Statistics and Its Interface Volume 4 (2011) 431–441

Optimal two-stage sequential robust design for gene-intervention studies

Lihan K. Yan*, Aiyi Liu, Zhaohai Li and Gang Zheng

Gene-intervention studies investigate the responsiveness to therapies according to individuals' genetic profiles. We propose a two-stage sequential design for these studies and investigate the cost of the sample size versus the statistical power. In a typical sequential design, a single normally distributed test statistic is used. For a genetic study, the robust test is used because of the uncertainty of the underlying genetic model (e.g. the recessive, additive or dominant models). The robust test statistic that we consider in the two-stage sequential design is the maximum of three correlated normally distributed statistics, each which is optimal under the corresponding genetic model. We study various factors that affect minimizing the average sample number (ASN) or maximizing the power of a gene-intervention study under the two-stage sequential design and make recommendations for the optimal solutions under different scenarios.

AMS 2000 subject classifications: 92D10.
Keywords and phrases: Optimal two-stage design, MAX, Group sequential, Gene-intervention.

1. INTRODUCTION

1.1 Background

There has been a great interest in gearing therapies according to individuals' genetic profiles. As a result, the link between a specific genetic marker and responsiveness to a therapy is often investigated in a clinical study, referred to as a gene-intervention study. The design of a gene-intervention study is analogous to a case-control study design except that the study is conducted prospectively. For a diallelic marker with alleles a and A, the three genotypes are denoted by aa, Aa and AA. Assume A is the allele of interest, an analysis can be performed to examine the association between responsiveness to therapy and the A allele. The Cochran-Armitage trend tests (CATTs) are often used to detect associations (Armitage, 1955; Cochran, 1954; Sasieni, 1997). Depending on the underlying genetic model, i.e., the recessive, additive (multiplicative) or dominant models, three CATTs are available, each of which is optimal for the corresponding genetic model (Slager and Schaid, 2001; Freidlin et al., 2002; Zheng et al., 2003). The disadvantage of using a single CATT is that, when the true genetic model is unknown, misspecification of a genetic model can result in substantial power loss for the CATT. Freidlin et al. (2002) proposed a robust test statistic which is the maximum of the three CATTs over the recessive, additive and dominant models, denoted by MAX3. Other robust tests, including the constrained likelihood ratio test (Wang and Sheffield, 2005) and maximum of likelihood ratio tests (Gonzalez et al., 2008), have also been studied in the literature. Note that under the null hypothesis H_0 which is no genetic effect for the candidate gene, MAX3 no longer follows $N(0,1)$ asymptotically. Its asymptotic distribution can be obtained by simulation (Zheng and Chen, 2005) or its tail probability can be approximated (Li et al., 2008).

Because of timing, cost and ethical reasons, group sequential designs have been widely used in clinical trials. Such designs are often used to monitor accumulated data at regular sample recruiting intervals and allow early stopping under the null and/or the alternative hypotheses (Jennison and Turnbull, 2000) while the pre-specified Type I errors are controlled. Statistical methods for group sequential analysis have been extensively applied in the situation where a single test statistic is involved, which has a known distribution, such as normal or t-distributions. Applications of group sequential analysis to genetic studies for the purpose of improving efficiency have been also reported in the literature. For example, Konig et al. (2001, 2003) demonstrated the sample size savings in linkage and association studies by utilizing a design with stopping boundaries based on the mean test and the transmission disequilibrium test (TDT), respectively. Konig and Ziegler (2003) extended the applications of sequential analysis to case-control studies using a normally distributed test statistic. The MAX statistic is more robust to the unknown genetic models than normally distributed statistics.

Two-stage optimal sequential designs have been studied in medical studies and clinical trials. Simon (1989) proposed a two-stage optimal design that minimizes the expected sample size for a Phase II clinical trial. Shu et al. (2007) studied the optimal designs for sequential evaluation of a medical diagnostic test. Although two-stage designs have also been studied in case-control genetic studies, they are not typical two-stage sequential studies and often arise from large association studies for marker selection. For example, Satagopan and Elston (2003) proposed an optimized

*Corresponding author.

two-stage design for case-control studies by genotyping all markers using a portion of samples at the first stage, and selected the most promising markers to be genotyped using the remaining samples at the second stage. Their approach has been further extended in different situations with different optimality criteria (Satagopan et al., 2004; Muller et al., 2007). See also Elston et al. (2008) for multi-stage designs for genetic case-control studies.

Recently, Nguyen et al. (2009) proposed a two-stage optimal design that incorporates the genotyping cost and statistical power in genome-wide association studies (GWAS). A robust maximal statistic was considered in the design. The design targets optimal power or cost in a setting where a portion of the samples are genotyped for the whole marker set in the first stage and then the rest of the samples are genotyped for the promising markers in the second stage. The authors found in one example that the allocation fraction of 0.48 was optimal and that under a fixed power, increasing the sample size by 13% could reduced the cost by 34.5%. However, this design does not apply for a single marker study in which the genotyping cost is not an issue but optimization of the sample size or power are more focused for varying design parameters of type I error spending and interim looks. Below are two of the real world studies that motivated us to consider a robust group sequential design in a single-marker gene-intervenetion study.

1.2 Motivational examples

The first example is a study entitled "Long Acting Beta Agonist Response by GEnotype" (LARGE) (Wechsler et al., 2009) which was conducted by the Asthma Clinical Research Network (ACRN) and sponsored by the Division of Lung Diseases (DLD) of the National Heart, Lung and Blood Institute. The objective of the study was to examine the effects of regularly scheduled long-acting beta-agonist in a group of asthmatic patients harboring the B16-Arg/Arg genotype and in a separate group of baseline matched patients harboring the B16-Gly/Gly genotype at the 16th aminoacid position of the β_2-adrenergic receptor (a candidate-gene). The study showed a significant genotype difference in responsiveness to methacholine between the two genotypes and suggested a further investigation.

Another example is "The CardioGene Study". This is a study of restenosis in bare metal stents (BMS) for the treatment of coronary artery disease (Ganesh et al., 2004). The objective of the study is to identify the genetic profile of patients at risk for in-stent restenosis (ISR). The study endpoint is the presence/absence of ISR at 6 months. Again this is also a genomic clinical trial in which individuals are enrolled prospectively. In such a study design, interim analysis can be performed after enrolling 50% of the patients. Prospective risk stratification would allow for the rational selection of specialized treatments against the development of ISR. Because the mode of inheritance of ISR is not clear, a robust test is considered for the sequential study design with interim analysis.

Both examples above require a study design where all of the study subjects are provided the target therapy and the responsiveness to therapy was evaluated subsquently. The study objective is to assess the association between a genetic marker and responsiveness to therapy and an intervention may take place in the future based on the results. In order to save time and cost, such a study may be assessed sequentially in two (or more) stages to potentially allow stopping during the interim of the study. Because of the unknown nature of the genetic mode for the marker, a robust test may be desirable.

1.3 Overview

In this paper, we consider a classical sequential design for a single genetic marker using the robust statistic MAX3 in a gene-intervention study setting, and study how to allocate samples (or information) in order to achieve minimum ASN or maximum power. We investigate the operating characteristics of the two-stage sequential design under a variety of parameters. We take into account not only the design specific parameters, such as the allocation fraction of samples for the first stage, an error spending function and alternative differences, but also genetic-related parameters including the allele frequency and underlying genetic model. We present the power and the ASN under each scenario and make recommendations for the optimal two-stage sequential design for gene-intervention studies.

The subsequent sections of this paper are arranged as follows: Section 2 provides the detailed description of a two-stage sequential design in a gene-intervention study (Section 2.1), the test statistics used for hypothesis testing (Section 2.2), the statistical methods for obtaining the critical values in a group sequential design (Section 2.3), and the parameters considered for optimization (Section 2.4). Section 3 presents simulations results and finally Section 4 provides comments and discussions.

2. METHODS

2.1 Two-stage group sequential design in a gene-intervention study

A typical gene-intervention study investigates association between the responsiveness to therapy and a candidate marker (noted as Allele A). Suppose the study is conducted in two stages. At Stage 1, a portion of the subjects are genotyped and identified to have one of the three genotypes (AA, Aa, aa, where the Allele a represents any other allele). All subjects are treated with the therapy and at the end the results are tabulated in a 2 by 3 table (Table 1). If the results indicate a statistically significant advantage or disadvantage for the subjects with the candicate allele, the study may be stopped at Stage 1. Otherwise, the study

Table 1. Gene-intervention study results

	aa	Aa	AA	Total
Responders	$r0$	$r1$	$r2$	r
Non-responders	$s0$	$s1$	$s2$	s
Total	$n0$	$n1$	$n2$	n

proceeds to Stage 2 and continues recruiting the rest of the subjects and the cumulative data are analyzed again at the end of the study. Significant association results may result in potential intervention, i.e., gearing the therapy to include (or exclude) the subjects in order to maximize the benefit from the targeted therapy.

2.2 Cochran-Armitage trend tests (CATT) and MAX3

As in population-based case-control genetic studies, CATT can also be a powerful statistical test to detect the association between an allele and response to therapy in a gene-intervention study setting. Assume A is the allele of interest and also the one with the minor allele frequency MAF. In Table 1 the responders $(r_i, i = 0, 1, 2)$ and non-responders $(s_i, i = 0, 1, 2)$ can be tabulated for genotypes aa, Aa, and AA, respectively. These counts are further determined by the corresponding genotype frequencies $(p_i, i = 0, 1, 2$ among responders and $q_i, i = 0, 1, 2$ among non-responders), i.e., $(r_0, r_1, r_2) \sim Multinomial(r; p_0, p_1, p_2)$ and $(s_0, s_1, s_2) \sim Multinomial(s; q_0, q_1, q_2)$. The null hypothesis is $H_0 : p_i = q_i, i = 0, 1, 2$. The CATT proposed by Sasieni (1997) can be formulated as:

$$(1) \quad Z_\theta = \frac{n^{1/2} \sum_{i=0}^{2} x_i (s r_i - r s_i)}{\left\{ rs[n \sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2] \right\}^{1/2}},$$

where $x = (x_0, x_1, x_2) = (0, \theta, 1)$ is the set of scores for the genotypes (aa, Aa, AA). Z_θ is asymptotically normally distributed. The optimal sets of scores for the dominant, additive (multiplicative), and recessive models are $x = (0, 1, 1)$, $(0, \frac{1}{2}, 1)$ and $(0, 0, 1)$, respectively. For a given θ, the CATT in (1) follows asymptotically $N(0, 1)$ under H_0.

The association between the candidate allele and responsiveness to therapy can be expressed in terms of genotype relative risks (GRRs): $\gamma_i = f_i/f_0$ for being a responder and $\delta_i = (1 - f_i)/(1 - f_0)$ for being a non-responder, $i = 1, 2$, where the penetrance f_i is the probability of being a responder given the genotype with i copies of the allele of interest. Then $p_i = \gamma_i g_i / \sum \gamma_i g_i$, and $q_i = \delta_i g_i / \sum \delta_i g_i$, where g_i, $i = 0, 1, 2$, are the genotype frequencies in the population for the three genotypes $G_0 = aa$, $G_1 = aA$ and $G_2 = AA$. The relationship between γ_1 and γ_2 under the dominant, additive and recessive models follows $\gamma_1 = \gamma_2$, $\gamma_1 = (1 + \gamma_2)/2$, and $\gamma_1 = 1$, respectively.

The MAX3 statistic is given by MAX3 $= \max(|Z_0|, |Z_{1/2}|, |Z_1|)$ or MAX3 $= \max(Z_0, Z_{1/2}, Z_1)$, depending on whether or not the risk allele is known. When the risk allele is unknown, the one-sided MAX3 can still be used at the $\alpha/2$ level with each allele being treated as the risk allele. The asymptotic null correlations $(\mathrm{Corr}_{H_0}(Z_0, Z_{1/2}), \mathrm{Corr}_{H_0}(Z_{1/2}, Z_1), \mathrm{Corr}_{H_0}(Z_1, Z_0))$ among the three CATTs are used in applying sequential analysis using MAX3. These correlations were given in Freidlin et al. (2002):

$$\mathrm{Corr}_{H_0}(Z_0, Z_{1/2}) = \frac{p_2(p_1 + 2p_0)}{\{p_2(1 - p_2)\}^{1/2} \{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2\}^{1/2}},$$

$$\mathrm{Corr}_{H_0}(Z_{1/2}, Z_1) = \frac{p_0(p_1 + 2p_2)}{\{(p_0(1 - p_0)\}^{1/2} \{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2)\}^{1/2}},$$

$$\mathrm{Corr}_{H_0}(Z_1, Z_0) = \frac{p_0 p_2}{\{p_0(1 - p_0)\}^{1/2} \{p_2(1 - p_2)\}^{1/2}}.$$

2.3 Applying MAX3 in sequential design

Critical values for statistical testing in a group sequential design are usually determined by the distribution of the test statistic and a prespecified alpha spending function (ASF). Conventional alpha spending methods, such as the Pocock method (Pocock, 1977) and the O'Brien-Fleming method (O'Brien and Fleming, 1979), can be readily applied to obtain the critical values for a single test statistic with a known form of distribution, however, the test statistic MAX3 does not have an explicit form of distribution.

Using an approximation of the tail probability for maximal-type statistics studied by Efron (1997), Yan et al. (2008) proposed tools for obtaining the critical values when MAX3 was used as the test statistic in a two-stage group sequential design for a family-based genetic study. In the family-based association study, MAX3 has the same form as that for the case-control study. However, the test statistics Z_θ are different, so are their correlations. The goals of Yan et al. (2008) were to study how to find critical values using Efron's approaches and a specific alpha spending function so that the overall Type I error is controlled due to using MAX3 sequentially. It should be noted that in Yan et al. (2008), the design parameters were fixed so that the thresholds for the two stages can be determined to control Type I errors. In the next section, we will examine the performance (in terms of minimum ASN or maximum statistical power) of two-stage designs with changes of these design parameters.

Among several approximations studied by Efron (1997), Yan et al. (2008) found the two-point formula is simple to use and controls the Type I error reasonably well. We consider MAX3 $= \max(|Z_0|, |Z_{1/2}|, |Z_1|)$ with a target $\alpha/2$ level. The two-point formula can be written as:

$$P(\text{MAX3} > c) \le \bar{\Phi}(c) + \sum_{j=1/2,1} \left\{ \int_{-\infty}^{c} \bar{\Phi}\left[\frac{c - \rho_j t}{(1-\rho_j^2)^{1/2}} \right] \phi(t) dt \right\},$$

where ρ_j is the asymptotic null correlation between Z_j and $Z_{j-1/2}$ ($j = 1/2, 1$); ϕ and Φ are the density and distribution functions of $N(0,1)$, respectively; and $\bar{\Phi} = 1 - \Phi$.

Consider a two-stage sequential design with overall level α. Assume the sample sizes in stage i is N_i with level α_i, $i = 1, 2$, and $\alpha_1 + \alpha_2 = \alpha$. The sample allocation fraction is denoted by $\pi = N_1/N_2$ (the samples in stage 2 include all samples in stage 1). Given α and π, the levels α_1 and α_2 are determined by a prespecified alpha spending function (ASF). Three commonly used ASFs are considered here (Lan and Demets, 1983; Betensky, 1998): i) $\text{ASF}_1(t) = 2\{1-\Phi(z_{\alpha/2}t^{-1/2})\}$, ii) $\text{ASF}_2(t) = \alpha \log\{1+(e-1)t\}$, and iii) $\text{ASF}_3(t) = \alpha t$, where $t = \pi$ is the information fraction. The first and second functions are equivalent to the discretized O'Brien-Fleming and Pocock types of spending functions, respectively. The third function is a uniform spending function with regard to information fraction.

2.4 Optimal two-stage design for gene-intervention studies

In a two-stage sequential design, there is a trade-off between minimizing the sample size and maximizing the statistical power. Both are important when designing a group sequential study and decisions are often made to balance the trade-off between the two factors. Suppose one interim analysis is conducted at Stage 1 and one final analysis is done at Stage 2. Define the stopping rule as that the study stops if MAX3 is statistically significant at Stage 1, i.e., the study terminates at the first stage if $\text{MAX3}_1 > c_1$ and continues to the second stage if otherwise. Denote the allocation fraction $\pi = N_1/N_2$ as before, where N_1 and N_2 are the cumulative sample sizes at the two stages. The ASN for the two-stage sequential design can be calculated as:

$$\begin{aligned} \text{ASN} &= N_1 \text{P}(\text{Study stops at Stage 1}) \\ &\quad + N_2 \text{P}(\text{Study continues}) \\ &= N_2 - N_2(1-\pi) P(\text{MAX3}_1 > c_1). \end{aligned}$$

Accordingly, the Type I error and power can be respectively written as:

$$\begin{aligned} \alpha &= P(\text{MAX3}_1 > c_1 | H_0) \\ &\quad + P(\text{MAX3}_1 \le c_1 \text{ and } \text{MAX3}_2 > c_2 | H_0), \end{aligned}$$

$$\begin{aligned} \text{Power} &= P(\text{MAX3}_1 > c_1 | H_1) \\ &\quad + P(\text{MAX3}_1 \le c_1 \text{ and } \text{MAX3}_2 > c_2 | H_1), \end{aligned}$$

where MAX3_1 and MAX3_2 are the test statistics, and c_1 and c_2 are the critical values obtained using the method described in Yan et al. (2008).

The goals of our optimal designs are to minimize the ASN or to achieve the maximum power while the Type I error is controlled, and find the ranges of parameter values to achieve the minimum ASN or maximum power. Note that a parameter considered in the design is a function of other parameters in the design. Using the notation in Section 2.1, given values of the MAF (and hence the genotype frequency g_i, $i = 0, 1, 2$ under Hardy-Weinberg equilibrium (HWE)), the power of MAX3 for a given sample size is a function of the GRRs and the underlying genetic model. In a two-stage design, the ASN is determined by the power of detecting the difference at each stage, which is determined not only by the above parameters but also the critical values, which are further determined by the alpha spending function and the allocation fraction. Here we study the optimal designs such that either the minimum ASN or the maximum power are achieved. First, we present the optimal designs under a fixed target sample size and study the impact on ASN and power for different combinations of the specified values of the allele frequency, genetic model (dominant, additive, or recessive), the GRRs (γ_1, γ_2), and ASF. Then, from a different perspective, we provide recommended sample sizes for optimal designs under the specified values of those parameters when the power is fixed. The results are also compared with those for a single-stage design.

3. SIMULATION STUDIES AND EXAMPLES

In this section, we first present results when parameter values change. Then we apply the results using values from real examples in the simulation. Each simulation was replicated 10,000 times. The critical values for MAX3 were obtained given the values of the allocation fraction π from 0.1 to 0.5 with an increment of 0.05 and an ASF as described before. When the target sample size was fixed, the simulation was generated, and the ASN and power were calculated. When the power was fixed, the simulation was repeated until the sample size that achieved the specified power was found. The ASN and power were then calculated based on the simulated datasets. For all simulations, all tests were two-sided at an overall alpha level of 0.05.

3.1 Fix the sample size

Given the target sample size of 2,000 subjects: 1,000 responders and 1,000 non-responders assuming the response rate is 50%, we simulated results to obtain the ASNs and the powers under the alternative hypotheses for each of the three ASFs. The results are presented in Tables 2 through 4 for three different ASFs. In each table, results are presented under different allele frequencies MAF = 0.1, 0.3, or 0.5, a genetic model, dominant (DOM), additive (ADD), or recessive (REC), and GRRs (γ_1, γ_2). The ASNs and powers and the corresponding allocation fractions π are presented for two scenarios: 1) when the minimum ASN is achieved (Columns 5 through 7); and 2) when the maximum power is

Table 2. Optimal allocation fractions π to achieve the minimum ASN or the maximum power given the target sample size for a single-stage design ($N = 1{,}000$ per group): ASF $= \mathrm{ASF}_1$. The alternatives are specified by the GRRs (γ_1, γ_2) under three genetic models with different allele frequencies MAF

				Min ASN is achieved			Max Power is achieved			Power for a single
MAF	Model	γ_1	γ_2	ASN	Power	π	Power	ASN	π	stage study
0.1	DOM	1.5	1.5	786	0.978	0.5	0.978	1000	0.15	0.975
		2	2	468	1.000	0.4	1.000	468	0.4	1.000
	ADD	1.25	1.5	964	0.598	0.5	0.608	1000	0.25	0.601
		1.5	2	745	0.991	0.5	0.991	916	0.35	0.989
	REC	1	1.5	998	0.146	0.5	0.156	1000	0.15	0.151
		1	2	988	0.430	0.5	0.454	1000	0.15	0.451
0.3	DOM	1.5	1.5	678	0.997	0.5	0.998	723	0.45	0.997
		2	2	426	1.000	0.4	1.000	426	0.4	1.000
	ADD	1.25	1.5	878	0.895	0.5	0.911	973	0.35	0.902
		1.5	2	561	1.000	0.5	1.000	561	0.5	1.000
	REC	1	1.5	918	0.827	0.5	0.831	985	0.35	0.823
		1	2	600	1.000	0.5	1.000	600	0.5	1.000
0.5	DOM	1.5	1.5	802	0.968	0.5	0.973	1000	0.2	0.972
		2	2	528	1.000	0.45	1.000	528	0.45	1.000
	ADD	1.25	1.5	863	0.919	0.5	0.919	968	0.35	0.915
		1.5	2	573	1.000	0.5	1.000	586	0.45	1.000
	REC	1	1.5	736	0.993	0.5	0.993	736	0.5	0.990
		1	2	439	1.000	0.4	1.000	439	0.4	1.000

Table 3. Optimal allocation fractions π to achieve the minimum ASN or the maximum power given the target sample size for a single-stage design ($N = 1{,}000$ per group): ASF $= \mathrm{ASF}_2$. The alternatives are specified by the GRRs (γ_1, γ_2) under three genetic models with different allele frequencies MAF

				Min ASN is achieved			Max Power is achieved			Power for a single
MAF	Model	γ_1	γ_2	ASN	Power	π	Power	ASN	π	stage study
0.1	DOM	1.5	1.5	635	0.962	0.45	0.973	943	0.10	0.977
		2	2	340	1.000	0.30	1.000	340	0.30	1.000
	ADD	1.25	1.5	869	0.501	0.50	0.581	985	0.10	0.605
		1.5	2	597	0.981	0.40	0.987	933	0.10	0.988
	REC	1	1.5	968	0.119	0.45	0.137	997	0.10	0.153
		1	2	914	0.351	0.50	0.425	995	0.10	0.452
0.3	DOM	1.5	1.5	547	0.994	0.45	0.998	727	0.20	0.998
		2	2	291	1.000	0.25	1.000	291	0.25	1.000
	ADD	1.25	1.5	722	0.852	0.50	0.891	958	0.10	0.903
		1.5	2	431	1.000	0.35	1.000	440	0.30	1.000
	REC	1	1.5	783	0.745	0.50	0.810	970	0.10	0.827
		1	2	469	0.999	0.35	1.000	755	0.15	1.000
0.5	DOM	1.5	1.5	654	0.945	0.50	0.966	948	0.10	0.970
		2	2	397	1.000	0.30	1.000	397	0.30	1.000
	ADD	1.25	1.5	711	0.870	0.50	0.907	955	0.10	0.918
		1.5	2	441	1.000	0.35	1.000	480	0.45	1.000
	REC	1	1.5	592	0.982	0.45	0.989	926	0.10	0.990
		1	2	306	1.000	0.25	1.000	306	0.25	1.000

Table 4. Optimal allocation fractions π to achieve the minimum ASN or the maximum power given the target sample size for a single-stage design ($N = 1{,}000$ per group): ASF $= \mathrm{ASF}_3$. The alternatives are specified by the GRRs (γ_1, γ_2) under three genetic models with different allele frequencies MAF

				Min ASN is achieved			Max Power is achieved			Power for a single
MAF	Model	γ_1	γ_2	ASN	Power	π	Power	ASN	π	stage study
0.1	DOM	1.5	1.5	652	0.964	0.45	0.977	960	0.10	0.979
		2	2	354	1.000	0.30	1.000	354	0.30	1.000
	ADD	1.25	1.5	882	0.526	0.50	0.584	979	0.15	0.602
		1.5	2	613	0.983	0.45	0.988	831	0.20	0.988
	REC	1	1.5	977	0.115	0.50	0.143	996	0.15	0.154
		1	2	928	0.355	0.50	0.422	997	0.10	0.450
0.3	DOM	1.5	1.5	566	0.996	0.45	0.997	928	0.10	0.998
		2	2	300	1.000	0.25	1.000	300	0.25	1.000
	ADD	1.25	1.5	743	0.863	0.50	0.898	969	0.10	0.905
		1.5	2	444	1.000	0.35	1.000	444	0.35	1.000
	REC	1	1.5	794	0.778	0.50	0.822	982	0.10	0.832
		1	2	487	1.000	0.40	1.000	799	0.15	1.000
0.5	DOM	1.5	1.5	661	0.953	0.50	0.967	959	0.10	0.971
		2	2	411	1.000	0.35	1.000	411	0.35	1.000
	ADD	1.25	1.5	729	0.876	0.50	0.914	936	0.15	0.918
		1.5	2	460	1.000	0.35	1.000	460	0.35	1.000
	REC	1	1.5	609	0.986	0.50	0.990	744	0.25	0.991
		1	2	321	1.000	0.25	1.000	321	0.25	1.000

achieved (Columns 8 through 10). The last column presents the power of a single-stage study with 1,000 responders and 1,000 non-responders.

While the optimal designs that achieve the minimum ASNs suggest the allocation fractions of between 0.3 and 0.5 for Stage 1, depending on the ASFs and powers, the allocation fractions for the designs where the maximum powers are achieved are lower, ranging from 0.1 to 0.3. The gain of the power is relatively small for having extra samples, especially when the study power is high. For example, in Table 3, under the additive model ($\gamma_1 = 1.25$ and $\gamma_2 = 1.5$) where MAF=0.3, the minimum ASN is 722 subjects per group when the allocation fraction is 0.5 with the corresponding power of 85%. The ASN rises to 958 subjects per group when the allocation fraction is 0.1, the power is maximized and increased to 89%, only 4% more than the power achieved earlier. The power for a single-stage study with a sample size of 1,000 per group is 90%.

For a given ASF, higher allele freqency (MAF) results in lower ASN and higher power. For example, for ASF $=$ ASF_2, under the additive model ($\gamma_1 = 1.25$ and $\gamma_2 = 1.5$), the minimum ASN when MAF=0.1 is 869 with a power of 50% whereas the correponding minimum ASNs and powers are 722 and 85% when $p = 0.3$, and 711 and 87% when MAF=0.5.

Among the three genetic models, when the GRR for genotype AA (γ_2) is fixed, the test has the highest power under the dominant model and has the lowest power under the recessive model. Focusing on the same section of results where ASF $= \mathrm{ASF}_2$, MAF $= 0.3$ and $\gamma_2 = 1.5$ in Table 3, the minimum ASN is 547 with a power of 99% under the dominant model, whereas it is 722 with 85% power under the additive model, and 783 with 75% power under the recessive model.

Finally, across different ASFs, the allocation fractions are consistently suggested to be around 0.5 to achieve the minimum ASN when the alpha spending is conservative at the first stage, e.g. ASF $= \mathrm{ASF}_1$ or the O'Brien-Fleming type. When the allowance for Type I error increases, the allocation fractions that achieve the minimum ASN go slightly lower (ranging from 0.25 to 0.5). The Pocock type of ASF (ASF_2) appears to result in the smallest ASNs with the powers similar to those in the corresponding setting for the other ASFs.

3.2 Fix the statistical power

For two-stage sequential designs, when the power is given at, say 80%, the required sample size to achieve this given power under a certain alternative can be obtained along with the allocation fraction. Under similar simulation procedures, we present in Tables 5 through 7 the planned sample sizes (N), as well as the ASNs that achieve the power 80% under different scenarios. Columns 5 through 8 show results when the ASN is the minimum while Columns 9 through 12 show the results when the ASN is the maximum as a comparison, reflecting the less conservative nature of the designs in early stopping.

When the allocation ratio is set around half ($\pi = 0.5$), the probability of stopping at Stage 1 is the highest and the ASN reaches the minimum. On the other hand, when the allocation ratio is set low (between 0.1 and 0.15), the required sample size is relatively smaller because while the probability of stopping at the first stage is small, the chance of mistakenly stopping (Type I error) is small as well. The ASNs, however, are about 7% higher than those in the case where the allocation fraction is 0.5.

The results also confirm the conservativeness among the three ASFs. When the O'Brien-Fleming type of spending ($\text{ASF} = \text{ASF}_1$) is utilized, the required sample sizes are the smallest but the ASNs are maintained to be close to the targeted sample size because the probability of stopping at Stage 1 is smaller than those using the other ASFs.

In summary, for a two-stage sequential study, although the target sample size is generally required to be larger than what it is for a single-stage study, the ASN from a sequential study is smaller than that in a single study if a moderate risk exists. Figure 1 presents a simultaneous view of ASN and power varying by ASF over different allocation fractions for $p = 0.3$. The results show an earlier look, interim before 50% of the samples, and more aggressive type I error spending can result in substantial sample size saving with minimal power loss.

4. DISCUSSION

In this paper, we studied the operating characteristics of a group sequential gene-intervention study of the association between a single candidate marker and responsiveness to therapy using the robust statistic MAX3. The results have shown advantages of having a two-stage design on savings on average sample size while maintaining the power at a slightly reduced rate. Our results indicated that the typical allocation fraction of half often balances the trade-off between the sample sizes and powers. The choice of alpha spending function can impact the sample size. Overly conservative alpha spending, such as that of the O'Brien-Fleming type, helps little in sample size saving in a two-stage design. On the other hand, when the underlying difference is large, e.g., for mild genotype relative risk (e.g., $\gamma = 2.0$ for Genotype AA), the optimal design can be achieved by an earlier look (e.g., allocation fraction = 0.3) and more aggressive type I error spending in the first stage (e.g. the Pocock type of spending function). Such a design can result in a reduction of more than half for sample size compared to a single stage study. In the framework where a genetic marker is studied in either a retrospective or a prospective fashion, it is our opinion that alpha spending during interim analyses does not need to be as conservative as it is in a classical clinical trial where treatment effect is evaluated. Therefore, we recommend using the Pocock spending method in genetic studies rather than the conservative ones such as the O'Brien-Fleming spending method. Moreover, since the magnitude of the association between a genetic marker and the study endpoint, e.g. response to therapy or disease status, are frequently rather moderate, it is also beneficial for sample size savings to schedule an interim look time at 30% of the target samples.

As there have been increasing interests in the area of personalized medicines and gene-intervention therapies, our

Table 5. Sample size and optimal allocation fraction to achieve the power 80%: ASF $= \text{ASF}_1$

				Min ASN is achieved				Max ASN is resulted			
MAF	Model	γ_1	γ_2	π	N	ASN	P(stopping at Stage 1)	π	N	ASN	P(stopping at Stage 1)
0.1	DOM	1.5	1.5	0.5	523	481	0.158	0.3	523	521	0.006
		2	2	0.5	165	152	0.154	0.1	164	164	0.000
	ADD	1.25	1.5	0.5	1572	1446	0.161	0.2	1547	1547	0.000
		1.5	2	0.5	451	417	0.153	0.25	450	449	0.000
	REC	1	1.5	0.5	7659	7074	0.153	0.25	7581	7573	0.002
		1	2	0.5	2186	2039	0.135	0.3	2190	2187	0.002
0.3	DOM	1.5	1.5	0.5	377	348	0.153	0.1	368	368	0.000
		2	2	0.5	133	122	0.155	0.1	132	132	0.000
	ADD	1.25	1.5	0.5	747	690	0.153	0.25	745	744	0.002
		1.5	2	0.5	237	217	0.166	0.2	240	240	0.000
	REC	1	1.5	0.5	947	872	0.159	0.25	947	947	0.001
		1	2	0.5	284	263	0.151	0.15	289	289	0.000
0.5	DOM	1.5	1.5	0.5	552	510	0.152	0.25	557	556	0.001
		2	2	0.5	210	194	0.150	0.15	209	209	0.000
	ADD	1.25	1.5	0.5	714	656	0.163	0.15	708	708	0.000
		1.5	2	0.5	250	230	0.154	0.2	253	253	0.000
	REC	1	1.5	0.5	445	410	0.157	0.1	440	440	0.000
		1	2	0.5	142	131	0.153	0.1	143	143	0.000

Table 6. Sample size and optimal allocation fraction to achieve the power 80%: ASF = ASF_2

				Min ASN is achieved				Max ASN is resulted			
MAF	Model	γ_1	γ_2	π	N	ASN	P(stopping at Stage 1)	π	N	ASN	P(stopping at Stage 1)
0.1	DOM	1.5	1.5	0.4	596	469	0.354	0.1	541	527	0.027
		2	2	0.45	194	148	0.428	0.1	169	166	0.023
	ADD	1.25	1.5	0.5	1845	1385	0.499	0.1	1615	1574	0.028
		1.5	2	0.5	540	410	0.483	0.1	465	454	0.027
	REC	1	1.5	0.5	8830	6712	0.480	0.1	7903	7718	0.026
		1	2	0.45	2566	1967	0.425	0.1	2254	2229	0.013
0.3	DOM	1.5	1.5	0.5	440	330	0.500	0.15	401	377	0.071
		2	2	0.5	157	118	0.488	0.1	139	135	0.026
	ADD	1.25	1.5	0.45	869	659	0.439	0.1	781	756	0.035
		1.5	2	0.5	287	215	0.505	0.1	249	243	0.030
	REC	1	1.5	0.5	1123	844	0.497	0.1	986	959	0.030
		1	2	0.5	338	256	0.482	0.15	303	288	0.058
0.5	DOM	1.5	1.5	0.5	654	495	0.489	0.1	577	559	0.033
		2	2	0.4	235	185	0.354	0.1	216	211	0.023
	ADD	1.25	1.5	0.45	827	630	0.435	0.1	740	719	0.032
		1.5	2	0.5	296	221	0.508	0.1	258	251	0.030
	REC	1	1.5	0.45	518	393	0.437	0.1	459	447	0.030
		1	2	0.5	172	129	0.503	0.1	149	145	0.027

Table 7. Sample size and optimal allocation fraction to achieve the power 80%: ASF = ASF_3

				Min ASN is achieved				Max ASN is resulted			
MAF	Model	γ_1	γ_2	π	N	ASN	P(stopping at Stage 1)	π	N	ASN	P(stopping at Stage 1)
0.1	DOM	1.5	1.5	0.5	586	460	0.430	0.1	525	516	0.020
		2	2	0.45	184	146	0.370	0.15	173	168	0.035
	ADD	1.25	1.5	0.45	1737	1382	0.371	0.1	1601	1571	0.021
		1.5	2	0.5	508	400	0.425	0.15	467	451	0.039
	REC	1	1.5	0.5	8512	6676	0.431	0.1	7698	7569	0.019
		1	2	0.45	2471	1986	0.357	0.1	2235	2219	0.008
0.3	DOM	1.5	1.5	0.5	422	328	0.444	0.1	381	374	0.020
		2	2	0.5	150	117	0.439	0.15	138	132	0.045
	ADD	1.25	1.5	0.5	850	661	0.445	0.1	752	736	0.024
		1.5	2	0.5	272	211	0.449	0.1	249	244	0.021
	REC	1	1.5	0.45	1045	832	0.371	0.1	962	946	0.019
		1	2	0.5	322	254	0.425	0.1	291	288	0.013
0.5	DOM	1.5	1.5	0.45	618	492	0.373	0.1	565	554	0.023
		2	2	0.45	235	186	0.379	0.1	212	208	0.020
	ADD	1.25	1.5	0.5	810	627	0.454	0.1	724	706	0.027
		1.5	2	0.45	279	220	0.387	0.1	255	250	0.019
	REC	1	1.5	0.5	496	387	0.438	0.1	450	440	0.023
		1	2	0.5	162	126	0.439	0.1	146	143	0.019

analysis and results may provide some guidelines for design and analysis of future genomic clinical trials. Through this research, one may be about to utilize the findings to build a clinical prediction model for future personalized medicine. Some of our results are similar to those in a GWAS setting in Nguyen et al. (2009) as the authors also used the robust statistic. For example, both papers agreed on a possible allocation fraction of less than 0.5, e.g., 0.30 for some additive models. Our paper, however, provided more in-depth investigations on type I error spending and cost of sample size and power. While simulations provide the closest results in real world settings, we can also potentially apply and evaluate the accuracy of the general optimization functions used in Nguyen et al. (2009).

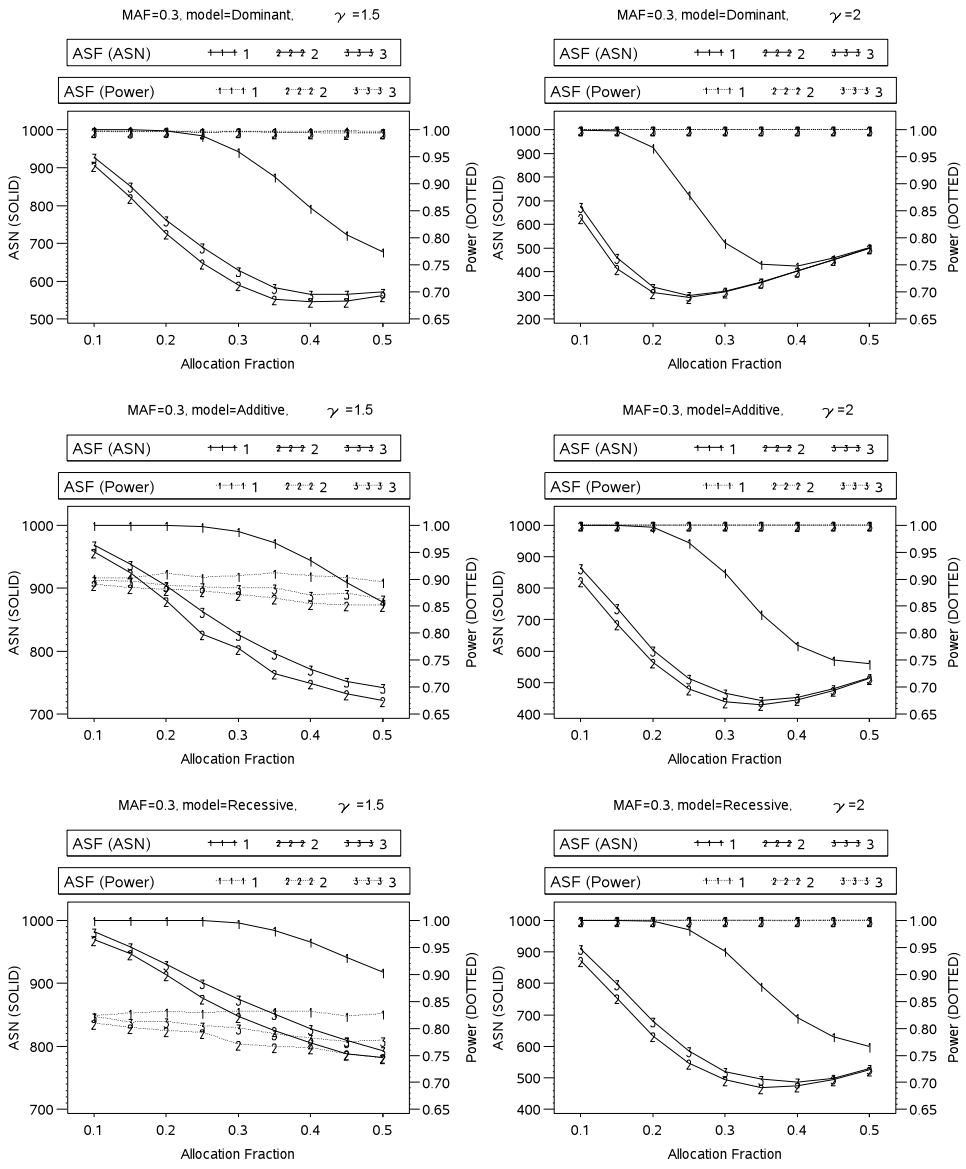

Figure 1. ASNs and power by ASN and genotype relative risk (γ).

We proposed a solution to ease the cost of enrolling individuals through a two-stage group sequential design. This design can naturally be extended to multiple-stage when multiple interim analyses are used. Also, the idea of having a two-stage sequential design under strong alternatives that allows stopping for significance in order to reduce average sample size can be extended to the opposite direction of the stopping rule, i.e., the study stops for futility or no association. Our simulation results also infer savings on sample size for such a design although careful design on the allocation of Type II error (β) needs to be considered. Furthermore, the study can be designed to allow stopping for both significance and non-significance, where the benefit may reach the most in terms of sample size and power. Our preliminary results of the impact factors on sample sizes can be a starting point for further research for optimal choices in the genetic study setting. Our computer programs which are used to determine critical values, and simulate sample sizes and powers under different pre-specified values of the design parameters can also be potentially used in practice for designing any multi-stage group sequential study of the association between a genetic marker and responsiveness to therapy.

Note that in this article the calculation of critical values for MAX3 is an approximation. Although the approximation method has been proved through simulation to work well for a two-stage study, it can be overly conservative if more stages are planned and therefore affects the choice of required sample sizes. Further investigation may be needed to compare different alpha spending functions in terms of

their impact on the designs of the study given the distribution of MAX3 under different alternatives with different values of allele frequency and genetic models.

Throughout, the optimal sample sizes are determined based on pre-specified differences in genotype frequencies between responders and non responders. If apart from the true values, these pre-specified frequencies may substantially decrease the power of the study. When the observed genotype frequencies are smaller than the pre-specified ones, but also of interest to the investigators, it may be desirable that the sample size be increased so that the study is adequately powered to detect smaller yet meaningful differences in genotype frequencies. To this end, the adaptive strategies developed in the clinical trial settings can be adopted. To elaborate a bit more, we conduct the first interim analysis with possible stopping when a pre-specified proportion of the initial sample size is reached. Data from the interim analysis are then used to estimate the genotype frequencies among both responders and non-responders. Based on the observed genotype frequencies the optimal final sample size is updated, and the final test statistic can be defined as a weighted average of the test statistics from the two stages; see, e.g. Bauer and Kohne (1994), Cui, Hung and Wang (1999), and Jennison and Turnbull (2003).

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments which strengthened the presentation of this paper. We would also like to thank Ms. Heather Liu for her editorial help. Research of A. Liu is supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health. The opinions expressed in the article are those of the authors, not necessarily of the National Institutes of Health, nor the Food and Drug Administration.

Received 13 July 2010

REFERENCES

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11** 375–386.

Betensky, R. A. (1998). Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **54** 1061–1071.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50** 1029–1041.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* **10** 417–451. MR0067428

Cui, L., Hung, H. M. J. and Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55** 321–324.

Efron, B. (1997). The length heuristic for a simultaneous hypothesis tests. *Biometrika* **84** 143–147. MR1450198

Elston, R. C., Lin, D. and Zheng, G. (2007). Multistage sampling for genetic studies. *Annu. Rev. Genomics Hum. Genet.* **8** 327–342.

Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum. Hered.* **53** 146–152. (Erratum 2009 69, 220.)

Ganesh, S. K., Skelding, K. A., Mehta, L., O'Neill, K., Joo, J., Zheng, G., Goldstein, J., Simari, R., Billings, E., Geller, N. L., Holmes, D., O'Neill, W. W. and Nabel, E. G. (2004). Rationale and study design of the CardioGene Study: genomics of in-stent restenosis. *Pharmacogenomics* **5** 952–1004.

Gonzalez, J. R., Carrasco, J. L., Dudbridge, F., Armengol, L., Estivill, X. and Moreno, V. (2008). Maximizing association statistics over genetic models. *Genet. Epidemiol.* **32** 246–254.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL. MR1710781

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22** 971–993.

Konig, I. R., Schafer, H., Müller, H. and Ziegler, A. (2001). Optimized group sequential study designs for tests of genetic linkage and association in complex diseases. *Am. J. Hum. Genet.* **69** 590–600.

Konig, I. R., Schafer, H., Ziegler, A. and Muller, H. (2003). Reducing sample sizes in genome scans: Group sequential study designs with futility stops. *Genet. Epidemiol.* **25** 339–349.

Konig, I. R. and Ziegler, A. (2003). Group sequential study designs in genetic epidemiological case-control studies. *Hum. Hered.* **56** 63–72.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663. MR0725380

Li, Q., Zheng, G., Li, Z. and Yu, K. (2008). Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann. Hum. Genet.* **72** 397–406.

Li, Q., Yu, K., Li, Z. and Zheng, G. (2008). MAX-rank: A simple and robust genome-wide scan for case-control association studies. *Hum. Genet.* **123** 617–623.

Müller, H. H., Pahl, R. and Schäfer, H. (2007). Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genet. Epidemiol.* **31** 844–852.

Nguyen, T. T., Pahl, R. and Schäfer, H. (2009). Optimal robust two-stage designs for genome-wide association studies. *Ann. Hum. Genet.* **73** 638–651.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25** 149–157.

Satagopan, J. M., Venkatraman, E. S. and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60** 589–597. MR2089433

Slager, S. L. and Schaid, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum. Hered.* **52** 149–153.

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

Shu, Y., Liu, A. and Li, Z. (2007). Sequential evaluation of a medical diagnostic test with binary outcomes. *Stat Med.* **26** 4416–4427. MR2410051

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics* **53** 1253–1261. MR1614374

Tarone, R. E. and Gart, J. J. (1980). On the robustness of combined tests for trends in proportions. *J. Am. Stat. Assoc.* **75** 110–116.

Wang, K. and Sheffield, V. C. (2005). A constrained-likelihood approach to marker-trait association studies. *Am. J. Hum. Genet.* **77** 768–780.

Wechsler, M. E. et al. (2009). Effect of β_2-adrenergic receptor polymorphism on response to longacting 2 agonist in asthma (LARGE trial): A genotype-stratified, randomised, placebo-controlled, crossover trial. *The Lancet* **374** 1754–1764.

Yan, L. K., Zheng, G. and Li, Z. (2008). Two-stage group sequential robust tests in family-based association studies: controlling type I error. *Ann. Hum. Genet.* **72** 557–565.

Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical J.* **45**(3) 335–348. MR1973305

Zheng, G. and Chen, Z. (2005). Comparison of maximum statistics for hypothesis testing when a nuisance parameter is only present under the alternative. *Biometrics* **61** 254–258. MR2135868

Lihan K. Yan
Office of Biostatistics and Epidemiology
Center for Biologics Evaluation and Research
The Food and Drug Administration
Rockville, MD
USA
E-mail address: lihan.yan@fda.hhs.gov

Aiyi Liu
Biostatistics and Bioinformatics Branch
Eunice Kennedy Shriver
National Institute of Child Health and Human Development
Rockville, MD
USA
E-mail address: liua@mail.nih.gov

Zhaohai Li
Department of Statistics
The George Washington University
Washington, DC
USA
E-mail address: zli@gwu.edu

Gang Zheng
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD
USA
E-mail address: zhengg@nhlbi.nih.gov