STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:4145–4160
Published online 28 April 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3300

Optimized multi-stage designs controlling the false discovery or the family-wise error rate

Sonja Zehetmayer, Peter Bauer and Martin Posch*, †

Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

SUMMARY

When a large number of hypotheses are investigated, we propose multi-stage designs where in each interim analysis promising hypotheses are screened, which are investigated in further stages. Given a fixed overall number of observations, this allows one to spend more observations for promising hypotheses than with single-stage designs, where the observations are equally distributed among all considered hypotheses. We propose multi-stage procedures controlling either the family-wise error rate (FWER) or the false discovery rate (FDR) and derive asymptotically optimal stopping boundaries and sample size allocations (across stages) to maximize the power of the procedure. Optimized two-stage designs lead to a considerable increase in power compared with the classical single-stage design. Going from two to three stages additionally leads to a distinctive increase in power. Adding a fourth stage leads to a further improvement, which is, however, less pronounced. Surprisingly, we found only small differences in power between optimized integrated designs, where the data of all stages are used in the final test statistics, and optimized pilot designs where only the data from the final stage are used for testing. However, the integrated design controlling the FDR appeared to be more robust against misspecifications in the planning phase. Additionally, we found that with increasing number of stages the drop in power when controlling the FWER instead of the FDR becomes negligible.

Our investigations show that the crucial point is not the choice of the error rate or the type of design, but the sequential nature of the trial where non-promising hypotheses are dropped in the early phases of the experiment.

KEY WORDS: group sequential designs; gene-disease association; multiple testing; microarray

1. INTRODUCTION

In multi-stage designs for gene association or expression studies, early stages are used to screen promising genes out of all initially studied genes. Only the selected genes are investigated in further

*Correspondence to: Martin Posch, Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
†E-mail: Martin.posch@meduniwien.ac.at

Contract/grant sponsor: Austrian FWF-Fund; contract/grant number: P18698-N15

Received 12 October 2007
Accepted 10 March 2008

stages based on additional observations. There is a rich literature on the special case of two-stage designs that demonstrate the superiority of the sequential approach compared with the conventional single-stage designs where the observations are evenly distributed among the considered hypotheses [1–8]. In this paper, we explore the gain in efficiency when increasing the number of stages. We give a general framework for the implementation of multi-stage designs for testing problems with a large number of hypotheses controlling either the false discovery rate (FDR) or the family-wise error rate (FWER). We cover designs with deterministic stage-wise sample sizes as well as designs where the overall number of observations is fixed and at each stage a pre-specified fraction of the observations is evenly distributed among the selected hypotheses. The latter approach leads to data-dependent stage-wise sample sizes for each individual hypothesis. For designs with a fixed overall number of observations, we derive asymptotically (letting the number of hypotheses go to infinity) optimal stopping boundaries and asymptotically optimal allocations of observations across stages. Here, we compare *pilot designs*, where the final test statistics are based only on data from the last stage, as well as *integrated designs*, where the pooled data from all stages enter the final test decision.

Multi-stage designs can be applied in gene–disease association studies, where a large number of marker loci are investigated in order to identify genes, when conferring for a disease of interest. Here, the design constraint is often the total cost, represented by the total number of gene evaluations rather than the total number of individuals (cf. [3]). We show that increasing the number of stages leads to a pronounced gain in efficiency regardless of the error criterion (FDR or FWER) or the type of design (pilot or integrated). When applying asymptotically optimal design parameters in experiments with three or four stages, we observed only minor differences in power between the different types of designs and error rates. However, integrated designs and designs controlling the FDR appear to be more robust to misspecifications of the design parameters.

The paper is structured as follows: In Section 2 we specify the testing problem, shortly review the estimation of the FDR in single-stage designs (Section 2.1) and introduce the pilot (Section 2.2) and the integrated design (Section 2.3). Here the per-hypothesis sample sizes in each stage are assumed to be fixed. As of Section 3 the overall number of observations is assumed to be fixed and thus the per-hypothesis sample sizes are random variables. In Section 4 we derive asymptotic expressions for the power and compute asymptotically optimal stage-wise sample sizes and selection thresholds for several settings. We investigate the robustness of the optimal designs to deviations from the planning assumptions and give a real data example. Results are summarized in a short discussion.

2. DESIGNS WITH FIXED STAGE-WISE PER-HYPOTHESIS SAMPLE SIZES

We consider an experiment with m_1 one-sided null hypotheses for the means $\mu^{(i)},\ i=1,\ldots,m_1$ of independent, normally distributed observations with known variances $(\sigma^{(i)})^2$, assuming also independence across hypotheses. We test the hypotheses

$$H_0^{(i)}:\ \mu^{(i)}=0 \text{ against } H_1^{(i)}:\ \mu^{(i)}>0, \quad i=1,\ldots,m_1$$

and assume that for a fraction π_1 of the m_1 hypotheses the null hypothesis holds. Below we investigate single- and multi-stage designs to test the m_1 hypotheses controlling either the FDR, defined as the expected proportion of Type I errors among the rejected hypotheses (e.g. [9]), or the FWER, defined as the probability of at least one Type I error. Consider a multi-stage experiment with k stages and $k-1$ interim analyses. At each interim analysis, only hypotheses with promising

DOI: 10.1002/sim

effect sizes are carried on to the subsequent stage for being further observed. The other hypotheses are dropped and accepted. Let $m_2 \geqslant \cdots \geqslant m_k$ denote the number of hypotheses considered at stages $2, \ldots, k$. Note that for $t>1$ the m_t are random variables that depend on the outcomes of the preceding stages.

In this section, we consider the case of deterministic sample sizes balanced over the selected hypotheses per stage. As of Section 3, designs with a fixed overall number of observations and consequently random stage-wise per-hypothesis sample sizes are considered.

2.1. Single-stage designs

We first consider single-stage designs ($k=1$). The stage-wise per-hypothesis sample sizes are given by n_1; $s_1^{(i)}$ denotes the sum of observed values for hypothesis i. The p-values are then given by $p_1^{(i)} = 1-\Phi\{s_1^{(i)}/(\sigma^{(i)}\sqrt{n_1})\}$, where $\Phi(.)$ denotes the cumulative distribution function of the standard normal distribution. For the control of the FWER at level α, we use the Bonferroni adjustment and apply individual significance levels $\gamma = \alpha/m_1$ for each hypothesis. To control the FDR, the expected fraction of erroneously rejected null hypotheses among all rejected hypotheses is estimated. If all hypotheses with an individual p-value $p_1^{(i)}$ smaller than a critical value γ are rejected, the resulting FDR can be estimated by [10]

$$\widehat{\mathrm{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_1 \gamma m_1}{\max(\#\{p_1^{(i)} \leqslant \gamma\}, 1)}, \quad i = 1, \ldots, m_1 \tag{1}$$

Here $\hat{\pi}_1$ is an estimator of π_1 given by

$$\hat{\pi}_1 = \#\{p_1^{(i)} > \lambda\}/\{(1-\lambda)m_1\} \tag{2}$$

where $\#\{p_1^{(i)} > \xi\}$ denotes the number of p-values exceeding ξ. $\lambda \in [0,1]$ is a tuning parameter chosen *a priori*. Increasing λ reduces the bias of $\hat{\pi}_1$ at the cost of a higher variance. Storey [10] suggested to set $\lambda = 0.5$, which gives a good trade-off of bias and variance. Alternatively, he suggested a Bootstrap algorithm to choose λ. To perform a test with a specified FDR α, the largest γ is determined such that $\widehat{\mathrm{FDR}}_\lambda(\gamma) \leqslant \alpha$. Storey *et al.* [11, Theorem 3] showed that this procedure controls the FDR if the p-values corresponding to the true null hypotheses are independent and uniformly distributed.

2.2. Pilot designs

In the pilot design, several single-stage experiments are performed consecutively. At each interim analysis hypotheses are dropped and accepted based on the stage-wise p-values of the preceding stage. At the final stage k, the remaining hypotheses are tested based on the p-values $p_k^{(i)}$ from the last stage. More formally, let $(\gamma_1, \ldots, \gamma_{k-1})$ denote a vector of critical values. The stage-wise per-hypothesis sample sizes are denoted by $n_1, \ldots, n_k$, and $s_t^{(i)}$ denotes the sum of observations for hypothesis i from stage t. The stage-wise p-values are then given by $p_t^{(i)} = 1-\Phi\{s_t^{(i)}/(\sigma^{(i)}\sqrt{n_t})\}$. At interim analysis t, $t=1, \ldots, k-1$ all hypotheses that reached stage t and for which $p_t^{(i)} \leqslant \gamma_t$ are selected for stage $t+1$. In the final analysis, all hypotheses that reached stage k and for which $p_k^{(i)} \leqslant \gamma_k$ are rejected. To control the FWER of the procedure overall we set $\gamma_k = \alpha/m_k$, where m_k denotes the number of hypotheses that reached stage k. To control the FDR of the experiment, the

DOI: 10.1002/sim

critical value γ_k is given by the maximal γ (as in the single-stage design) such that

$$\frac{\hat{\pi}_k \gamma m_k}{\max(\#\{p_k^{(i)} \leqslant \gamma\}, 1)} \leqslant \alpha \tag{3}$$

where $\hat{\pi}_k$ is the estimated proportion of true null hypotheses in stage k given by (2) with $p_1^{(i)}$ replaced by $p_k^{(i)}$. It is the special feature of the pilot design that previous stages are only used for selection of hypotheses but not for testing. Therefore, a test procedure controlling the FDR (FWER) at the last ('inferential') stage controls the FDR (FWER) overall.

A related approach was investigated by Van den Oord and Sullivan [12]. For each stage they fixed the FDR and then applied the corresponding critical values.

2.3. Integrated design

The pilot design uses at each interim analysis only the data of the directly preceding stage. Although this approach is appealing for its simplicity, efficiency may be gained by using designs that employ the accumulated data at each stage. Such designs can be realized with group sequential plans [13] for each hypothesis i. Let $\tilde{s}_t^{(i)}$ denote the sum of all observations for hypothesis i from stage 1 to stage t (in the following the symbol '$\sim$' always denotes the cumulative quantity). Then $\tilde{p}_t^{(i)} = 1 - \Phi(\tilde{s}_t^{(i)} / (\sigma^{(i)} \sqrt{\sum_{j=1}^{t} n_j}))$ gives the (cumulated) p-value for hypothesis i at stage t. To define the stopping boundaries, we specify continuation probabilities $g_t, t = 1, \ldots, k-1$, which give the probability under the null hypothesis that the trial continues at least to stage $t+1$. The continuation probabilities define critical values $\tilde{\gamma}_t$ for the cumulative p-values $\tilde{p}_t^{(i)}$: a hypothesis is accepted early at stage $t < k$ (and excluded from further consideration) if $\tilde{p}_t^{(i)} > \tilde{\gamma}_t$. $\tilde{\gamma}_t$ at stages $t = 1, \ldots, k-1$ are recursively defined by $\tilde{\gamma}_1 = g_1$, and the solutions of

$$g_t = P_{H_0^{(i)}} \left[\bigcap_{l=1}^{t} \{\tilde{P}_l^{(i)} \leqslant \tilde{\gamma}_l\} \right] \tag{4}$$

in $\tilde{\gamma}_t$. Here $\tilde{P}_l^{(i)}$ denote random variables. To specify the rejection region, for each individual hypothesis i we derive a sequential p-value based on a stage-wise ordering of the sample space [14]. Let $\tau^{(i)}$ denote the final stage for hypothesis i. Hence, $\tau^{(i)} < k$ if hypothesis i was accepted early and $\tau^{(i)} = k$ if it reached the final stage. Assume that hypothesis i stopped at stage $\tau^{(i)} = t^{(i)}$ and the p-value $\tilde{p}_{t^{(i)}}^{(i)}$ is observed. Then the sequential p-value is defined by

$$\tilde{p}^{(i)} = P_{H_0^{(i)}} \left[\bigcap_{l=1}^{t^{(i)}-1} \{\tilde{P}_l^{(i)} \leqslant \tilde{\gamma}_l\} \cap \{\tilde{P}_{t^{(i)}}^{(i)} \leqslant \tilde{p}_{t^{(i)}}^{(i)}\} \right] \tag{5}$$

The probabilities in (4) and (5) can be calculated with standard software for group sequential tests as, e.g. the R-package 'seqmon' [15]. Note that this p-value has the following monotonicity property: $\tilde{p}^{(i)}$ lies in the interval $(g_t, g_{t-1}]$ if and only if the test for hypothesis i stopped for futility at stage t. Additionally, this p-value does not depend on sample sizes beyond the observed stopping stage [13, p. 181].

To control the FWER of the experiment, in the final analysis we reject all hypotheses whose sequential p-values fall below α/m_1. To control the FDR, we reject all null hypotheses whose sequential p-values $\tilde{p}^{(i)}$ fall below a critical value $\tilde{\gamma}$, where $\tilde{\gamma}$ is the maximal critical value such

DOI: 10.1002/sim

that

$$\frac{\hat{\pi}_1 \tilde{\gamma} m_1}{\max(\#\{\tilde{p}^{(i)} \leqslant \tilde{\gamma}\}, 1)} \leqslant \alpha \tag{6}$$

and $\hat{\pi}_1$ is defined as in (2) with $p_1^{(i)}$ replaced by $\tilde{p}^{(i)}$. Under the null hypothesis $H_0^{(i)}$, the sequential p-values are uniformly distributed [14, 16] and the p-values are stochastically independent across hypotheses because of the independence of observations across hypotheses. Thus, by Theorem 3 in [11] the procedure controls the FDR.

3. DESIGNS WITH A FIXED TOTAL NUMBER OF OBSERVATIONS

We now consider designs where the total number of observations across stages and hypotheses, N, is fixed and thus the stage-wise per-hypothesis samples sizes are random (as in [3, 5]). For example, N could denote the total number of genotyping or the total costs. At each stage a (predetermined) fraction r_t ($\sum_{t=1}^{k} r_t = 1$) of the total number of observations N is distributed among all hypotheses selected for this stage. Thus, given a hypothesis has not been dropped before the tth stage, its sample size at stage t is given by $n_t = r_t N / m_t$, with m_t denoting the number of hypotheses selected for stage t. Since the m_t are stochastic, this results in random sample sizes $n_2, \ldots, n_k$ (for simplicity, we use the same notation as in the case of deterministic stage-wise per-hypothesis sample sizes). Only the first-stage sample size is deterministic as m_1 is prefixed. For the pilot designs, this has no impact on the type I error rate since the p-values at each stage are evaluated separately and the stage-wise p-values are uniformly distributed under the null and independent across hypotheses. Hence, the proposed procedures to control the FDR and the FWER can still be directly applied. For the integrated design with sample sizes given by $n_t = r_t N / m_t$ the arguments are more involved. Assume that the stopping boundaries $\tilde{\gamma}_t$ and the sequential p-values $\tilde{p}^{(i)}$ are computed as in (4) and (5) plugging in the observed sample sizes (ignoring the fact that they are random). In Theorem 1 in Appendix A, we prove that the resulting sequential p-values corresponding to the true null hypothesis are still independent and uniformly distributed. Thus in applying the Bonferroni-corrected significance levels to the sequential p-values, the FWER is controlled. Additionally, by Theorem 3 in [11] the procedure of Storey applied to the sequential p-values controls the FDR.

In case $n_t = r_t N / m_t$ is not an integer, we propose to round first downwards and to distribute the remaining observations across randomly selected hypotheses.

4. OPTIMAL DESIGN PARAMETERS FOR DESIGNS WITH A FIXED TOTAL NUMBER OF OBSERVATIONS

Given a fixed total number of observations N, the design parameters of multi-stage designs are the fractions of observations r_t to be spent at each stage as well as the futility bounds γ_t (for the pilot design) or the continuation probabilities g_t (for the integrated design). Again let $\pi_1 < 1$ denote the proportion of true null hypotheses among the m_1 considered hypotheses. For the $(1-\pi_1)m_1$ remaining hypotheses, we assume a common effect size of Δ/σ. We aim to optimize the individual power, which is the probability for an alternative hypothesis i to be selected and rejected. Since

DOI: 10.1002/sim

we assumed the same effect size for all alternative hypotheses, the individual power is identical for all alternative hypotheses. Additionally, the individual power is equal to the expected proportion of alternative hypotheses that are rejected. Although the exact power is difficult to compute, we can derive asymptotic expressions for the individual power, which we then optimize in the design parameters.

4.1. Optimal pilot designs

Let $\gamma_1,\ldots,\gamma_{k-1}$ and $r_1,\ldots,r_k$ be fixed and assume that the (for $t\geqslant 2$ stochastic) stage-wise per-hypothesis sample sizes are given by $n_t=r_tN/m_t$. We derive asymptotic expressions for the individual power letting $m_1\to\infty$ and assuming that $N=dm_1$ for some $d>0$. We show by induction that the stage-wise per-hypothesis sample sizes n_t and the probabilities

$$\Pi_j=P_{\mu^{(i)}=\Delta/\sigma}(P_j^{(i)}\leqslant\gamma_j)=1-\Phi(c_{1-\gamma_j}-\sqrt{n_j}\Delta/\sigma)$$

to select an alternative hypothesis for stage $j+1$ given that it reached stage j converge almost surely and derive their limits. Here, c_β denotes the β-quantile of the standard normal distribution. First note that $n_1=r_1N/m_1=r_1d$ and Π_1 are deterministic. Assume now for $j=1,\ldots,t$ that $n_j\to\bar{n}_j$ (where $\to$ denotes almost sure convergence for $m_1\to\infty$ and $N=dm_1$) and $\Pi_j\to\bar{\Pi}_j$ where

$$\bar{\Pi}_j=1-\Phi(c_{1-\gamma_j}-\sqrt{\bar{n}_j}\Delta/\sigma) \tag{7}$$

Then, using the strong law of large numbers, we have $m_{t+1}/m_1\to\pi_1\prod_{j=1}^t\gamma_j+(1-\pi_1)\prod_{j=1}^t\bar{\Pi}_j$. Thus,

$$n_{t+1}=\frac{r_{t+1}N}{m_{t+1}}\to\bar{n}_{t+1}:=\frac{r_{t+1}d}{\pi_1\prod_{j=1}^t\gamma_j+(1-\pi_1)\prod_{j=1}^t\bar{\Pi}_j} \tag{8}$$

and consequently also $\Pi_{t+1}\to\bar{\Pi}_{t+1}$, where $\bar{\Pi}_{t+1}$ is defined in (7). Then, asymptotically the probability for an alternative hypothesis to reach the final stage is $\prod_{t=1}^{k-1}\bar{\Pi}_t$.

Control of the FDR: The proportion of true null hypotheses in stage k converges almost surely to $\pi_k=\pi_1\prod_{j=1}^{k-1}\gamma_j/[\pi_1\prod_{j=1}^{k-1}\gamma_j+(1-\pi_1)\prod_{j=1}^{k-1}\bar{\Pi}_j]$. For $\lambda<1$ the estimator $\hat{\pi}_k$ (defined as in (2)) is biased and converges almost surely to, say, $\pi_k(\lambda)\geqslant\pi_k$. Since $\lim_{\lambda\to 1}\pi_k(\lambda)=\pi_k$, the critical value for the final analysis γ_k asymptotically (letting $m_1\to\infty$, $N=d\,m_1$ and then $\lambda\to 1$) satisfies the equation [11]

$$\alpha=\frac{\pi_k\gamma_k}{\pi_k\gamma_k+(1-\pi_k)P_{\mu^{(i)}=\Delta/\sigma}(p_k\leqslant\gamma_k)}=\frac{\pi_1\prod_{j=1}^k\gamma_j}{\pi_1\prod_{j=1}^k\gamma_j+(1-\pi_1)\prod_{j=1}^k\bar{\Pi}_j} \tag{9}$$

Note that γ_k enters the equation also via $\bar{\Pi}_k$ defined in (7). The individual power is asymptotically given by $\bar{\Pi}=\prod_{t=1}^k\bar{\Pi}_t$, where γ_k solves (9).

Control of the FWER: To get an approximate expression for the procedure that controls the FWER, we set $\gamma_k=\alpha/(m_1\prod_{t=1}^{k-1}\bar{\Pi}_t)$.

To obtain optimal design parameters, we numerically optimize the $\bar{\Pi}$ with respect to $\gamma_1,\ldots,\gamma_{k-1}$ and $r_1,\ldots,r_k$. It is easy to see that $\bar{\Pi}$ and consequently also the optimal parameters depend on N, m_1, Δ and σ only via $\sqrt{N/m_1}\Delta/\sigma$ and for the FWER additionally on m_1 (via the rejection boundary).

DOI: 10.1002/sim

4.2. Optimal integrated designs

Analogous to the pilot design, we derive an approximation for the individual power of the integrated design letting $m_1 \to \infty$ and assuming that $N = dm_1$ for some $d > 0$. First let $g_1, \ldots, g_{k-1}$ and $r_1, \ldots, r_k$ be fixed and assume that the (for $t \geqslant 2$ stochastic) stage-wise per-hypothesis sample sizes are given by $n_t = r_t N / m_t$. $P_{\mu^{(i)}=\Delta/\sigma, \bar{n}_1, \ldots, \bar{n}_t}\{\tilde{p}^{(i)} \leqslant g_t\}$ is the probability for an alternative hypothesis to reach stage $t+1$ in a group sequential test with fixed stage-wise sample sizes $\bar{n}_1, \ldots, \bar{n}_t$ (this probability depends only on the first t stage-wise sample sizes). By analogous arguments as for the pilot design, we derive limits for the per-hypothesis sample sizes and the proportions of selected hypotheses m_t / m_1 (for $m_1 \to \infty$ and $N = dm_1$)

$$\frac{m_{t+1}}{m_1} \to \pi_1 g_t + (1 - \pi_1)\, P_{\mu^{(i)}=\Delta/\sigma, \bar{n}_1, \ldots, \bar{n}_t}\{\tilde{p}^{(i)} \leqslant g_t\}$$

$$n_{t+1} \to \bar{n}_{t+1} = \frac{r_{t+1} N}{\bar{m}_{t+1}}$$

with $\bar{n}_1 = n_1$. To control the FDR, the critical value γ for the sequential p-value $\tilde{p}^{(i)}$ asymptotically satisfies the equation (additionally letting $\lambda \to 1$):

$$\alpha = \frac{\pi_1 \gamma}{\pi_1 \gamma + (1 - \pi_1) P_{\mu^{(i)}=\Delta/\sigma, \bar{n}_1, \ldots, \bar{n}_k}\{\tilde{p}^{(i)} \leqslant \gamma\}} \tag{10}$$

The asymptotic individual power is then given by $\bar{\Pi} = P_{\mu^{(i)}=\Delta/\sigma, \bar{n}_1, \ldots, \bar{n}_k}\{\tilde{p}^{(i)} \leqslant \gamma\}$, where γ solves (10) to control the FDR. To control the FWER, we approximate the power $\bar{\Pi}$ by setting $\gamma = \alpha / m_1$.

To obtain optimal design parameters, we numerically optimize the objective function $\bar{\Pi}$ with respect to $g_1, \ldots, g_{k-1}$ and $r_1, \ldots, r_k$. As in the pilot design, $\bar{\Pi}$ and the optimal parameters depend on N, m_1, Δ and σ only via $\sqrt{N/m_1}\, \Delta/\sigma$ and for the FWER additionally on m_1 (via the rejection boundary).

4.3. Examples of optimized experiments

Here, we present numerical optimization results for integrated and pilot designs for different scenarios. Consider a test for $m_1 = 5000$ hypotheses with a total of $N = 8m_1 = 40\,000$ observations. At each stage, a predetermined fraction of the total observations is equally allocated to the selected hypotheses (see Section 3). Assuming that $\sigma^2 = 1$ and that for 50 of the hypotheses the alternative $\Delta = 1$ holds (i.e. $\pi_1 = 0.99$), we computed the optimal r_t and γ_t (resp. g_t) for pilot or integrated multi-stage designs with up to four stages either controlling the FDR or the FWER at level $\alpha = 0.05$. Table I lists the asymptotically optimal design parameters for the different designs and shows the obtained asymptotic power values. The optimal design parameters apply to all scenarios with $\sqrt{N/m_1}\Delta/\sigma = \sqrt{8}$ and $\pi_1 = 0.99$ for FDR controlling procedures. Given $N = 40\,000$ and $m_1 = 5000$, the per-hypothesis sample sizes are not integer but using designs with stage-wise integer sample size (first rounded downwards and randomly choosing hypotheses where the rounded sample size is increased by one) yields practically identical power values within the simulation error.

For all types of designs, the power increases with the number of stages. The advantage of the integrated design over the pilot design is moderate but increases with the number of stages. In contrast, the power difference between procedures for the control of the FWER *versus* the control of the FDR decreases with the number of stages: For the four-stage design, it makes hardly

DOI: 10.1002/sim

Table I. Optimal power values and design parameters for pilot and integrated (int) designs controlling the FDR and the FWER for $N=40000$, $m_1=5000$, $\pi_1=0.99$, $\Delta/\sigma=1$ and $\alpha=0.05$.

Stages	Error rate	Design	Power in per cent	r_1 in per cent	r_2 in per cent	r_3 in per cent	r_4 in per cent	γ_1 or g_1	γ_2 or g_2	γ_3 or g_3
1	FWER		7.5	1						
1	FDR		18.9	1						
2	FWER	Pilot	79.1	68.6	31.4			0.073		
2	FWER	Int	80.1	69.3	30.7			0.077		
2	FDR	Pilot	84.7	68.1	31.9			0.112		
2	FDR	Int	85.9	68.7	31.3			0.123		
3	FWER	Pilot	88.5	56.1	30.2	13.7		0.223	0.098	
3	FWER	Int	90.9	57.1	28.6	14.3		0.276	0.036	
3	FDR	Pilot	90.3	56.3	30.5	13.2		0.255	0.147	
3	FDR	Int	92.8	56.9	28.3	14.8		0.330	0.065	
4	FWER	Pilot	90.2	50.0	28.5	13.2	8.3	0.308	0.236	0.121
4	FWER	Int	93.8	49.8	25.6	15.4	9.2	0.441	0.150	0.022
4	FDR	Pilot	90.8	51.1	28.7	13.6	6.6	0.323	0.266	0.178
4	FDR	Int	94.8	50.2	24.7	15.9	9.2	0.484	0.194	0.043

The optimizations were performed with the function 'optim' in the R-program [15] using the method 'L-BFGS-B' of Byrd *et al.* [17].

any difference if one controls the FDR or the FWER. The optimal design parameters are fairly equal. For all scenarios, the optimized per-hypothesis sample size in the first stage is rather small and increases in the later stages. The expected number of hypotheses under observation clearly decreases with increasing stage.

These qualitative features persist also for other scenarios. Figure 1 shows optimal power values for $\pi_1=\{0.99, 0.97\}$, $\Delta=\{0.5, 1\}$ and experiments with 1–4 stages. Clearly, the single-stage power is the same for pilot and integrated designs. To assess finite sample properties, we performed a simulation study using asymptotically optimal design parameters and the corresponding rounded sample sizes as specified in Section 3. The simulated power is practically identical to the asymptotic power (10^5 simulation runs).

4.4. Robustness of multi-stage designs

The optimal design parameters depend on Δ/σ and π_1, which are typically unknown. We investigated the robustness of the design with respect to misspecifications in the planning phase. Consider an optimal three-stage design for $\Delta/\sigma=0.5$ and $\pi_1=0.99$. The lines in Figure 2(a) and (b) show the impact for the power values if Δ/σ deviates from these assumptions. For this purpose, the difference between the optimal power that could be obtained if the trial would have been planned based on the actual values of Δ/σ and π_1 of each scenario and the non-optimal design (applying the parameters optimal for $\Delta/\sigma=0.5$) is shown. The plots (c) and (d) in Figure 2 show the deviation between optimal and non-optimal power values for the optimal design for deviations of π_1. The plots indicate that designs controlling the FDR are more robust than those controlling the FWER, particularly for decreasing π_1. Additionally, deviating from the optimal values of $r_1, r_2, \ldots$ has only a small impact on the integrated design controlling the FDR, a larger impact on the integrated

DOI: 10.1002/sim

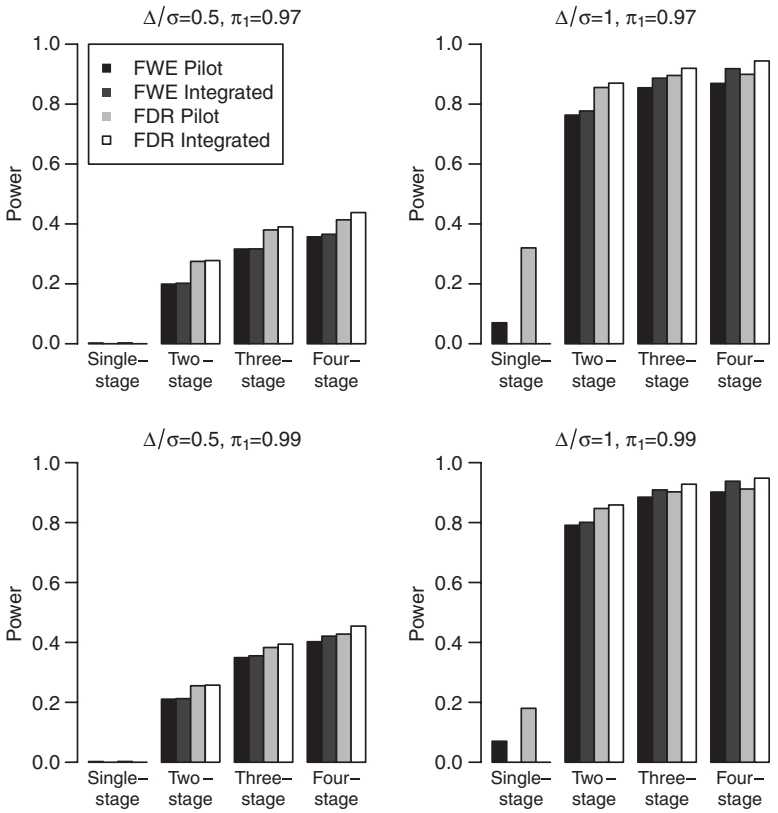

Figure 1. Asymptotically optimal power values for experiments with 1–4 stages. Results for the pilot and the integrated designs controlling the FWER and the FDR are given for $N=40\,000$, $m_1=5000$ and $\alpha=0.05$.

design controlling the FWER but causes a striking decrease in power for the pilot designs (data not shown).

4.5. Extension to the case of unknown variance and two-sided hypotheses

If the variance is unknown, the group sequential t-test can be applied. However, the exact computations of group sequential p-values for the t-test are numerically difficult. As an alternative we propose an approximate procedure using the critical values from the model with known variances in Section 2 applied to the p-values of the t-test at each stage [18].

For the case of two-sided hypotheses, sequential p-values are computed as in (5) with $\tilde{p}^{(i)}_{t^{(i)}}$ replaced by the respective two-sided p-value. Similarly, the critical boundary is calculated from equation (4) adapted to the two-sided case.

4.6. Real data application

We emulate a two- and three-stage design based on a data set from an experiment by Tian *et al.* [19], resulting from the comparison of gene expression measurements of 36 patients with

DOI: 10.1002/sim

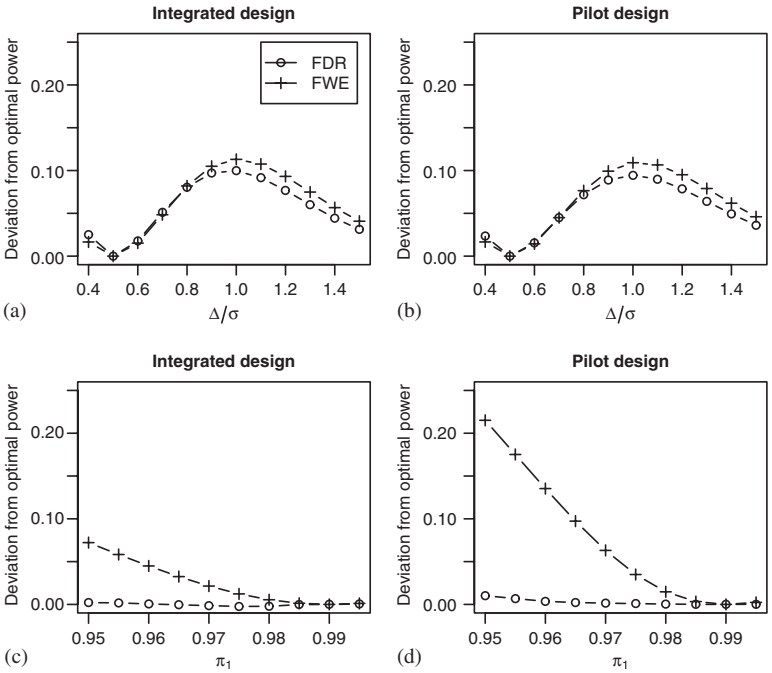

Figure 2. Deviation from optimal power. The differences between optimal and non-optimal designs are given for pilot and integrated designs controlling the FDR or the FWER, respectively, for $N=40\,000$, $m_1=5000$ and $\alpha=0.05$ and varying values of π_1 and Δ/σ. For (a) and (b) for each scenario the optimal parameters for $\Delta/\sigma=0.5$, $\pi_1=0.99$ are applied, for (c) and (d) the optimal parameters for $\Delta/\sigma=1$, $\pi_1=0.99$.

bone lytic lesions with a control group of equal size without such lesions. The original data were generated with Affymetrix Human U95A chips, each containing 12 625 probe sets. We used the post-processed data published by Jeffery *et al.* [20].

In the re-analysis, we assumed that the total number of gene expression measurements is limited to only $8\times 12\,625$ measurements per group. Thus, in a single-stage design, the gene expressions from all 12 625 probe sets would be available for 8 patients. Performing a two-sided two-sample t-tests controlling either the FWER or the FDR at level $\alpha=0.05$, this single-stage analysis showed no significantly different gene expression measurements between the two groups. Below we compare this single-stage test with multi-stage procedures (also two-sided).

4.6.1. Two-stage design. We consider a two-stage pilot design with parameters $r_1=0.66$ and $\gamma_1=0.1$ and an integrated design with $r_1=0.66$ and $\tilde{\gamma}_1=0.1$. Thus, we have equal sample sizes and selected probe sets for the pilot and the integrated design. We include 5 patients per group in the first stage. For each of these patients, all $m_1=12\,625$ probe set gene expression measurements are taken. Thus, $r_1=0.625$. For the second stage, only probe sets with a two-sided p-value below $\tilde{\gamma}_1=\gamma_1=0.1$ are selected. In this data set, these are $m_2=1502$ probe sets. The number of patients at the second stage is given by $[(8-5)\times 12\,625]/1502=25.2$. We assume that for (rounded off) 25 patients per group a microarray chip with the 1502 selected probe set was designed. In the

DOI: 10.1002/sim

Table II. Number of significant probe sets for a two-stage and a three-stage design for the integrated and the pilot designs controlling the FWER and the FDR with $\alpha=0.05$.

# Stages	Pilot FWER	Pilot FDR	Integrated FWER	Integrated FDR
2	2	74	14	107
3	4	52	26	195

For the two-stage design, $r_1=0.625$ and $\gamma_1=\tilde{\gamma}_1=0.1$; for the integrated three-stage design $r_1=0.5$, $r_2=0.3$, $\tilde{\gamma}_1=g_1=0.25$, $g_2=0.03$, whereas for the pilot design $\gamma_1=0.25$ and $\gamma_2=0.13$.

re-analysis, we use the measurements of the selected probe sets taken from the full arrays (discarding the measurements from all other probe sets). Table II shows the number of rejections for the pilot and integrated designs controlling the FWER and the FDR.

4.6.2. Three-stage design. We consider a three-stage pilot design with parameters $r_1=0.5$, $\gamma_1=0.25, r_2=0.3$ and $\gamma_2=0.13$ and an integrated design with parameters $r_1=0.5, g_1=\tilde{\gamma}_1=0.25$, $r_2=0.3$ and $g_2=0.03$. In the first stage, the sample size and selected probe sets are the same for the pilot and integrated designs. This also holds for the second but not for the third stage. In the first stage, 4 patients are included where all $m_1=12\,625$ expression measurements are taken for each patient. For $m_2=3297$ probe sets, the respective first-stage p-value falls below $\gamma_1=\tilde{\gamma}_1$. In the second stage for $n_2=0.3\times 8\times 12\,625/m_2=9$ (which is rounded off) patients per group, a chip with the selected genes is designed. The algorithm to select probe sets for the third stage is different for the integrated and the pilot designs. In the pilot design, p-values are calculated only from the second-stage data. In the example, $m_3=883$ p-values fall below $\gamma_2=0.13$, which results in a third-stage sample size of $n_3=22$ patients. For the integrated design, we first compute the critical value $\tilde{\gamma}_2$. This gives $\tilde{\gamma}_2=0.053$. In the integrated design, all probe sets whose p-value calculated from the first and second stage falls below $\tilde{\gamma}_2$ are selected for the third stage. In this scenario, $m_3=916$, which also results in $n_3=22$ patients for the third stage.

It can be seen from Table II that for this example the integrated design appears to have a much larger power than the pilot design. In addition, the difference between the control of the FDR and that of the FWER is pronounced. Finally, the three-stage design gives a distinct improvement over the two-stage design. Clearly, the effect size and the proportion of true null hypotheses are unknown in the real data set. Yet, calculations in Section 4.4. showed that designs are rather robust against such misspecifications in these parameters. However, the misspecifications for this example are different: Due to the restricted number of patients in the real data set from the literature (36 per group), we could not perform an optimization as in Section 4 and the design parameters r_1, r_2, r_3, γ_1 and γ_2 were chosen so that the sample sizes in the second and third stages did not exceed the available number of observations in the data set. Asymptotic numerical power calculations show a clear superiority of the integrated design controlling the FDR compared with the pilot design and compared with the FWE controlling procedure when choosing design parameters that lead to too small sample sizes in the last stage of the trial. In contrast if, e.g. in the two-stage design a smaller value of r_1 is chosen, the differences between the integrated and the pilot designs as well as the difference between designs controlling the FDR and the FWER decrease. This effect also occurs, if γ_1 is decreased. Additionally, for a very large number of hypotheses m_1, the discrepancy between procedures controlling the FDR and the FWER becomes larger because for $m_1\to\infty$ (and N increasing proportionally) the power of the FWER controlling procedure tends to zero. However,

DOI: 10.1002/sim

even for the number of hypotheses as large as $m_1 = 10^5$, the power of the FWER controlling procedures in the optimal design hardly decreases compared with the scenarios in Table I (ca. 3 per cent decrease in power for the two-stage design, 1 per cent for the three-stage design).

5. DISCUSSION

Multi-stage procedures are of increased interest in gene-expression or gene-association studies since, e.g. the technical equipment for designing individual micro-arrays have become available. For two-stage designs, remarkable improvements in the power to detect influential markers have been shown by different authors. In this paper, we have investigated the statistical properties of multi-stage designs either controlling the FDR or the FWER. Two different concepts of the sequential designs have been considered: In the pilot design, the test decisions are exclusively derived from the sample at the last stage. The previous stages are only used to screen the promising markers that are then tested at the final stage. For the integrated design, the test decisions for the hypotheses selected for the last stage are based on the observations from all stages. The main goal was to construct asymptotically optimal multi-stage designs and investigate their behavior depending on the concept (pilot *versus* integrated design), the type of error control (FDR *versus* FWER), the number of stages and the *a priori* assumptions (proportions of true null hypotheses, effect size under the alternative). Generally, going from two to three stages may lead to a worthwhile increase in power, both for the pilot and the integrated designs and for the control of the FDR or the FWER. As expected, the improvement when going from three to four stages decreases and may not be worthwhile considering the logistic and computational burden arising from a further increase in the number of stages. In all scenarios, the number of hypotheses decreases and the per-hypothesis sample size increases strikingly with increasing stage and the last stage power is nearly 1.

The difference in the statistical properties between the optimal pilot and integrated design is surprisingly small in designs with more than two stages. However, if in the planning phase misspecifications occur so that a non-optimal design is applied for an experiment the loss of power of the pilot design may become noticeably larger than for the integrated design. This is plausible because the integrated design uses information from all stages to make the decision, whereas the decision in the pilot design may be based on a non-optimal last-stage sample size.

This result is emphasized by the real data application. Owing to the lack of knowledge of π_1 and the effect sizes, we generally will not apply the optimal parameters. Here, we found that the integrated design leads to more rejections than the pilot design. In addition, the multi-stage procedure based on the FDR is more robust than that based on the FWER, which here leads to a pronounced increase in power. The basic result that screening leads to a high improvement in the power as compared with conventional single-stage designs can be seen in the application to the real data set.

For optimal designs, we found a further interesting result for increasing number of stages: The difference in power of the optimal designs controlling the FDR or FWER, respectively, becomes smaller (both for the pilot and the integrated designs). The optimal design for controlling the FDR selects more hypotheses for the final stage by applying higher critical boundaries γ_i at earlier stages than the optimal design for controlling the FWER.

We also investigated the case of unknown variances and distributed alternatives as in [5] and in all simulations the average FDR was very close to the nominal level (data not shown). Optimized designs where the total number of observations is constrained and the proportion of the total number of observations allocated to each stage is equally distributed among the selected hypotheses

DOI: 10.1002/sim

(leading to random sample sizes) showed similar power as designs where deterministic stage-wise per-hypothesis sample sizes are applied.

The integrated design can be easily extended to allow for early rejection of null hypotheses [21]. Yet, for the considered scenarios where π_1 is close to one, the possibility of early rejection gave no noticeable improvement in power. Moreover, it makes sense to continue with promising hypotheses to confirm early significant results with a larger sample.

Under optimal conditions, our investigations show that the crucial point is not the choice of the error rate, the type of design (integrated or pilot) or the number of stages (three or four stages), but skipping non-promising hypotheses in the early phases of the experiment. Then test decisions among the selected hypotheses can be based on considerably larger sample sizes than in the single-stage design distributing the total number of observations equally among all candidate markers. This is in line with the finding that even randomly dropping hypotheses for a single-stage design may increase the power in case of many hypotheses and limited resources [22].

APPENDIX A

A.1. Proof of Theorem 1

The proof is given in two steps. First, for the case of deterministic stage-wise per-hypothesis sample sizes, we derive the joint distribution of the sequential p-values and a random variable specifying the stage where the hypothesis is dropped. Then (Theorem 1), we prove that this distribution is the same in the design with random stage-wise per-hypothesis sample sizes. This is shown via modified designs, where the stage-wise per-hypothesis sample sizes depend only on the outcomes of alternative hypotheses.

Without limitation of generality, we assume that for the first i_0 hypotheses the null hypothesis holds. For the remaining hypotheses, the alternative is true. Let $\tilde{\mathbf{p}}=(\tilde{p}^{(i)})_{i=1}^{i_0}$ denote the vector of sequential p-values corresponding to the true null hypotheses. For a set A we denote the i-ary Cartesian product by A^i and the characteristic function by $\mathbf{1}_{\{A\}}$.

A.1.1. On the distribution of the sequential p-value for fixed stage-wise per-hypothesis sample sizes. Define the random variables $\tau^{(i)}$ that denote the stage where hypothesis i stops. Set $\tau^{(i)}=k$ if the trial continues to the final stage. Let $\boldsymbol{\tau}=(\tau^{(i)})_{i=1}^{i_0}$.

Lemma 1

In a trial with fixed stage-wise per-hypothesis sample sizes under H_i for all vectors of time points $\mathbf{s}=(s^{(i)})_{i=1}^{i_0}\in\{1,\ldots,k\}^{i_0}$ and all real vectors $\mathbf{q}=(q_i)_{i=1}^{i_0}\in[0,1]^{i_0}$, we have $P(\{\tilde{\mathbf{p}}\leqslant\mathbf{q}\}\cap\{\boldsymbol{\tau}=\mathbf{s}\})=\prod_{i=1}^{i_0}b(q^{(i)},s^{(i)})$, where, for $t=1,\ldots,k$,

$$b(x,t)=\begin{cases}0 & \text{if } x<g_t\\ x-g_t & \text{if } x\in(g_t,g_{t-1}]\\ g_{t-1}-g_t & \text{if } x>g_{t-1}\end{cases}$$

Proof

By (5) $\{\tau^{(i)}=t\}=\{\tilde{p}^{(i)}\in(g_t,g_{t-1}]\}$. Since the $\tilde{p}^{(i)}$ are uniformly distributed, $P(\tilde{p}^{(i)}\leqslant x,\tau^{(i)}=t)=b(x,t)$. Since the p-values are independent, the result follows. □

DOI: 10.1002/sim

A.1.2. The case of fixed overall number of observations and random per-hypothesis sample sizes. Let $(\Omega, \mathscr{P})$ denote a probability space and define an m_1-dimensional stochastic process $(n_t, z_t^{(i)})_{i=1}^{m_1}$, $t=1,\ldots,k$, where n_t denotes the per-hypothesis sample size in stage t and $z_t^{(i)}=s_t^{(i)}/(\sigma^{(i)}\sqrt{n_t})$ the z-statistic of hypothesis i computed from the observations at stage t. Let $(\mathscr{F}_t)_{t=1}^k$ denote the filtration generated by $(n_t, z_t^{(i)})_{i=1}^{m_1}, t=1,\ldots,k$. For $t=1$ the number of considered hypotheses m_1 and consequently $n_1=r_1 N/m_1$ and by (4) $\tilde{\gamma}_1$ are constants. Consequently, the $z_1^{(i)}$ are independent and $\mathrm{N}(\mu_i\sqrt{n_1}, 1)$ distributed. For $t>1$ the m_t and n_t are defined recursively by

$$m_t=\sum_{i=1}^{m_1}\min_{j=1,\ldots,t-1}\mathbf{1}_{\{1-\Phi(\sum_{l=1}^{j}z_l^{(i)}\sqrt{n_l/\tilde{n}_j})\leqslant\tilde{\gamma}_j\}}$$

where $\tilde{n}_t=\sum_{j=1}^{t}n_j$ gives the cumulated sample size until stage t. Note that $m_t, n_t=r_t N/m_t$ and by (4) also $\tilde{\gamma}_t$ are $\mathscr{F}_{t-1}$-measurable and the $z_t^{(i)}, i=1,\ldots,m_1$ conditional on $\mathscr{F}_{t-1}$ are independent and $\mathrm{N}(\mu_i\sqrt{n_t}, 1)$ distributed. Define the stopping time for hypothesis i:

$$\tau^{(i)}=\min\left\{t\left|1-\Phi\left(\sum_{l=1}^{t}z_l^{(i)}\sqrt{n_l/\tilde{n}_t}\right)>\tilde{\gamma}_t\right.\right\}$$

where $\tau^{(i)}=k$ if no stopping boundary is crossed and set $\boldsymbol{\tau}=(\tau^{(i)})_{i=1}^{i_0}$.

Theorem 1

Assume that the data are independently distributed across hypotheses. In the integrated design with fixed overall number of observations (and random per-hypothesis sample sizes), the sequential p-values $\tilde{p}^{(i)}$, $i=1,\ldots,i_0$, corresponding to true null hypotheses are independent and uniformly distributed.

Proof

We show that in the experiment with random per-hypothesis sample sizes for all $\mathbf{s}\in\{1,\ldots,k\}^{i_0}$ and $\mathbf{q}\in[0,1]^{i_0}$

$$P(\{\tilde{\mathbf{p}}\leqslant\mathbf{q}\}\cap\{\boldsymbol{\tau}=\mathbf{s}\})=\prod_{i=1}^{i_0}b(q^{(i)}, s^{(i)}) \tag{A1}$$

By Lemma 1 this implies that the joint distribution of the p-values and stopping times of true null hypotheses is the same as in the case of deterministic sample sizes. Hence, this holds also for the joined distribution of the p-values alone. This proves the theorem. To prove (A1) we will define a modified experiment where the sample sizes are independent of the outcomes from the data of the true null hypotheses.

As above, let $\mathbf{s}\in\{1,\ldots,k\}^{i_0}$ denote a realization of $\boldsymbol{\tau}$ and define the stopping time τ_s to be the first time point t where for *any* true null hypothesis i, the stopping indicator $\mathbf{1}_{\{\tau^{(i)}\geqslant t\}}$ differs from the stopping indicator $\mathbf{1}_{\{s^{(i)}\geqslant t\}}$ (corresponding to stopping at stage $s^{(i)}$). More formally,

$$\tau_s=\min\left\{t\left|\max_{i=1,\ldots,i_0}|\mathbf{1}_{\{\tau^{(i)}\geqslant t\}}-\mathbf{1}_{\{s^{(i)}\geqslant t\}}|\neq 0\right.\right\}$$

where $\tau_s=k$ if $|\mathbf{1}_{\{\tau(i)\geqslant t\}}-\mathbf{1}_{\{s^{(i)}\geqslant t\}}|=0$ for all t and i. By definition, τ_s is a stopping time.

DOI: 10.1002/sim

To specify the modified experiment consider an m_1-dimensional stochastic process $(n'_t, z'^{(i)}_t)_{i=1}^{m_1}$, $t=1,\ldots,k$, defined as follows: for all $t\leqslant\tau_s$ let $(n'_t, z'^{(i)}_t)_{i=1}^{m_1}=(n_t, z^{(i)}_t)_{i=1}^{m_1}$. For $t>\tau_s$ we set

$$n'_t=\frac{r_t N}{m'^A_t+\sum_{i=1}^{i_0}\mathbf{1}_{\{s^{(i)}\geqslant t\}}} \tag{A2}$$

where m'^A_t denotes the number of alternative hypotheses that have been continued to stage t in the modified experiment. For $t>\tau_s$, $z'^{(i)}_t$ are defined as the standardized stage-wise means of n'_t-independent observations for each continued hypothesis. Hence, the sample sizes are determined as if the stopping times for the true null hypotheses were equal to $s^{(1)},\ldots,s^{(i_0)}$. Note that actually also for $t\leqslant\tau_s$ n'_t is given by (A2). This follows, since $n_t=r_t N/(m^A_t+m^N_t)$ and for $t\leqslant\tau_s$ we have $m^A_t=m'^A_t$ and $m^N_t=\sum_{i=1}^{i_0}\mathbf{1}_{\{s^{(i)}\geqslant t\}}$. Here m^A_t and m^N_t denote the number of alternative and true null hypotheses in the original experiment that are continued to stage t. Thus, in the modified experiment the sample sizes are defined by (A2) for all $t=1,\ldots,k$. Hence, in the modified experiment, the sample sizes are independent of the z-statistics of the true null hypotheses.

Let $\tilde{p}'^{(i)}$ denote the sequential p-values in the modified experiment and let $\tau'^{(i)}$ denote the corresponding stopping times, i.e. $\tau'^{(i)}$ denotes the stopping stage for hypothesis i in the modified experiment. The sequential p-values $\tilde{p}^{(i)}$ and $\tilde{p}'^{(i)}$ are $\mathscr{F}_{\tau^{(i)}}$ and $\mathscr{F}_{\tau'^{(i)}}$ measurable and $\{\boldsymbol{\tau}=\mathbf{s}\}$ is F_{τ_s}-measurable. Thus, since for $t\leqslant\tau_s$ the modified and the original stochastic processes are identical, it follows that $(\tilde{p}^{(i)})_{i=1}^{i_0}\mathbf{1}_{\{\boldsymbol{\tau}=\mathbf{s}\}}$ and $(\tilde{p}'^{(i)})_{i=1}^{i_0}\mathbf{1}_{\{\boldsymbol{\tau}=\mathbf{s}\}}$ are $\mathscr{F}_{\tau_s}$ measurable. Now, again since for $t\leqslant\tau_s$, the modified and the original stochastic processes are identical and since $\{\boldsymbol{\tau}=\mathbf{s}\}=\{\boldsymbol{\tau}'=\mathbf{s}\}$ it follows that

$$P(\{\tilde{\mathbf{p}}\leqslant\mathbf{q}\}\cap\{\boldsymbol{\tau}=\mathbf{s}\})=P(\{\tilde{\mathbf{p}}'\leqslant\mathbf{q}\}\cap\{\boldsymbol{\tau}=\mathbf{s}\}) \tag{A3}$$

Since in the modified experiment the sample sizes and z-statistics corresponding to true null hypotheses are independent, the joined distribution of the p-values corresponding to true null hypothesis is in the modified experiment the same as in the experiment with deterministic per-hypothesis sample sizes (as derived in Lemma 1). Hence,

$$P(\{\tilde{\mathbf{p}}'\leqslant\mathbf{q}\}\cap\{\boldsymbol{\tau}=\mathbf{s}\})=\prod_{i=1}^{i_0}b(q^{(i)},s^{(i)})$$

Together with (A3) this implies (A1). □

ACKNOWLEDGEMENTS

The authors thank Werner Brannath and the referees for their valuable suggestions.

REFERENCES

1. Bukszar J, Van den Oord E. Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for Pearson's statistic. *Biometrics* 2006; **62**:1132–1137.
2. Ohashi J, Clark AG. Application of the stepwise focusing method to optimize the cost-effectiveness of genome-wide association studies with limited research budgets for genotyping and phenotyping. *Annals of Human Genetics* 2005; **69**:323–328.

DOI: 10.1002/sim

3. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association studies. *Biometrics* 2002; **58**:163–170.
4. Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology* 2003; **25**:149–157.
5. Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* 2005; **21**:3771–3777.
6. Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for genetic case-control association studies. *Statistics in Medicine* 2006; **25**:3134–3149.
7. Goll A, Bauer P. Two-stage designs applying methods differing in costs. *Bioinformatics* 2007; **23**:1519–1526.
8. Mueller HH, Pahl R, Schaefer H. Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genetic Epidemiology* 2007; **31**:844–852.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
10. Storey JD. A direct approach to false discovery rate. *Journal of the Royal Statistical Society*, *Series B* 2002; **64**:479–498.
11. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society*, *Series B* 2004; **66**:187–205.
12. Van den Oord EJ, Sullivan PF. A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. *Human Heredity* 2003; **56**:188–199.
13. Jennison C, Turnbull B. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: London, Boca Raton, 2000.
14. Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**:797–803.
15. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2005. ISBN: 3-900051-07-0.
16. Brannath W, Bauer P, Posch M. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
17. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 1995; **16**:1190–1208.
18. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
19. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD. The role of the Wnt-signaling antagonist DKKI in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine* 2003; **349**:2483–2494.
20. Jeffery IB, Higgins DG, Culhance AC. Comparison and evaluation of methods for generating differentially expressed genes lists from microarray data. *BMC Bioinformatics* 2006; **7**:359–375.
21. Victor A, Hommel G. Combining adaptive designs with control of the false discovery rate—a generalized definition for a global p-value. *Biometrical Journal* 2007; **49**:94–106.
22. Futschik A, Posch M. On the optimum number of hypotheses to test when the number of observations is limited. *Statistica Sinica* 2005; **15**:841–855.

DOI: 10.1002/sim