Optimal multistage designs—a general framework for efficient genome-wide association studies

ROMAN PAHL, HELMUT SCHÄFER, HANS-HELGE MÜLLER*

Institute of Medical Biometry and Epidemiology, Philipps University, Marburg, Germany
muellerh@med.uni-marburg.de

SUMMARY

Genome-wide association studies (GWAS) have become increasingly affordable but they are still costly. Therefore, cost saving 2-stage designs were proposed in the literature. The restriction to 2 stages, however, seems artificial and does not exploit the full potential of the underlying methods. We extend the 2-stage approach to the general framework of any number of stages. Based on the theory of group sequential methods, we derive optimal multistage designs. With current genotyping cost structures, our results suggest that up to 4 stages are sufficient in order to get feasible and efficient designs. Furthermore, we consider the problem of choosing the optimal number of stages depending on the costs of the statistical interim analysis at each stage and provide guidelines for planning the number of stages in practice. In particular, we found that in the majority of cases both 3-stage designs and 4-stage designs are more efficient than 2-stage designs. Although prices for marker panels are showing a continuing downward trend, we still recommend implementing and using optimal multistage designs in practice. In addition to the immediate benefit, it will be necessary to acquire know-how regarding the application of multistage designs in order to be able to adapt the general framework of multistage designs to upcoming technologies in the area of GWAS.

Keywords: Group sequential methods; Maximal power; Minimal study costs; Optimal 2-stage design.

1. INTRODUCTION

By now genome-wide association studies (GWAS) with a case-control design are widely used in order to detect genes involved in complex diseases. Using large marker sets, allele or genotype frequency differences between cases and controls are studied for every single marker. The significance of association between a marker and the disease is determined by calculating an appropriate test statistic such as Armitage’s test for trend (Sasieni, 1997).

A straightforward genome-wide–testing approach requires genotyping large numbers of markers for every individual (1-stage design). Despite the prices for Single Nucleotide Polymorphism (SNP) arrays going down, cost is still a concern in GWAS. In addition, since no more than a few markers are expected...
to be potential disease markers, genotyping the entire marker set is inefficient. Hence in practice, analysis of GWAS is often performed in 2 (or sometimes more) stages. First, at stage 1, the full marker set is genotyped for a number of individuals. Second, at stage 2, only the most promising markers (e.g. markers that meet some genome-wide statistical significance threshold at stage 1) are processed for additional individuals using smaller SNP arrays. The sample sizes for the first and second stage can be specified in advance by splitting the total available sample of individuals into 2 groups. Another common approach is to derive the sample size of the second stage from the results of the analysis of the first stage, in which case the second stage is conducted as a follow-up study. The statistical analyses of 2-stage designs are usually performed using a replication-based approach instead of using a joint analysis approach combining the 2 stages, which is more efficient from a statistical point of view (Skol \textit{and others}, 2006). Furthermore, although optimal 2-stage designs for GWAS based on the principles of joint analysis appeared in the literature already some years ago (Satagopan \textit{and others}, 2002; Satagopan and Elston, 2003), they were rarely applied in the analysis of GWAS, probably because of both the complexity of the method and the sparse availability of user friendly software in order to construct optimal 2-stage designs. There is a stand-alone program, CATS (Skol \textit{and others}, 2006), which calculates optimal 2-stage designs under the constraint of fixed sample sizes. In addition, there are some simple \textbf{R} implementations available that can handle power calculation for 2-stage designs based on false-discovery rate (Kraft, 2006; Wang and Stram, 2006).

Two-stage designs discard markers after stage 1 and thus show a small loss of power as compared to genotyping the entire marker set, which, however, can be always compensated by slightly increasing the total sample size of the study (Wang \textit{and others}, 2006). The basic idea of 2-stage designs arises from group sequential methods as used in clinical trials (Jennison and Turnbull, 2000). Group sequential designs are used mainly to reduce the duration of a clinical trial. Therefore, in the course of a clinical trial, interim analyses are performed at pre-planned time points where in each time interval a certain number (i.e. group) of individuals is processed. Then, a test statistic is calculated sequentially at every interim analysis, and as soon as the cumulative test statistic has reached a statistically significant value, the trial can be stopped early. As a matter of fact, 2-stage designs for GWAS can be considered as a special case of general group sequential designs, consisting of 1 single interim analysis (stage 1) and a final analysis (stage 2). Furthermore, instead of reducing individuals, the main objective in GWAS is to reduce the overall amount of genotyping. As a result, the analysis of GWAS has not to be restricted to 2 stages but can be extended to multiple stages.

In this paper, we extend the special case of 2-stage designs to the general framework of multistage designs. We derive optimal multistage designs with respect to both minimal overall study costs and maximal study power. We show that using more than 2 stages can decrease the overall study costs significantly. In fact, if the costs of the statistical analyses are not taken into account, one could always decrease the overall study costs by including an additional interim analysis. Thus, the theoretically optimal procedure would perform an interim analysis after every single observation (i.e. after every genotyped individual). However, for practical applications, the number of stages will be rather small due to practical limitations of time, organization effort, and costs of interim analyses. For this reason, we consider the optimization problem in that we treat the number of stages as an additional free parameter. That is, we derive the optimal number of stages depending on the costs of the interim analyses. Since it may be difficult in practice to reliably specify costs of interim analyses in advance, we also provide practical guidelines for choosing an appropriate number of stages. Multistage designs furthermore rely on technologies that allow for processing SNPs in both whole-genome panels and smaller custom designed sets. We therefore present multistage designs based on currently available technology (Butler and Ragoussis, 2008; Steemers and Gunderson, 2007) and common prices for Illumina® SNP chips. We show that in general up to 4 stages will be sufficient in order to get feasible and efficient designs.

In this article, we focus on designs that are derived under a hypothesis-testing paradigm, that is, our designs maintain both a prespecified genome-wide type-I error rate, $\alpha$, and power, $1 - \beta$. Two-stage
approaches with respect to other statistical frameworks have also been proposed (Thomas and others, 2004; Bukszár and van den Oord, 2006; Zuo and others, 2006; Zheng and others, 2007).

2. METHODS

2.1 Statistical model

Statistical models for GWAS have been widely discussed in the setting of 2-stage designs. Since they can be applied to multistage designs in much the same way, we briefly introduce the used test statistic, and for more details, we refer the reader to the references. Consider $n$ individuals in a case–control design, for simplicity assuming $n/2$ cases and $n/2$ controls. Furthermore, consider evaluating $m$ markers from which $d$ are truly associated with the disease. We want to detect a disease marker by examining the difference of allele frequencies between cases and controls. For each marker, let $Y(i)$ and $X(i)$ denote the observed quantities of cases and controls that carry $i = 0, 1, 2$ copies of the allele under question. For example, $Y(2)$ denotes the number of cases that carry 2 risk alleles (for a marker) and $Y(0) + Y(1) + Y(2) = n/2$ corresponds to the total number of cases. Following Müller and others (2007), we define the linear trend statistic

$$Z = \sum_{i=0}^{2} \frac{w_i(Y(i) - X(i))}{\sqrt{ns^2}},$$

where $w_i$ are scores for the 3 genotypes with 0, 1, or 2 copies of the risk allele and $ns^2$ is an estimator of the variance of the numerator under the null hypothesis. Following the results of Zheng and Gastwirth (2006), in this article we use $s^2 = \sum_{i=0}^{2} w_i^2X(i)/\bar{n} - \left(\sum_{i=0}^{2} w_iX(i)/\bar{n}\right)^2$, which considers only the controls for the estimation of the variance. For more details regarding the possible choices of $s^2$, we refer to Zheng and Gastwirth (2006).

If $n$ is sufficiently large, $Z$ is approximately normally distributed with a location parameter $\mu \sqrt{n}$ and a dispersion parameter $\sigma^2$. Under the null hypothesis ($H_0$) of no association of the marker, $\mu = 0$ and $\sigma = 1$. Under the alternative hypothesis, $\mu = \mu_A$ and $\sigma^2 = \sigma_A^2$ depend on the weights $w_i$ and on the allele frequency and the penetrances of the 3 genotypes. For example, for a multiplicative odds model under a rare disease assumption, a reasonable choice of the weights is $w_0 = 0$, $w_1 = 1$, and $w_2 = 2$. Furthermore, if we let $p$ denote the population frequency of the risk allele and let $\Psi$ and $\Psi^2$ be the odds ratios for heterozygous and homozygous carriers of the risk allele, respectively, we get

$$\mu_A = \sqrt{\frac{1}{2}p(1-p)(\Psi - 1)} \frac{1}{1 + p(\Psi - 1)}, \quad \sigma_A^2 = \frac{(\frac{1}{2} + p(1-p))(\Psi - 1)^2 + 4\Psi}{4(1 + p(\Psi - 1))^2}.$$ (2.2)

In order to obtain $\mu$ and $\sigma$ for other genetic models, we refer to the appendix of Müller and others (2007).

2.2 One-stage design

For every marker locus, we aim at testing the null hypothesis that the marker is not associated with the disease ($H_0$: $\mu = 0, \sigma^2 = 1$) against the alternative hypothesis that the marker is associated with the disease ($H_A$: $\mu = \mu_A, \sigma^2 = \sigma_A^2$). To simplify matters, we suppose independent markers and 1-sided testing. In addition, the overall significance level $\alpha$ will be evenly distributed among all $m$ markers. One-stage designs are characterized by genotyping all $m$ markers on all $n$ individuals. Let $Z_{\text{full}}$ denote the test
statistic of a “single” marker in the “full” sample of all \( n \) individuals being calculated according to (2.1). Hence, the marker-wise type-I error rate and the power of the 1-stage design are given by \( \alpha = P_{H_0}(Z_{\text{full}} > z_{1-\alpha}) \) and \( 1 - \beta = P_{H_A}(Z_{\text{full}} > z_{1-\alpha}) \), respectively. Here, \( z_x \) denotes percentiles of the standard normal distribution, that is, \( \Phi(z_x) = x \).

### 2.3 Multistage design

Consider a multistage design of 2 or more, say \( K \), stages. A multistage or \( K \)-stage (subsequently, the terms “multistage” and “\( K \)-stage” are used interchangeably) design consists of a total number of \( n \) individuals and a starting marker set of \( m \) markers. To be exact, these \( m \) markers assumingly include \( d \) disease markers, which are truly associated with the disease, and \( m - d \) null markers, which are not associated with the disease. Furthermore, for every \( K \)-stage design, one can derive the overall type-I error rate \( \alpha \) and the power \( 1 - \beta \). Following group sequential design terminology, a \( K \)-stage design for a GWAS consists of \( K - 1 \) interim analyses at stages 1, 2, \ldots, \( K - 1 \) and is completed by a final analysis at stage \( K \). Generally, the starting marker set is shrinking after every stage. In particular, the most promising markers are carried forward to subsequent stages while statistically futile markers are dropped. At each stage, the decision whether to keep or drop a marker is made by calculating a \( p \)-value for that marker. Each marker with a \( p \)-value smaller than \( \alpha_i \) (\( i = 1, \ldots, K \)) will survive the \( i \)th stage. Markers that do not pass the threshold of \( \alpha_i \) are dropped and thus not treated in future stages. For every marker that happens to survive all \( K - 1 \) stages and also passes the threshold \( \alpha_K \) at the last stage, significant association with the disease will be declared.

The \( p \)-value for a marker at some stage \( k \) is based on the set of individuals that was accumulated up to this stage. For \( k = 1, \ldots, K \), let \( n_k \) denote the cumulative number of individuals at stage \( k \), where \( n = n_K \). Analogous to (2.1), we define

\[
Z_k = \frac{\sum_{i=0}^{2} w_i(Y_k(i) - X_k(i))}{\sqrt{n_k s_k^2}}
\]

(2.3)

to be the test statistic per stage \( k \). Here, \( Y_k(i) \) and \( X_k(i) \) denote the cumulative numbers of cases and controls, respectively, observed up to stage \( k \) with \( i = 0, 1, 2 \) copies of the risk allele. Let \( \pi_1 < \pi_2 < \cdots < \pi_K = 1 \) denote the cumulative sample proportions of the individuals at the corresponding stages (i.e. \( \pi_k = n_k/n \)). The joint distribution of \( Z_1, Z_2, \ldots, Z_K \) is approximated by a multivariate normal distribution with mean vector \( \mu = \mu \sqrt{n} \cdot (\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_{K-1}}, 1)^T \) and covariance matrix elements \( \Sigma_{i,j} = \sigma^2 \cdot \frac{\sqrt{\pi_i}/\pi_j}{\sqrt{\pi_j}} \), where \( 1 \leq i < j \leq K \). Given a set of \( k \) critical values \( \alpha_1, \alpha_2, \ldots, \alpha_k \), the cumulative type-I error rate at stage \( k \) under \( H_0 \) is given by

\[
a_k^* = P_{\mu=0, \sigma=1}(Z_1 > z_{1-\alpha_1}, Z_2 > z_{1-\alpha_2}, \ldots, Z_k > z_{1-\alpha_k}).
\]

(2.4)

In other words, (2.4) expresses the probability for a single null marker to survive stages 1 through \( k - 1 \). The cumulative power at stage \( k \) is calculated accordingly as

\[
1 - \beta_k^* = 1 - P_{\mu=\mu_A, \sigma=\sigma_A}(Z_1 > z_{1-\alpha_1}, Z_2 > z_{1-\alpha_2}, \ldots, Z_k > z_{1-\alpha_k})
\]

(2.5)

where \( \mu_A \) and \( \sigma_A \) are obtained from (2.2). For the numerical integration of the multivariate normal distributions, we used the package “mvtnorm” from the statistical programming language R (R D. C. Team, 2008).
2.4 Cost function

The overall costs $c$ for a GWAS basically consist of cost for phenotyping individuals and genotyping markers. For the sake of simplicity, we assume that phenotyping has been done already and focus on genotyping costs. Let $g_1, \ldots, g_K$ denote the costs of genotyping a single marker at stages $1, \ldots, K$. In case of a 1-stage design, each marker is genotyped on each individual with per marker costs of $g_1$. Thus, the overall costs of a 1-stage design using $n$ individuals result in $C_1 = n \cdot g_1 \cdot m$. If we let $E(m_2) \geq E(m_3) \geq \cdots \geq E(m_K)$ denote the expected number of markers at stages $2, \ldots, K$ to be genotyped “per individual,” the expected overall study costs $C_K$ of a $K$-stage design is given by

$$C_K = n(\pi_1 g_1 m + \sum_{k=2}^K (\pi_k - \pi_{k-1}) g_k E(m_k)),$$

The expected value of $m_k$ depends on the critical values $\alpha_k$ of the multistage design and is calculated by $E(m_k) = (m - d)\alpha_k^* + d(1 - \beta_k^*)$. Taken together, the overall expected study costs of $K$-stage designs can be described by a general cost function $C_K(\cdot)$ as

$$C_K(n, \pi_K, \alpha_K) = n \left( \pi_1 g_1 m + \sum_{k=2}^K (\pi_k - \pi_{k-1}) g_k ((m - d)\alpha_{k-1}^* + d(1 - \beta_{k-1}^*)) \right),$$

where $\pi_K = (\pi_1, \ldots, \pi_{K-1}, \pi_K = 1)^T$ and $\alpha_K = (\alpha_1, \ldots, \alpha_K)^T$ denote the vectors of the sample proportions and the “nominal” significance levels, respectively.

2.5 Optimization

Depending on the choice of the sample proportions $\pi_K$ and significance thresholds $\alpha_K$, any multistage design can be constructed. However, $\pi_K$ and $\alpha_K$ usually have to be set under the constraint in that the desired type-I error is maintained so that the resulting probability under the null hypothesis (Equation 2.4) equals some prespecified marker-wise significance threshold $\alpha$. Generally, all optimization problems of multistage designs are carried out under this constraint. In this article, we cover 2 important optimization problems of multistage designs. First, we minimize the overall expected study costs under a predefined overall power. Second, we treat the optimization problem of maximizing the overall power under a fixed overall budget.

In order to obtain a “cost-optimal” multistage design under a predefined overall power, we must search for the set of parameters $(n, \pi_K, \alpha_K)$ that minimize the overall costs, that is, $C_K(n, \pi_K, \alpha_K) \rightarrow \min$. In addition, the given constraints require the parameters to be chosen such that both the desired type-I error and the study power is maintained. In particular, $\alpha^*_K = \alpha$ and $1 - \beta^*_K = 1 - \beta$ for 2 prespecified values of $\alpha$ and $\beta$, where $\alpha^*_K$ and $1 - \beta^*_K$ are defined in $(2.4)$ and $(2.5)$. As a result, a cost-optimal multistage design consisting of $K$ stages leads to an optimization problem in $2K$ unknowns (i.e. $n, \pi_1, \ldots, \pi_{K-1}, \alpha_1, \ldots, \alpha_K$). More precisely, two of the $2K$ unknowns are defined by the 2 above constraints so that ultimately the optimization has to be carried out in $2K - 2$ unknowns. Note that the sample size $n$ can be either a free parameter as part of the minimization problem or a fixed parameter as often is the case in practice, in which case the optimization problem is reduced to $2K - 3$ unknowns. In the following, we denote the parameters that belong to an optimal multistage design by $\pi_K^{\text{opt}}, \alpha_K^{\text{opt}}$, and $n_{\text{opt}}$. If the sample size is fixed, we use $n_{\text{fix}}$.

The demand for a “power-optimal” multistage design arises when the researcher has a fixed overall budget, $b_{\text{fix}}$, which can be spend for the GWAS, and he searches for the study design with maximal power under the given budget constraint. To simplify matters, we assume a fixed total sample size $n_{\text{fix}}$ in this optimization problem. First of all, we are not interested in scenarios where the available budget is greater than $C_1$ (i.e. $b_{\text{fix}} > n_{\text{fix}} \cdot g_1 \cdot m$). In this case, all markers can be genotyped for every individual, which as a matter of course leads to the maximal achievable study power. Hence, we are interested in cases

Optimal multistage designs
where the costs of a 1-stage design exceed the available budget \((C_1 > b_{\text{fix}})\). In such cases, a 1-stage design is no longer appropriate because in order to maintain the budget one had reduce the number of genotyped individuals to \(\tilde{n} = b_{\text{fix}}/(g_1 \cdot m)\). Thus, if \(C_1 > b_{\text{fix}}\), multistage designs are clearly more powerful because they can always incorporate all available individuals into the analysis, controlling the overall costs by discarding markers for an appropriate proportion of the individuals. The corresponding power-optimal \(K\)-stage design is computed similar to the cost-optimal design above except that, instead of treating \(C_K(n_{\text{fix}}, \pi_K, \beta_K) \rightarrow \min\), we perform optimization with respect to maximizing the overall power. That is, we search for \(\pi_K\) and \(\alpha_K\) such that \((1 - \beta_K) \rightarrow \max\) while maintaining \(\alpha_K = \alpha\) and \(C_K(n_{\text{fix}}, \pi_K, \alpha_K) = b_{\text{fix}}\) for 2 prespecified values \(\alpha\) and \(b_{\text{fix}}\).

As outlined before, calculating the joint probabilities for (2.4) and (2.5) requires numerical integration so that the entire optimization is done numerically too. In order to get a first approximate estimate about the global location of the optimal sample proportions \((\pi_1^{\text{opt}}, \ldots, \pi_{K-1}^{\text{opt}}, \pi_K = 1)\), we first determine the optimal sample proportions of the \((K - 1)\)-stage design, that is, \((\pi_1^{\text{opt}}, \ldots, \pi_{K-2}^{\text{opt}}, \pi_{K-1} = 1)\). This leads to recursive evaluations down to \(K = 2\). Using \((\pi_{K-1}^{\text{opt}}, 1)\) as starting values for \(\pi_K^{\text{opt}}\), we perform a naive search for the optimum on a grid near these starting values. Moreover, for every occurrence \(\pi_K\) treated in this first search, we find the corresponding optimal \(\alpha_K\) using the Nelder–Mead (1965) simplex algorithm, implemented in the R function optim. Once the area of the global minimum is narrowed down, we finish optimization using the Nelder–Mead simplex algorithm for the entire set of parameters \(\pi_K\) and \(\alpha_K\). Furthermore, in every single optimization step, we apply 1D numerical root finding (R function uniroot) in order to obtain the 2 parameters out of the 2\(K\) unknowns that are defined by the constraints (i.e. type-I error and overall power or overall budget, respectively). For more details on the optimization routines, we refer the reader to the well-documented source code, which was completely written in the programming language R and is freely available upon request.

3. Results

Conducting a multistage design in practice requires different genotyping platforms at the different stages. In order to reflect the current genotyping cost structure, for stage 1, we take the example of Illumina® Infinium® whole-genome panel, which consists of \(m = 610\,000\) markers per panel (Illumina® DNA Analysis, 2008). For stage 2, we assume that the Illumina® iSelect™ SNP chip is applied, which provides interrogation of up to 60 800 SNPs from 12 samples simultaneously (Illumina® DNA Analysis, 2007). For all subsequent stages \((3, \ldots, K)\), we treat the Illumina® GoldenGate® assay, providing 96, or from 384 to 1 536 SNPs per sample, 16 or 96 samples in parallel (Illumina® SNP Genotyping, 2006). Other genotyping platforms are likely to yield similar results. In addition, the assumptions about applying certain SNP chips to certain stages are not meant to be absolute but may vary due to the conditions of a GWAS. For example, the expected number of markers at the different stages could determine both the technology and the platform to be used. Based on the setting above and current prices for SNP chips, we set the genotyping costs “per SNP” at the different stages to be \(g_1 = 0.0005\), \(g_2 = 0.01\), \(g_3 = 0.02\), and \(g_{(K \geq 4)} = 0.05\) so that genotyping costs at stage 4 as compared to stage 1 are assumed to increase by a factor of 100.

We set the genome-wide type-I error rate to 0.05/2 (one-sided) and study power to 80%. Furthermore, we treat \(m = 610\,000\) markers and set the number of disease markers to \(d = 1\). Note that our results basically do not depend on the actual number of the disease markers because they are always comparatively small in numbers as compared with the number of null markers. To account for multiple testing, one could apply marker-wise Bonferroni correction yielding a marker-wise \(\alpha = 0.05/2m\). Since there may be more than a few causal variants in reality, Bonferroni correction could be conservative so that we decided to evaluate the performance of the multistage designs under a more relaxed type-I error rate of \(\alpha = 10 \times 0.05/2m \approx 4.1 \times 10^{-7}\).
3.1 Cost-optimal multistage designs

Calculating multistage designs requires assumptions about the underlying genetic effects and models. To simplify matters, we derive all subsequently presented results using the multiplicative odds model as outlined in Section 2.1. Note that altering the genetic model would lead to different formulas for $\mu_A$ and $\sigma^2_A$ and thus can be seen as simply adjusting $\mu_A$ and $\sigma^2_A$. The conclusions about the presented multistage designs therefore do not depend on specific genetic models. The impact of varying genetic parameters (risk allele frequency, genetic effect size) can be seen as varying $\mu_A$ and $\sigma^2_A$ as well and was already analyzed for the case of 2-stage designs by Wang and others (2006) and Müller and others (2007). Since 2-stage designs are special cases of multistage designs, the basic findings from these works carry forward to multistage designs. As a result, if the sample size $n$ is a free parameter in the optimization problem, we found that cost-optimal multistage designs (i.e. fixed power, costs $\rightarrow$ min) do not depend on varying genetic parameters (results not shown). That is, the design-specific parameters such as sample proportions $\pi_k$ at the different stages as well as significance thresholds $\alpha_k$ do not change upon varying allele frequencies or genetic effect sizes, respectively. In contrast, if the sample size $n$ is fixed in advance, as often is the case

Table 1. Dependence of the sample proportions of optimal multistage designs with fixed sample size upon varying genetic parameters $p$ (risk allele frequency) and $\Psi$ (odds ratio for heterozygous carriers of the risk allele). The parameter setup is $m = 610,000$, $a = 10 \times 0.05/2m$ (1-sided), $g_1 = 0.0005$, $g_2 = 0.01$, $g_3 = 0.02$, and $g_4 = g_5 = g_6 = 0.05$. For each multistage design, the power is 80%

<table>
<thead>
<tr>
<th>Number of stages</th>
<th>Sample proportion</th>
<th>Number of expected markers</th>
<th>Expected costs ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1$ $\pi_2$ $\pi_3$ $\pi_4$ $\pi_5$</td>
<td>$E(m_2)$ $E(m_3)$ $E(m_4)$ $E(m_5)$ $E(m_6)$</td>
<td></td>
</tr>
<tr>
<td>$n_{\text{fix}} = 6000$, $p = 0.1$, $\Psi = 1.4$ ($\mu_A = 0.082$ and $\sigma^2_A = 1.316$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.37 1</td>
<td>2595</td>
<td>767 800</td>
</tr>
<tr>
<td>3</td>
<td>0.33 0.56 1</td>
<td>5939 437</td>
<td>704 200</td>
</tr>
<tr>
<td>4</td>
<td>0.32 0.51 0.74 1</td>
<td>6895 856 73</td>
<td>694 000</td>
</tr>
<tr>
<td>5</td>
<td>0.32 0.49 0.68 0.80 1</td>
<td>7237 1042 162 44</td>
<td>690 900</td>
</tr>
<tr>
<td>6</td>
<td>0.32 0.49 0.66 0.73 0.84</td>
<td>7425 1144 225 102 30</td>
<td>689 500</td>
</tr>
<tr>
<td>$n_{\text{fix}} = 6000$, $p = 0.1$, $\Psi = 1.5$ ($\mu_A = 0.101$ and $\sigma^2_A = 1.394$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.24 1</td>
<td>1515</td>
<td>502 700</td>
</tr>
<tr>
<td>3</td>
<td>0.20 0.40 1</td>
<td>4802 216</td>
<td>438 900</td>
</tr>
<tr>
<td>4</td>
<td>0.19 0.34 0.57 1</td>
<td>6048 616 34</td>
<td>427 800</td>
</tr>
<tr>
<td>5</td>
<td>0.19 0.32 0.49 0.63 1</td>
<td>6516 827 103 19</td>
<td>424 400</td>
</tr>
<tr>
<td>6</td>
<td>0.19 0.32 0.46 0.55 0.69</td>
<td>6616 915 155 57 13</td>
<td>423 000</td>
</tr>
<tr>
<td>$n_{\text{fix}} = 3000$, $p = 0.3$, $\Psi = 1.4$ ($\mu_A = 0.116$ and $\sigma^2_A = 1.139$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.34 1</td>
<td>2460</td>
<td>359 600</td>
</tr>
<tr>
<td>3</td>
<td>0.30 0.53 1</td>
<td>5955 412</td>
<td>325 500</td>
</tr>
<tr>
<td>4</td>
<td>0.29 0.47 0.71 1</td>
<td>7019 856 69</td>
<td>319 700</td>
</tr>
<tr>
<td>5</td>
<td>0.29 0.46 0.64 0.77 1</td>
<td>7399 1065 162 41</td>
<td>317 900</td>
</tr>
<tr>
<td>6</td>
<td>0.29 0.45 0.61 0.69 0.80</td>
<td>762 1187 233 101 29</td>
<td>317 000</td>
</tr>
<tr>
<td>$n_{\text{fix}} = 3000$, $p = 0.3$, $\Psi = 1.5$ ($\mu_A = 0.141$ and $\sigma^2_A = 1.168$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.23 1</td>
<td>1516</td>
<td>247 700</td>
</tr>
<tr>
<td>3</td>
<td>0.19 0.39 1</td>
<td>4978 216</td>
<td>213 700</td>
</tr>
<tr>
<td>4</td>
<td>0.19 0.32 0.54 1</td>
<td>6375 649 33</td>
<td>207 500</td>
</tr>
<tr>
<td>5</td>
<td>0.18 0.31 0.47 0.61 1</td>
<td>6863 888 109 19</td>
<td>205 600</td>
</tr>
<tr>
<td>6</td>
<td>0.18 0.30 0.44 0.52 0.66</td>
<td>7101 1012 172 61 12</td>
<td>204 700</td>
</tr>
</tbody>
</table>
in practice, multistage designs will depend on the genetic parameters as can be seen in Table 1. Basically, with increasing $\Psi$ and $p$, also the effect parameter $\mu_A$ increases, resulting in an increased study power. Thus, the sample proportions of the leading stages ($\pi_1, \pi_2, \ldots$) can be reduced because the increased power increases the probability of detecting the true disease markers and keeping them in subsequent stages.

For the most part, the presented results were derived using the genotyping cost structure from above. Since the costs of genotyping are known to have an impact on the parameters of 2-stage designs, we exemplarily provide results of multistage designs with both lower and higher genotyping costs than above (Table 2). Again, the basic results from 2-stage designs carry forward to multistage designs. That is, expensive genotyping in latter stages generally results in less markers being carried over to these stages as can be observed in the columns displaying the number of expected markers (Table 2). It is interesting to note that the genotyping cost structure might also vary depending on the number of stages. For example, consider the number of expected markers at stage 2 (column $E(m_2)$, Table 2). Since the number of expected markers increases with $K$, the genotyping costs might be reduced if cheaper SNP panels can be applied for larger marker set. In this case, increasing $K$ would lead to additional cost savings. We will not consider this subject further, however, as it required detailed assumptions about the applied SNP technology.

In practice, the sample size, that is, the number of available individuals, will be often fixed in advance or at least restricted. Hence, we calculated cost-optimal multistage designs for a range of fixed sample sizes. The resulting minimum achievable costs are depicted in Figure 1 for optimal multistage designs of up to 5 stages. Obviously, there is a trade-off between increasing $n$ and minimizing the costs. While increasing $n$ reduces the costs at first, at some point, the costs raise again. These turning points are located at the minimum of each curve (Figure 1) and correspond to the cost-optimal sample size for the respective number of stages. For example, if the number of available individuals is 6000 and the researcher wants to conduct a 3-stage design (based on parameter setting as given in Figure 1), it would be inefficient to use the entire sample of 6000 individuals because the cost-optimal sample size of the 3-stage design is $n_{opt} = 5346$. Intuitively, these turning points can be explained by the fact that multistage designs save costs by

---

**Table 2. Dependence of the sample proportions of optimal multistage design with fixed sample size upon varying per SNP genotyping costs $g_k$.** The parameter setup is $n_{fix} = 6000$, $m = 610,000$, $\alpha = 10 \times 0.05/2m$ (1-sided), risk allele frequency $p = 0.1$, and heterozygous odds ratio of $\Psi = 1.5$ ($\mu_A = 0.101$ and $\sigma_A^2 = 1.394$). For each multistage design, the power is 80%.

<table>
<thead>
<tr>
<th>Number of stages</th>
<th>Sample proportion</th>
<th>Number of expected markers</th>
<th>Costs ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1$ $\pi_2$ $\pi_3$ $\pi_4$ $\pi_5$</td>
<td>$E(m_2)$ $E(m_3)$ $E(m_4)$ $E(m_5)$ $E(m_6)$</td>
<td></td>
</tr>
<tr>
<td>$g_1 = 0.0005$, $g_2 = \cdots = g_6 = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.24 1</td>
<td>1515</td>
<td>502700</td>
</tr>
<tr>
<td>3</td>
<td>0.19 0.35 1</td>
<td>5841 500</td>
<td>426800</td>
</tr>
<tr>
<td>4</td>
<td>0.18 0.25 0.42 1</td>
<td>9191 2444 259</td>
<td>403900</td>
</tr>
<tr>
<td>5</td>
<td>0.17 0.22 0.31 0.47 1</td>
<td>11542 4591 1333 164</td>
<td>393400</td>
</tr>
<tr>
<td>6</td>
<td>0.17 0.21 0.26 0.35 0.52</td>
<td>13179 6373 2693 836 115</td>
<td>387500</td>
</tr>
<tr>
<td>$g_1 = 0.0005$, $g_2 = \cdots = g_6 = 0.05$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.30 1</td>
<td>324</td>
<td>613300</td>
</tr>
<tr>
<td>3</td>
<td>0.26 0.40 1</td>
<td>1223 115</td>
<td>543000</td>
</tr>
<tr>
<td>4</td>
<td>0.25 0.32 0.48 1</td>
<td>1821 500 60</td>
<td>521500</td>
</tr>
<tr>
<td>5</td>
<td>0.24 0.28 0.37 0.53 1</td>
<td>2296 930 282 39</td>
<td>511600</td>
</tr>
<tr>
<td>6</td>
<td>0.24 0.27 0.32 0.41 0.57</td>
<td>2615 1287 556 180 28</td>
<td>506100</td>
</tr>
</tbody>
</table>
Fig. 1. Curves show overall genotyping costs of optimal multistage designs as a function of the sample size. The parameter setup is $m = 610,000$, $\alpha = 10 \times 0.05/2m$ (one-sided), $p = 0.1$, $\Psi = 1.5$ ($\mu_A = 0.101$ and $\sigma^2_A = 1.394$), $g_1 = 0.0005$, $g_2 = 0.01$, $g_3 = 0.02$, and $g_4 = g_5 = 0.05$. The power was set to 80%. The costs and sample size of a corresponding 1-stage design with 80% power is $1,049,061$ and $n = 3,440$, respectively. The horizontal dotted lines correspond to tangents of the curves minima in which also the vertical dotted lines originate indicating the cost-optimal sample size for the multistage designs with their respective number of stages.

discarding markers, which would reduce the power of the study. Thus, to compensate for the loss of power, the sample size is increased, which at the same time increases the number of genotyped markers just because of the growing sample size. As a result, the optimal sample size of a multistage design with a certain number of stages represents an optimal balance between discarding markers and increasing sample size.

Looking at the shape of the curves, it can be seen that they flatten noticeably with increasing number of stages. Hence, the more stages are used, the wider is the region of an “almost-optimal” sample size. Furthermore, the overall costs of genotyping can be seen to decrease with every added stage. At the same time, the additional costs savings become increasingly marginal with increasing number of stages. In fact, the curve for 6-stage designs was almost congruent with the 5-stage curve and thus is omitted in Figure 1.

### 3.2 Optimal number of stages

The results presented in Figure 1 solely refer to costs of genotyping. In practice, however, at every stage, the statistical analysis and evaluation of the interim results will cause additional costs, which we have disregarded so far. In order to account for these additional costs, we define the total study costs as $C_{\text{total}}(K, n, \pi_K, \alpha_K) = C_K(n, \pi_K, \alpha_K) + K \cdot a$, where $a$ denotes the additional costs of one analysis, which for simplicity is assumed to be equal at every stage. In order to determine the optimal number of stages, $K_{\text{opt}}$, we construct the cost-optimal $K$-stage designs for $K = 1, 2, \ldots$ and then determine $\min\{C_K(n_{\text{opt}}, \pi_K^{\text{opt}}, \alpha_K^{\text{opt}}) + K \cdot a | K = 1, 2, \ldots\}$. To give an example, we determine $K_{\text{opt}}$ for the previous parameter setting from Figure 1. We assume an existing sample of $n_{\text{fix}} = 4,500$ available individuals, which is shortly below the optimal sample size of the 2-stage design ($n_{\text{opt}} = 4,730$, see Figure 1). Figure 2 depicts the total study costs $C_{\text{total}}(K, n_{\text{fix}} = 4,500, \pi_K^{\text{opt}}, \alpha_K^{\text{opt}})$ as a function of $a$. In this scenario, if an interim analysis costs about $1k$, the 5-stage design will be already as equally expensive as the 6-stage design. Furthermore, from interim analysis costs of about $2k$ to $8k$, the 4-stage design will be the cheapest design, while from $8k$ on the 3-stage design is the cheapest choice until, at some point (about $48k$, not shown in Figure 2), the 2-stage design would obtain the total cost minimum. Another point worth mentioning is that additional stages can prolong a GWAS due to the additional genotyping and statistical
3.3 Power-optimal multistage designs

Our second optimization problem arises in practice when the researcher is restricted to a certain budget, which does not allow for genotyping the entire marker set on all available individuals ($n_{fix} \cdot g_1 \cdot m > b_{fix}$). In this case, the researcher is interested in extracting the most possible information under the budget constraint. Leaving costs of interim analyses aside, this means to maximize the overall study power while keeping the overall genotyping costs constant. Figure 3 displays maximal achievable power for optimal multistage designs of up to 5 stages. Again, we omit displaying the curve for the 6-stage design as it would overlap the 5-stage curve. While we still observe substantial power gains when comparing optimal 2-stage designs with optimal 3-stage designs, the results suggest once more that for practical purposes using up to 4 stages will be enough, even more so as Figure 3 does not yet include costs of interim analyses. Moreover, the power of the optimal multistage designs can be seen to quickly converge to the absolutely maximal achievable power (89.1%, dashed line in Figure 3). For example, using only half of the available budget, $0.5$ Mio, yields already about 79% power for multistage designs of 3 or more stages, which is only 10% below the maximum.

3.4 Algorithms and software

For the construction of optimal multistage designs, the most difficult part arises from the numerical optimization. The accuracy of the results mainly depends on 2 factors. First, it is directly connected with the numerical accuracy of the underlying algorithm that performs the numerical integration of the multivariate normal distributions. Second, the accuracy of the results depends on the performance of the optimization procedure in that it can find the global optimum. In any case, with growing number of stages, the optimization problem gets increasingly complex due to the growing set of parameters to optimize. Furthermore,
due to the large number of markers, the critical values, $\alpha_k$, tend to be very small, which poses additional numerical challenge.

We developed an algorithm, which as a matter of principle may compute multistage designs of any number of stages. However, with increasing number of stages, computational time is heavily increasing as well. Thus, we did not yet test our algorithm for more than 6 stages. The calculation of cost-optimal designs under a fixed sample size with 3, 4, 5, and 6 stages took about 0.5 h, 3 h, 5–10 h, and 24–48 h, respectively. The construction of power-optimal designs under a fixed sample size with 3, 4, 5, and 6 stages took about 7 min, 15 min, 60 min, and 4 h, respectively. Two-stage designs will take only a few seconds in both cases. All calculations were performed using an AMD Dual Core Opteron 270 with 2.0 GHz. If $n$ is optimized as a free parameter, the respective computation times have to be multiplied by a factor of about 20. Finally, our algorithm does not require the specification of any starting values for the optimized parameters.

Since we have constructed optimal designs with more than 2 stages for the first time, the results cannot be validated by comparison with existing methods. However, comparing the outcome of the different stages with each other, the results are reasonable. In particular, we are very confident that our optimization algorithm yields results with high accuracy for at least up to 4 stages, which we showed to be sufficient for practical purposes. Furthermore, to overcome the lack of user friendly software, we are currently developing an R package that enables the researcher to easily compute optimal multistage designs. The package will provide full functionality of all methods being discussed in the present article. In addition, the package comes along with a user friendly graphical user interface and will be made freely available on the comprehensive R archive network.

4. DISCUSSION

In order to investigate the additional benefit of adding stages, we have constructed designs of up to 6 stages. Our results show that the economical benefits of 2-stage designs can be expanded by using more than 2 stages. At the same time, the benefit of switching from 3 to 4 stages is less pronounced. According to this, further cost savings decrease with increasing number of stages. From a theoretical point
of view, using more than 2 stages will be always more efficient with respect to genotyping costs. In prac-
tice, however, applying multistage designs requires additional efforts and costs with every additional stage
(data processing, statistical, etc.). Therefore, we considered the problem of choosing the optimal number
of stages with respect to the total study costs, taking into consideration the cost of both genotyping and
statistical analysis.

Our results are based on the current genotyping cost structure and show that using no more than
4 stages will be sufficient for practical purposes. The investigator will have to consider the intrinsic condi-
tions of his or her study to determine an appropriate number of stages to be used, particularly, the available
technology, the available amount of time, and the costs of the statistical interim analyses. In general, the
more expensive the statistical interim analyses the less stages can be used.

We provide algorithms to construct optimal multistage designs and currently are developing an R
package that enables the investigator to compute optimal multistage designs in a user friendly software
environment. In practice, however, it may still be difficult for the investigator to plan and conduct a GWAS
using a multistage design because of the lack of experience with this approach in the area of GWAS. As
outlined earlier, the basic idea of multistage designs arises from group sequential methods as used in clin-
ical trials. Group sequential designs are by now widely applied in clinical research environment. On the
other hand, the basic methods of group sequential analysis were published years before clinical investiga-
tors started using them in clinical research practice. Furthermore, it took decades for group sequential
designs to become established as a standard method for the conduction of clinical trials. In this context, be-
sides the immediate economical benefits, we recommend using optimal multistage designs in the practice
of GWAS in order to both acquire expertise and establish the underlying methods. In doing so, the general
framework of multistage designs can be adapted faster to upcoming technologies in the area of GWAS
and, as a result, the benefit from multistage designs might be more sustainable for future technologies.

ACKNOWLEDGMENTS

We thank 2 reviewers for valuable comments that lead to considerable improvement of the paper. Conflict
of Interest: None declared.

FUNDING

German Research Foundation (Deutsche Forschungsgemeinschaft) (SCHA 542/9–4).

REFERENCES


York: Chapman and Hall.
Optimal multistage designs


[Received November 30, 2007; revised August 22, 2008; accepted for publication September 16, 2008]