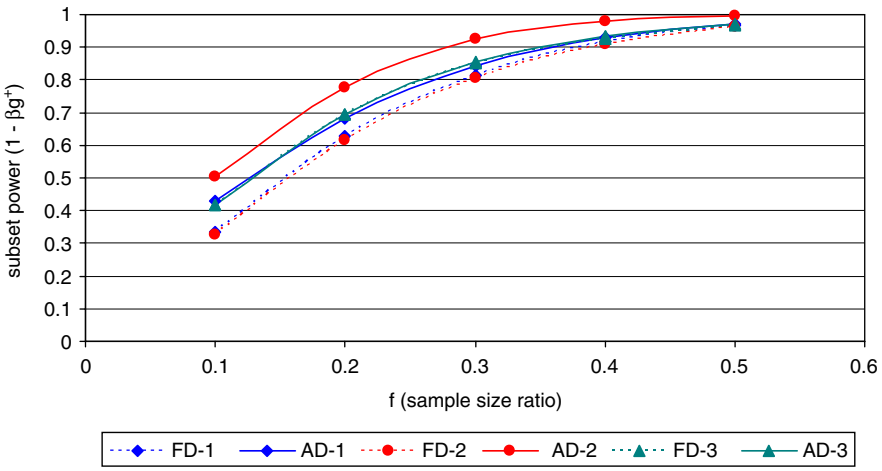

Figure 7. Power comparison for Δ_{g+} with Hochberg method (1 = (Δ, 0.4, 0); 2 = (0, 0.4, $\Delta_{g-} < 0$); 3 = (0.2, 0.4, Δ_{g-})).

method (Figure 7). Using either analysis method, the AD offers (much) more power than the non-adaptive approach for detecting Δ_{g+}. The most significant power gains occur when the genomic biomarker is predictive. As shown in Figure 6 for the split-alpha method, when $f = 0.1$, 0.2, and 0.3, the power gains are 17%, 16%, 12% for (0, 0.4, $-$) and 10%, 6%, 3% for ($+$, 0.4, 0). Note that for the FD, the three power curves depicted in dashed line are very similar resulting in overlapping curves. For Hochberg method (Figure 7), similar magnitudes of power gains with adaptive approach are observed. These findings suggest that the adaptive approach outperforms the non-adaptive approach when the genomic biomarker can predict which patient subset benefits from therapeutic intervention, e.g. $(\Delta, \Delta_{g+}, \Delta_{g-}) = (0.0, 0.4, -)$. The $1 - \beta_{g+}$ power advantage of the adaptive approach decreases and is almost absent when the overall treatment effect Δ reaches half that of the effect in the $g+$ subset, e.g. $(\Delta, \Delta_{g+}, \Delta_{g-}) = (0.2, 0.4, \Delta_{g-})$ and

$f < 0.5$, where the effect Δ_{g-} is greater than 0, but less than Δ_{g+}.

5. SOME CRITICAL ISSUES IN DESIGN AND ANALYSIS

5.1. Factors that impact misclassification associated with genomic biomarker

The AD as considered above assumes that the patients can be accurately classified into either $g+$ or $g-$ class based on the genomic composite biomarker that is believed to be predictive of treatment effect. Namely, there is no error in the classification of patients. In practice, the presence or absence of a genomic characteristic is generally determined by a genomic diagnostic assay. There are several issues with the genomic diagnostic assay. For instance, the reliability and reproducibility of the genomic samples and standardization of the diagnostic assay are often lacking; therefore, it is very difficult to ensure the consistency in assay testing across local and central diagnostic laboratories. Genomic samples are often acquired through voluntary consents. Convenient samples so obtained violate the intent-to-treat principle. Although the sensitivity and specificity of the genomic diagnostic assay may not be well established, the prospective drug trial assesses the predictive values of the genomic composite biomarker for its clinical utility based on the $g+$ versus $g-$ classification status.

When an imperfect genomic classification of a diagnostic assay is incorporated into the design planning, Maitournam and Simon [15] extensively discuss the impact on the efficiency of targeted clinical trials in terms of number of screened patients to meet planned number of subjects in the $g+$ and $g-$ subgroups for randomization. The authors modeled assay specificity and whether a treatment effect is expected for the less responsive group of patients separately, viz. whether Δ_{g-} is also greater than zero. Using the two scenarios: $(\Delta_{g-} = 0, \Delta_{g+} > 0)$ and $(\Delta_{g-} = \frac{1}{2}\Delta_{g+})$, and considering sensitivity $= 0.6,\ 0.8,\ 1.0$ and specificity $= 0.6,\ 0.8,\ 1.0$ and as a function of proportion of $g+$, the authors highlighted that the efficiency of targeted clinical trials is primarily assessed by assay specificity for the required number of randomized patients and assay sensitivity for the required number of screened patients.

The diagnostic assay test for classifying the patient into either $g+$ or $g-$, which is used in turn to assess whether the drug effect is attributed to all patients studied or only to the genomic subset, can be co-developed [16] with the evaluation of the drug. Often, drug and diagnostic assay test are developed separately and the emphasis within the drug trial is on the clinical utility measured by predictive value, whereas the emphasis within the diagnostic trial is on the performance characteristics of the diagnostic assay test.

5.2. Genomic predictive effect size Δ_{g+}

Although there is an ultimate interest to also evaluate the treatment effect in the pre-specified genomic subset, the magnitude of Δ_{g+} is often unclear, given the limited anecdotal evidence in the available scientific literature or prior studies. Suppose the sample size is planned based on the assumed overall treatment effect. The $g+$ subset sample size observed in the overall trial serves to capture the achievable effect size in the subset. The sample size ratio, $f = M/N$, can also be expressed as

$$f = \left[\frac{Z_{\alpha_{g+}} + Z_{\beta_{g+}}}{Z_{a_a} + Z_{\beta_a}}\right]^2 \left[\frac{\Delta}{\Delta_{g+}}\right]^2$$

We have

$$\frac{\Delta_{g+}}{\Delta} = \left[\frac{Z_{\alpha_{g+}} + Z_{\beta_{g+}}}{Z_{\alpha_a} + Z_{\beta_a}}\right] \Big/ \sqrt{f}$$

Given α_a and f, one can generate subset power versus effect size ratio Δ_{g+}/Δ at different levels of type II error (β_a) of the overall trial. The subset power is

$$1 - \beta_{g+} = \Phi\left(\sqrt{f}\left(\frac{\Delta_{g+}}{\Delta}\right)(Z_{\alpha_a} + Z_{\beta_a}) - Z_{\alpha_{g+}}\right)$$

The prevalence rate f of $g+$ is often empirically estimated from the given sample assuming balanced randomization without enriching the $g+$

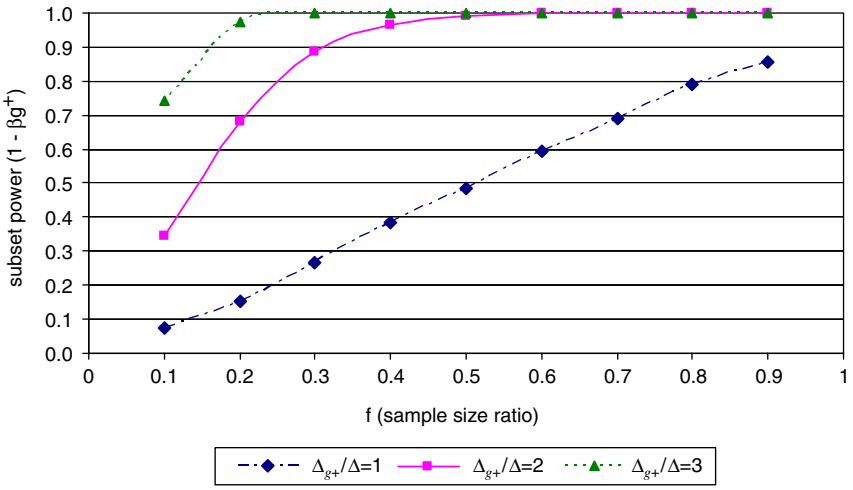

Figure 8. Subset power for $\Delta_{g+}/\Delta = 1$, 2, or 3, given $\alpha = 0.025$, $\alpha_a = 0.02$, $1 - \beta_a = 0.90$.

subgroup. This is generally achievable when the overall trial is large and the prevalence rate is not too small. It is also reasonable to use a conservative estimate of $g+$ prevalence. Of note, if $f = 0$ is used, the two null hypotheses, H_{0a} and H_{0g+}, are assumed disjoint. In this case, the effect size ratio Δ_{g+}/Δ does not depend on the sample size ratio, f.

Figure 8 depicts the subset power $1 - \beta_{g+}$ for $\alpha_{g+} = 0.005$. A smaller α_{g+} in the unequal split-alpha method aims at controlling a spurious and not well-understood subgroup finding. Here, the $g+$ subset power is computed assuming a one-sided overall experimentwise error rate of 0.025 and the power for the overall trial is 0.90. As shown in Figure 8 for the subset, at least a sample size of 30% of the overall trial is required to detect a $g+$ subset treatment effect that is twice the overall effect with 90% power, and at least 17% of the total sample size is needed for a three-fold increase over the overall treatment effect for the $g+$ subset.

5.3. Treatment by genomic interaction effect

The subgroup analysis or interaction analysis after trial completion is a popular exercise for the purpose of exploring the uniformity of any treatment effects found overall. Generally, such statistical tests for interaction are underpowered and if not planned for in advance, are likely to be so, e.g. [17]. A directional *post hoc* assessment of the consistency of a treatment effect across subgroups [18] is exploratory checking for prognostic value of the subgroup indicator.

In the setting of randomized controlled trial, it is generally anticipated that when the genomic factor is predictive, qualitative interaction between the treatment and the genomic factor will exist, which can be written as the treatment effect difference between the $g+$ and $g-$ subgroups, $\Delta_{g+} - \Delta_{g-}$. The extent of interaction may be explored. Some may argue that a preliminary test of interaction is always required before testing further the effect in the $g+$ subset.

To fully address whether the treatment effect is shown in the patient subset possessing specific genomic biomarker or in all patients studied, the testing plan that uses the interaction test result to decide on whether to test H_{0g-} and H_{0g+} separately or test H_{0a} ignoring the genomic biomarker status does not guarantee the strong control of the experimentwise type I error. This is because irrespective of rejecting or not rejecting the interaction null hypothesis of no interaction,

depending on the significance level applied to the interaction test, there is always the possibility that there is a significant treatment effect in one of the two genomic subgroups.

5.4. Clinical implication of the alternative hypotheses

If H_{0a} is rejected, but H_{0g+} is not, the treatment effect is often asserted for all patients studied and subgroup effects are for checking the consistency of the effect. When only H_{0g+} is rejected, the treatment effect appears to be only in the genomic subset and the genomic biomarker may be predictive of therapeutic response (e.g. Scenario A of Table I). When both H_{0a} and H_{0g+} are rejected, the genomic biomarker is likely to be prognostic of therapeutic effect if there is no qualitative interaction between the therapy and the genomic biomarker (e.g. Scenario C of Table I). However, if a qualitative interaction exists and depending on the prevalence of the genomic biomarker, the genomic biomarker may be considered predictive of treatment effect.

5.5. The futility analysis

Conceivably, one can build in a futility analysis based on all-patient population at an interim time of the trial. This approach can allow early termination of the entire study without giving an opportunity to detect substantial treatment effect in the $g+$ patient subset (e.g. Scenario A in Table I). Our justification of an interim futility analysis in the $g-$ subset has three folds. First, from a regulatory perspective, there is a concern about what the safety margin is or the therapeutic effect in the $g-$ genomic subgroup that eventually is not studied when an observed benefit in the selected $g+$ patient subgroup is declared. The proposed futility analysis addresses this concern. Secondly, if safety is not an issue, the interim futility analysis has small probability to exclude the $g-$ group if the utility of the genomic biomarker is prognostic of therapeutic effect. That is, unless the effect in $g-$ is too small to be of clinical utility, the chance of excluding the $g-$ genomic subgroup is (very) small, thus, allowing the study to continue with full-patient population. Therefore, the clinical composite hypothesis at issue can still be tested at the end of Stage II. With this adaptive approach, the opportunity to test whether the genomic biomarker is predictive or prognostic at the study end remains. Thirdly, if there is little therapeutic benefit, i.e. if safety is an issue and/or treatment effect in $g-$ is too small, the empirical data provide evidence for sound critical assessments to exclude the $g-$ genomic subgroup. In this case, the proposed two-stage AD allows for the flexibility to early terminate further accrual of genomic negative subgroup ($g-$) minimizing exposure of an ineffective/unsafe treatment to those who are unlikely to benefit from it.

The Stage I futility analysis builds in a sample size reallocation rule without increasing the total sample size based on the observed interim treatment effect estimate. Rather, the remaining sample size from the total sample size originally planned is reallocated to recruit only those patients who are likely to have reasonable benefit/risk balance in the second stage, that is a type of enrichment [19] that is applied only to the second stage. Indeed, from the perspectives of public health and cost effectiveness, if the new treatment only benefits the $g+$ genomic patient subset, the justification for a better genomic approach than the usual clinical approach should be based on a treatment effect size preferably bigger than those postulated based on clinical characteristics alone as discussed in Section 5.2.

6. CONCLUDING REMARKS

Valid statistical inference approaches require prospectively planned genomic subsets and valid alpha allocation to control for the multiple hypotheses [1,2,9,10]. In principle, instead of the conventional approach that H_{0g+} is tested only if H_{0a} is rejected, one can build in a multiplicity adjustment rule such that H_{0a} is tested only if H_{0g+} is rejected. There are several other alpha allocation

schemes. One may also account for the correlation between the hypotheses [12,20] as described in Section 3. To gain greater flexibility, AD may be employed, e.g. [9]. Most importantly, the choice of statistical method for the composite hypothesis and the choice of design type should be made prior to the start of the study.

The principle underlying the proposed futility analyses for the patient exclusion decision can be applied to evaluating surrogate endpoint. The problem is relatively more complicated than that associated with the use of clinical endpoint alone. The success of targeting genomically prone patient subset based on the surrogate endpoint heavily relies on the sample size, the prevalence of $g+$, and the association between the treatment effect on the surrogate endpoint and that on the clinical endpoint. If the predictive surrogacy for the clinical endpoint is well established and the genomic biomarker used for patient selection has the same predictivity on the surrogate endpoint and the clinical endpoint, the futility analysis based on the established surrogate endpoint might properly expedite exclusion to minimize the unnecessary exposure of patients to the wearisome new therapeutics. The decision, in turn, provides the opportunities to develop other treatments that may benefit those excluded patients. Such a pathway of patient selection has good potential to bring the clinical trial research one step closer to personalized medicine.

Selection of a genomically classified patient subset mid-trial may introduce a bias to the study when the interim unblinded results are used for selection. The notion of selecting a genomic patient subgroup that is very responsive to treatment is different from the notion of excluding a patient subgroup that does not appear to be responsive to treatment in midstream of an ongoing trial. The former may induce relatively more serious trial conduct biases than the latter, as ineffectiveness in the excluded subgroup is not clearly established. The latter preserves the trial objective of showing superiority, except that the objective may be shifted from H_0 to H_{0g+} in the study end; thus, it should be less likely to jeopardize the trial integrity.

To conclude, we have considered the comparative properties of an adaptive approach that incorporates the use of genomic data and a FD in the late phase randomized controlled clinical trial to evaluate the utility of a completely defined genomic composite biomarker, whether it is derived from a high-dimensional gene/protein expression studies or from a single protein biomarker.

ACKNOWLEDGEMENTS

This research work was supported by the RSR funds #02-06, #04-06, #05-2, and #05-14 awarded by the Center for Drug Evaluation and Research, U.S. Food and Drug Administration.

REFERENCES

1. Wang SJ. Genomic biomarker derived therapeutic effect in pharmacogenomics clinical trials: a biostatistics view of personalized medicine. *Taiwan Clinical Trials* 2006; **4**:57–66.
2. Simon R, Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. *The Pharmacogenomics Journal* 2006; **6**:166–173.
3. Baselga J. Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials. *Oncology* 2001; **61**(S2):14–21.
4. van'tVeer LJ, Dai H, Vijver MJVD, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**:530–536.
5. Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, Harrison JA, Bines J, Mook S, Decker N, Ravdin P, Therasse P, Rutgers E, van't Veer LJ, Piccart M. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nature Clinical Practice Oncology* 2006; **3**(10): 540–550.
6. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; **23**(9):2020–2027.
7. Bauer P, Brannath W, Posch M. Flexible two stage designs: an overview. *Methods of Information in Medicine* 2001; **40**:117–121.
8. Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* 2006; **48**:565–573.

9. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 2005; **11**(21):7872–7878.
10. Moye LA, Deswal A. Trials within trials: confirmatory subgroup analyses in controlled clinical experiments. *Controlled Clinical Trial* 2001; **22**: 605–619.
11. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
12. Hsu JC. *Multiple comparisons: theory and methods*. Chapman & Hall: London, 1996.
13. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857.
14. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**: 1833–1848.
15. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 2005; **24**:329–339.
16. U.S. FDA Draft Drug Diagnostic Co-Development Preliminary Concept Paper. http://www.fda.gov/cder/genomics/pharmacoconceptfn.pdf (8 April 2005).
17. Lachenbruch PA. A note on sample size computation for testing interactions. *Statistics in Medicine* 1988; **7**:467–469.
18. Cui L, Hung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics* 2002; **12**(3):241–252.
19. Temple RJ. Enrichment designs: efficiency in development of cancer treatments. *Journal of Clinical Oncology* 2005; **23**(22):4838–4839.
20. Wang SJ, Hung HMJ. Trials in trials: alpha allocation strategy and sub-trial planning. *Proceedings of the American Statistical Association*, *Biopharmaceutical Section* [*CD-ROM*]. American Statistical Association: Alexandria, VA, 2005.