BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 12 2007, pages 1519–1526
doi:10.1093/bioinformatics/btm140

Genetics and population analysis

Two-stage designs applying methods differing in costs

Alexandra Goll* and Peter Bauer

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

Received on February 9, 2007; revised on March 23, 2007; accepted on April 5, 2007

Advance Access publication April 26, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Two-stage pilot and integrated designs are powerful tools for investigating large numbers of hypotheses. Asymptotically, optimal two-stage designs controlling the familywise error or false discovery rate are considered when costs and effect sizes per measurement differ between stages and total costs are constrained.

Results: Depending on the cost and effect size ratios between the measurements, it is generally more powerful to apply two-stage procedures using one measurement method at both stages. For the practically relevant case that the same method is applied at both stages but designing the second-stage measurements raises extra costs, two-stage designs are more powerful than the single-stage design even for large costs ratios. The power of the optimal pilot and integrated two-stage designs generally are similar, however, the integrated approach is less sensitive even to severe design misspecifications in the planning phase.

Availability: R-programs (R, 2005) to calculate asymptotically optimal designs are available on: http://statistics.msi.meduniwien.ac.at/index.php?page=ao2stage

Contact: alexandra.goll@meduniwien.ac.at

1 INTRODUCTION

In gene expression and proteomic studies, we generally deal with large numbers of hypotheses, where only for a small fraction of the hypotheses noticeable effects exist. Due to limited resources, the number of observations per hypotheses in a conventional single-stage design is low which limits the power. It has been shown that two-stage (or multi-stage) designs are a good option to improve the power. In these sequential designs, early stages are used to screen for the promising hypotheses, which are further investigated in later stages. For example, Zehetmayer *et al*. (2005) proposed (optimal) two-stage designs for experiments with a large number of hypotheses and constraints on the total sample size which control the false discovery rate (FDR, see Benjamini and Hochberg, 1995). All hypotheses whose conventional univariate first-stage *P*-values fall below a certain common threshold are selected for the second stage. The final test decision is based on the observations pooled over both stages (‘integrated design’), see also Bukszar and Van den Oord (2006), Satagopan and Elston (2003), Satagopan *et al*. (2002), Satagopan *et al*. (2004), Van den Oord and Sullivan (2003), Zehetmayer *et al*. (2005) also investigated optimal ‘pilot designs’, where the final test is only based on the second-stage data. Further comparisons between the pilot and the integrated design can also be seen in Skol *et al*. 2006. In all these proposals, constant costs and effect sizes over stages have been assumed.

In the following, we investigate two-stage designs using a less accurate assay in early stages and more accurate ones in later stages for cost reasons (see also Wang *et al*., 2006). For example, a quasi-quantitative, global LC-MS profiling proteomics experiment may underestimate the true effect size due to saturation or sensitivity effects inherent in these multiplexed assays, whereas a targeted, calibrated assay (e.g. ELISA) can show an effect size generally larger than the profiling study. First, we consider such a scenario that the experimenter from the beginning may have the choice between two methods that differ in costs and effect sizes. In the second scenario, different costs per measurement may arise if the same method is applied at both stages but specific experimental devices have to be produced at higher costs per measurement for the selected markers at the second stage. In contrast to Wang *et al*. (2006) who constructed designs minimizing the overall costs for a given FWE rate and power, we assume that the total costs of the experiment are fixed, similar to Satagopan *et al*. (2002), Zehetmayer *et al*. (2005) or Ohashi and Clark, (2005). For limited total costs, we derive both integrated and pilot designs with an asymptotically optimal power (for an increasing number of null hypotheses), either controlling the FWE rate or the FDR. The test problem is defined in Section 2 and the corresponding single-stage procedures in Section 3. In Sections 4 and 5, we define the asymptotically optimal pilot and integrated design. In Section 6, we show for the first scenario that depending on the cost and the effect ratios between the methods it is preferable either to apply the low-cost or the high-cost method on both stages. The second scenario is investigated in Section 7 calculating cost ratios between stages for which it is worthwhile to use (optimal) two-stage designs. We further look how design misspecifications in the planning phase would change the power of two-stage designs as compared to the standard single-stage design. A short discussion including some results under less stringent distributional assumptions is given in Section 8.

2 TEST PROBLEM

Consider m_1 (null) hypotheses for the mean of independent normally distributed observations with known variance:

$H_{0i} : \mu_i = 0$ against $H_{1i} : \mu_i > 0$, $i = 1, ..., m_1$.

For deriving the test procedures, we assume independence of observations across hypotheses.

*To whom correspondence should be addressed.

3 THE SINGLE-STAGE DESIGN

We assume that there is a limit on the overall total costs C of the study. Without loss of generality, the costs per observation of the single-stage design are set to 1. In the standard single-stage design, we equally allocate $n = C/m_1$ observations to each of the m_1 hypotheses. The test statistics used for decisions are the P-values $p_i = 1 - \Phi(z_i),\ i = 1, \ldots, m_1$, where z_i is the standardized mean of the sample taken to test H_{0i} and Φ is the distribution function of the standard normal distribution. The P-values are compared to a common critical boundary γ: If $p_i < \gamma$ the null hypothesis H_{0i} is rejected, otherwise it is accepted. We further assume that for a fraction π_0 of the m_1 hypotheses considered the null hypothesis is true. To simplify later calculations, we also assume that the same mean $\mu_i = \Delta\sigma$ holds true for all the alternatives, where σ^2 is the common known variance.

To control the FWE rate (the probability to reject at least one true null hypothesis irrespective of how many and which are in fact true), we apply the critical Bonferroni boundary $\gamma = \alpha/m_1$. The power of such a single-stage design is defined by $\prod_s = 1 - \beta(\gamma) = 1 - \Phi_{\sqrt{(C/m_1)}\Delta,1}(c_{1-\gamma})$, where $\beta(\gamma)$ denotes the type 2 error as a function of the rejection boundary γ, Φ_{μ,σ^2} is the distribution function of the normal distribution with mean μ and variance σ^2 and $c_{1-\gamma}$ is the $(1-\gamma)$-quantile of the standard normal distribution. Note that under the assumption of a common alternative, the power is the expected fraction of null hypotheses correctly rejected.

To control the FDR (the expected proportion of erroneous rejections among all rejections), we apply the method of Storey, (2002) estimating the FDR. The critical value γ is determined as the maximal γ such that

$$\frac{\hat{\pi}_0 \gamma m_1}{\max(\sharp\{p_i < \gamma\}, 1)} \leq \alpha. \tag{1}$$

Here, $\hat{\pi}_0$ is the estimated proportion of true null hypotheses given by

$$\hat{\pi}_0 = \sharp\{p_i > \lambda\}/\{(1-\lambda)m_1\}, \tag{2}$$

where $\lambda, 0 < \lambda < 1$, is a constant chosen a priori and $\sharp\{p_i > \xi\}$ denotes the number of P-values exceeding ξ. Hence, the critical boundary is determined from the sample such that the estimated FDR never exceeds the targeted value α. Using the method of Storey the critical boundary is a random variable. Asymptotically, for large m_1, γ can be determined from the equation

$$\alpha = \frac{\pi_0\gamma}{\pi_0\gamma + (1-\pi_0)(1-\beta(\gamma))}$$

and plugged into the formula for $\prod_s$ to approximate the real power.

4 THE PILOT DESIGN

4.1 The test procedure

We consider the same test problem as described in Section 2. Again, we assume there is a limit of overall total costs C for the study. Now, a fraction r of the total costs C is used for the first stage for testing the m_1 hypotheses. Thus, for balanced sample size allocation the sample size of the first stage per hypothesis is $n_1 = rC/m_1$. The first-stage P-values are given by $p_i^{(1)} = 1 - \Phi(z_i^{(1)})$ where $z_i^{(1)}$ is the first-stage mean of the observations for hypothesis H_{0i}, $i = 1, \ldots, m_1$, standardized by using the common known first-stage SD σ_1. All null hypotheses are selected, whose P-values fall below a threshold γ_1 ($p_i^{(1)} < \gamma_1$). All others are accepted. Hence, a random number of m_2 hypotheses are selected for the second stage. Assume the sampling costs vary between the two stages due to applying a high-cost method in the second stage, so that the total costs are $m_1n_1 + m_2n_2c_2 = C$ for some constant $c_2 \geq 1$. The remaining costs $(1-r)C$ are equally allocated over the selected null hypotheses so that the second-stage sample size n_2 is given by $n_2 = (C - m_1n_1)/(m_2c_2) = ((1-r)C)/(m_2c_2)$. Let $z_i^{(2)}$ denote the mean of the second-stage sample for hypothesis H_{0i}, now standardized by using the common known second-stage SD σ_2. Consequently, $p_i^{(2)} = 1 - \Phi(z_i^{(2)})$ denotes the second-stage P-value for the selected null hypothesis H_{0i}. Remember that in the pilot design the P-value used for decisions after the second stage is only calculated from the second-stage sample. A selected hypothesis H_{0i} is rejected if the second-stage P-value falls below the boundary γ_2 ($p_i^{(2)} < \gamma_2$). Otherwise it is accepted.

4.2 Optimal designs controlling the FWE rate

To control the FWE rate, we simply apply the Bonferroni method to determine the rejection boundary for the second-stage P-value $p_i^{(2)}$, but in contrast to the single-stage design, the adjustment refers to the number of selected hypotheses m_2: $\gamma_2 = \alpha/m_2$. Since m_2 is independent of the second-stage data, this procedure clearly controls the FWE rate at the level α.

We now will try to determine a γ_1 and r which maximizes the power of the two-stage design controlling the FWE rate. We assume that at stage 1 for all alternative hypotheses the same mean $\mu_{1i} = \Delta\sigma_1$ and at stage 2 the same mean $\mu_{2i} = k\Delta\sigma_2, k \geq 1$, holds true, respectively. Here, k is the ratio of the effect sizes between the two stages, and we assume that the high-cost method at the second stage never provides a smaller effect size than the low-cost method at stage one. The first-stage power (the probability of being selected) for a true alternative is given by

$$1 - \beta_1(\gamma_1) = P_{\mu_{1i}=\Delta\sigma_1}(p_i^{(1)} < \gamma_1) = 1 - \Phi_{\sqrt{n_1}\Delta,1}(c_{1-\gamma_1}).$$

Note that under the assumption of a common alternative, this is the expected proportion of correctly selected null hypotheses among all null hypotheses for which the alternative holds.

For the second stage we select m_2 hypotheses which, for large m_1, is given by

$$m_2 = m_1(1-\pi_0)(1-\beta_1(\gamma_1)) + m_1\pi_0\gamma_1.$$

Because of the independence between the two stages, the overall power of the pilot design, i.e. the expected fraction of null hypotheses correctly rejected after the second stage, is asymptotically given by

$$\begin{aligned}\prod\nolimits_p &= (1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2)) \\ &= \left(1-\Phi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(c_{1-\gamma_1})\right)\left(1-\Phi_{\sqrt{\frac{(1-r)C}{m_2c_2}}\Delta k,1}(c_{1-\frac{\alpha}{m_2}})\right).\end{aligned} \tag{3}$$

Given an FWE rate α, an initial number of hypotheses m_1, overall costs C, the cost ratio c_2 between stages, the proportion of true null hypotheses π_0, the effect size Δ and the effect size ratio k between stages we can optimize $\prod_p$ in the two design parameters r and γ_1. Considering r as a continuous variable, the optimal sample sizes per stage (n_1 and n_2) in general will be non-integer. It is easy to see that the optimal γ_1 and r depend on C, m_1, Δ and k via $\sqrt{\frac{C}{m_1}}\Delta$ and $k/\sqrt{c_2}$.

4.3 Optimal designs controlling the FDR

To control the FDR, the second-stage critical boundary γ_2 is determined as in formulas 1 and 2 replacing m_1 by m_2. Asymptotically, for large m_1, the first-stage selection boundary γ_1 and the second-stage rejection boundary γ_2 in the pilot design have to adhere to the equation

$$\alpha = \frac{\pi_0\gamma_2\gamma_1}{\pi_0\gamma_2\gamma_1 + (1-\pi_0)(1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2))} \tag{4}$$

where, $(1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2))$ is the power $\prod_p$ of the pilot design defined in (3) using γ_2 instead of α/m_2. Again $\prod_p$ can be optimized as function of r and γ_1, where γ_2 follows from condition (4).

5 THE INTEGRATED DESIGN

5.1 The test procedure

We address the same test problem as in Section 2. Also, the screening step of the test procedure at the first stage is identical to the pilot design in the previous section. The only difference to the pilot design is that the final test decisions based on the selected null hypotheses are derived from integrated P-values $p_i = 1-\Phi(z_i)$ which are based on the data from both stages. An obvious way to construct single combination test statistics z_i from both stages is to combine the stagewise standardized means by suitable weights as applied for adaptive multi-stage clinical trials (e.g. Lehmacher and Wassmer, 1999):

$$z_i = \sqrt{w_1}z_i^{(1)} + \sqrt{1-w_1}z_i^{(2)}. \tag{5}$$

Now the test decision is again very simple: a selected null hypothesis H_{0i} is rejected in the final test if $p_i < \gamma$. Otherwise it is accepted. Optimizing the non-centrality parameter $(\sqrt{w_1}\sqrt{n_1} + \sqrt{1-w_1}\sqrt{n_2}k)\Delta$ of the test statistics z_i leads to the optimal weight

$$w_1 = \frac{n_1}{n_1 + n_2k^2}. \tag{6}$$

If the same method (with the same effect size, $k=1$) is used at both stages, then the weight $w_1 = n_1/(n_1+n_2)$ corresponds to that used in a group sequential two-stage design. Note that using 'non-optimal' weights may lead to a larger power of the pilot design as compared to the integrated design when the effect size in the second stage is much larger than in the first stage (as already pointed at by Skol *et al.*, 2006).

5.2 Optimal designs controlling the FWE rate

For the control of the FWE rate, the corresponding γ is the solution of:

$$\begin{aligned}\gamma_s &= P_{H_{0i}}(p_i^{(1)} < \gamma_1, p_i < \gamma) \\ &= \int_{c_{1-\gamma_1}}^{\infty}\left[1-\Phi\left(\frac{c_{1-\gamma} - \sqrt{w_1}z}{\sqrt{1-w_1}}\right)\right]\varphi(z)dz\end{aligned} \tag{7}$$

where γ_s is set to α/m_1. φ denotes the density function of the standard normal distribution. Note again that n_2 is random because it depends on the number of selected hypotheses (which also is random). By re-formulating the test decisions in terms of a sequential P-value p_{si} based on the Tsiatis–Mehta–Rosner ordering, (H_{0i} is rejected if $p_{si} < \gamma_s$) one can show that this procedure with the predefined sample size reallocation rule for the selected null hypotheses controls the FWE rate because under the null hypothesis they follow a uniform distribution (Zehetmayer *et al.*, 2005). The overall power is given by

$$\begin{aligned}\prod\nolimits_{int} &= P_{\mu_{1i}=\Delta\sigma_1,\, \mu_{2i}=k\Delta\sigma_2}(p_i^{(1)} < \gamma_1, p_i < \gamma) \\ &= \int_{c_{1-\gamma_1}}^{\infty}\left[1-\Phi_{\sqrt{n_2}k\Delta,1}\left(\frac{c_{1-\gamma} - \sqrt{w_1}z}{\sqrt{1-w_1}}\right)\right]\varphi_{\sqrt{n_1}\Delta,1}(z)dz,\end{aligned} \tag{8}$$

where, φ_{μ,σ^2} is the density function of the normal distribution with mean μ and variance σ^2. Given the other quantities, we can optimize $\prod_{\text{int}}$ in the two design parameters r and γ_1. Note that the optimal γ_1 and r, as in the pilot design, depend on C, m_1, Δ and k via $\sqrt{C/m_1}\Delta$ and $k/\sqrt{c_2}$.

5.3 Optimal designs controlling the FDR

For the control of the FDR, asymptotically the rejection boundary for the P-values in the final test is given by the solution of

$$\alpha = \frac{\pi_0\gamma_s}{\pi_0\gamma_s + (1-\pi_0)(1-\beta(\gamma_s))} \tag{9}$$

where γ_s is a function of γ which is given by (7). Such a two-stage procedure with a predefined sample size allocation rule controls the FDR, since it can be shown that the resulting sequential P-values p_{si} are independent across hypotheses (Zehetmayer *et al.*, 2005). Again, optimal values of r and γ_1 can be determined by maximizing the power (8) under the constraint (9). The rejection boundary γ for the P-values p_i of the selected null hypotheses calculated from pooling stagewise z-scores (5) with optimal weights (6) can then be found numerically from solving Equation (7).

6 COMPARISON OF TWO-STAGE PROCEDURES

6.1 Pilot design

Assume first that the experimenter has two different candidate methods for the measurements from the very beginning, a low-cost standard method and a high-cost improved method. So he could apply the same method at both stages ('low–low' or 'high–high'), or he may switch to the more expensive method at the second stage ('low–high'). In the following, we investigate which of these three procedures is more powerful when controlling the FWE rate. Using the same

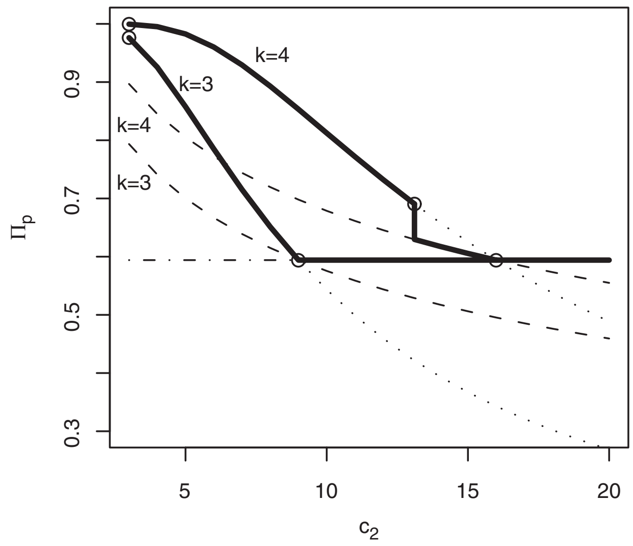

Fig. 1. Asymptotically optimal power of the low–low (dashed-dotted horizontal line), the low–high (dashed lines) and the high–high (dotted lines) procedure of the pilot design for varying c_2 and effect size ratios $k=3$ and $k=4$. The solid lines mark the respective maximum over the three procedures under the constraint $n_1 \geq 1$ for the high–high design. $C=20\,000$, $m_1=1000$, $\pi_0=0.99$, $\Delta=0.5$, FWE rate $\alpha=0.05$.

test statistics only with modified critical boundaries, we expect similar findings when controlling the FDR. The power of the pilot design controlling the FWE rate for the low–high procedure is given by (3). Clearly the power of a procedure using the low-cost method in both stages, $\prod_{p_{ll}}$, say, is given by setting $k=1$ and $c_2=1$; the power for the procedure using the high-cost method at both stages, $\prod_{p_{hh}}$, say, arises from (3) by using $(1-\Phi_{\sqrt{(rC/m_1c_2)}k\Delta,\,1}(c_{1-\gamma_1}))$ for the first-stage power leaving the second-stage power unchanged. It is easy to see that for $k=\sqrt{c_2}$ we get the identity $\prod_p \equiv \prod_{p_{ll}} \equiv \prod_{p_{hh}}$. Hence, the maxima of all three functions in r and γ_1 are identical. Since formula (3) is monotonic in c_2, the two-stage procedure applying the low-cost measurement method at both stages dominates the other two procedures ('low–high' and 'high–high') if the high-cost method is not sufficiently efficient, i.e. when $c_2 > k^2$. For $c_2 < k^2$, the high–high procedure dominates the other two. Hence, the important conclusion is that the procedure switching from the low-cost to the high-cost method is never the best procedure in terms of asymptotic power. However, it may be useful if the asymptotically optimal sample size (n_1) at the first stage is too small for the high–high procedure. Figure 1 shows the maximum asymptotically optimal power over the three procedures for the pilot design for varying c_2, given the constraint $n_1 \geq 1$. Two different effect ratios are assumed, $k=3$ and 4. The example $C=20\,000$, $m_1=1000$, $\pi_0=0.99$ and $\alpha=0.05$ (FWE), was used assuming an effect size for the low-cost measurement method of $\Delta=0.5$. The asymptotically optimal power is given for the three procedures. The solid lines mark the respective maximal power over the three procedures if at least one observation is left at the first stage for the optimal high–high procedure. Note that for the other two procedures, the asymptotically optimal n_1 is always larger than one. Obviously, the high–high procedure has the maximum power for relatively low costs c_2. For the cost ratio $k=4$, the solid curve jumps when the costs of the high-cost method get too large resulting in an asymptotic optimal $n_1 < 1$. Here, the region where the low–high procedure is preferable to both, the other is very small, for $k=3$ no such region exists. If we apply the constraint $n_1 \geq 2$, the region where the low–high procedure is preferable gets larger. For our example, such a region would even exist for an effect ratio of $k=3$ (data not shown). Note that the crossing point depends on the unknown effect size, and no procedure dominates the other two over the whole parameter space. Hence, in case of design misspecifications in the planning phase there will be other parameter constellations where the low–high type of strategy is in fact more powerful. However, when no misspecifications occur, the low–high procedure is preferable only if the high-cost method is too expensive so that the first-stage sample size for the high–high procedure is insufficiently small.

6.2 Integrated design

Comparing the three procedures for the integrated design, we have to modify the formula for the power $\prod_{\text{int}}$ given for the low–high procedure in (8). For the low–low procedure to calculate the power, we have to insert $k=1$ and $c_2=1$. For the high–high procedure, we have to replace $\sqrt{n_1}\Delta$ by $\sqrt{n_1/c_2}k\Delta$. It can be seen easily that again for $k=\sqrt{c_2}$ the three power functions are identical so that there is the same crossing point for the integrated design. Essentially, the results are very similar to those in Figure 1 for the pilot design (data not shown). Note that the common crossing point exists only if in the integrated low–high procedure the optimal weights (9) are used for combining the stagewise test statistics. The low–high procedure looses power when applying non-optimal weights (which will be the rule in applications).

6.3 Examples: optimal designs for $k=1$ and $c_2 \geq 1$

The previous sections have shown that if two methods are available, differing in costs and effect sizes, using two-stage designs applying the same method at both stages may be preferable. Asymptotically, optimal two-stage designs applying the same method at both stages ($k=1$) can be derived as in Zehetmayer *et al.* (2005) if the costs do not differ between stages ($c_2=1$) using appropriately defined total costs C. In the following, we focus on designs using the same methods at both stages; the second-stage measurement, however, raising extra costs $c_2 > 1$. When $c_2 > 1$, we have to use the power formulas (3) and (8) with $k=1$ to derive asymptotically optimal designs. Table 1 for $k=1$ and some c_2 gives the design parameters of optimal pilot and integrated designs and their power for controlling the FWE rate and the FDR. Note that the optimal power values given for the integrated designs are only slightly larger than those of the pilot designs. For comparison, the power of the (asymptotic) single-stage designs with equal total costs for the control of the FWE rate and FDR are also listed in Table 1. As one can see from the tables, the asymptotic optimal screening boundary γ_1 decreases with increasing costs c_2. For the same costs, the screening boundary γ_1 slightly increases with increasing Δ. At the same time, the

Table 1. Optimal two-stage designs controlling the FWE rate or FDR at $\alpha = 0.05$

			$\Delta = 0.5$			$\Delta = 0.75$			$\Delta = 1$		
	c_2	Design	r	γ_1	$\prod$	i	γ_1	$\prod$	r	γ_1	$\prod$
FWE rate	1	Pilot	0.635	0.074	0.594	0.718	0.092	0.926	0.774	0.097	0.995
		integrated	0.642	0.077	0.603	0.725	0.103	0.934	0.779	0.120	0.997
	5	Pilot	0.683	0.015	0.341	0.737	0.019	0.762	0.781	0.019	0.966
		integrated	0.697	0.016	0.351	0.759	0.021	0.783	0.806	0.024	0.974
	15	Pilot	0.685	0.006	0.214	0.701	0.007	0.589	0.722	0.007	0.893
		integrated	0.706	0.006	0.226	0.745	0.007	0.628	0.787	0.008	0.925
	Single-stage design				0.049			0.296			0.720
FDR	1	Pilot	0.632	0.096	0.641	0.715	0.119	0.943	0.772	0.121	0.997
		integrated	0.639	0.101	0.651	0.722	0.137	0.951	0.776	0.158	0.998
	5	Pilot	0.701	0.019	0.379	0.765	0.025	0.810	0.807	0.025	0.977
		integrated	0.707	0.020	0.387	0.778	0.029	0.828	0.824	0.033	0.983
	15	Pilot	0.716	0.007	0.242	0.766	0.009	0.673	0.799	0.009	0.936
		integrated	0.723	0.007	0.249	0.788	0.010	0.700	0.832	0.012	0.954
	Single-stage design				0.056			0.443			0.877

Asymptotically optimal parameters γ_1, r and the power $\prod$ for different c_2 and Δ. $k = 1$, $C = 20\,000$, $m_1 = 1000$, $\pi_0 = 0.99$. Power values of the corresponding single-stage designs are given for comparison.

proportion of costs used for the first stage increases with Δ. Note that due to the complexity of the power function there is a different dependence on costs for low and large effect sizes, which is also depending on FDR or FWE control. At least in the asymptotically optimal number of selected hypotheses m_2 increases with Δ and decreases with costs c_2 throughout the whole designs considered. Note that using designs with stage-wise integer sample size (first rounded downwards and randomly choosing hypotheses where the rounded sample size is increased by 1 in order to achieve constant total costs) does not noticeably decrease the power as compared to the optimal non-integer designs. Simulations (100 000 runs each) for the cases $C = 20\,000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.75$, $c_2 = 5$ and 15 from Table 1 show for the pilot design power values of $\prod_p = 0.753$ and 0.574, respectively for an FWE rate of $\alpha = 0.05$ and $\prod_p = 0.802$ and 0.660 for FDR control at the same level. It has to be mentioned that for large costs the number m_2 of selected hypotheses may become small, so that the finite sample size modification of formulas (1) and (2) proposed by Storey, *et al.* (2004) has to be used in order to guarantee control of the FDR. This leads to a slight decrease in power.

7 WHEN TO USE TWO-STAGE DESIGNS

7.1 Break even point in the cost ratio

It has been shown that for large m_1 and constraints on the total costs, the power of an asymptotic optimal two-stage design may be considerably larger than the power of the corresponding single-stage design (see Table 1). Again, the scenario is considered where the same method is applied at the two stages ($k = 1$) and the second stage measurement raises extra costs ($c_2 > 1$). We investigate when it is more efficient in terms of asymptotic power to use a two-stage design as compared to the single-stage design. We tackle the problem by asking whether a cost ratio c_2^* exists, where the power of the single-stage and the two-stage designs are the same. If the asymptotic power would be monotonically decreasing in c_2 for $c_2 > c_2^*$, the single-stage design would provide a larger power than the two-stage design. The first important answer is that for the integrated design such a finite c_2^* does not exist, because for given C, m_1, Δ, k and α and $c_2 \to \infty$ the power of the asymptotic optimal integrated design converges to the power of the single-stage design applying the low-cost measurement method. Hence, for the integrated approach theoretically the two-stage approach always pays off. However, in practice, if the optimal second-stage sample size gets too small, the two-stage design cannot be used. For the pilot design, the power converges to 0 as $c_2 \to \infty$. Hence, for the pilot design in general such a break even point c_2^* between the two-stage and single-stage designs exists. Figure 2 shows c_2^* for varying π_0 and Δ for the case of controlling the FWE rate or the FDR at $\alpha = 0.05$. Again, C was set to 20 000 and m_1 was set to 1000. The curves are fairly similar for control of the FWE rate and the FDR, the break even point varying more when the FDR is controlled. For large effect sizes, the power of the single-stage design and the pilot design are close to 1, and consequently c_2^* is small. For decreasing effect sizes, the break even point c_2^* is increasing. When the number of true alternatives decreases (π_0 increases) c_2^* increases. In both situations, a smaller number of null hypotheses is selected for the second stage (with larger sample sizes n_2) so that we can afford higher costs for the selected hypotheses. Note that the power when controlling the FDR is always slightly larger than when controlling the FWE rate. If there is a relatively large proportion of alternatives with substantial effects, the break even point is smaller for controlling the FDR than the FWE rate: the single-stage design controlling the FDR then is noticeably more powerful than the single-stage design controlling the FWE rate.

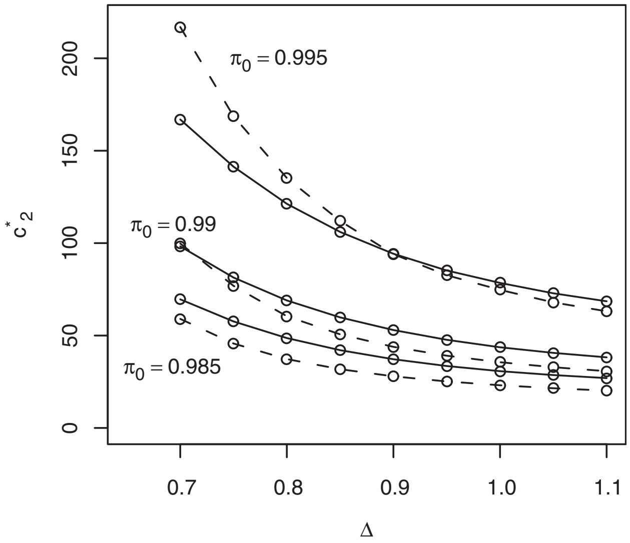

Fig. 2. Break even point c_2^* for the cost ratio between the asymptotically optimal pilot design and the single-stage design depending on Δ and π_0, for controlling the FDR (dashed lines) or the FWE rate (solid lines). $C = 20\,000$, $m_1 = 1000$, FWE rate and FDR both $\alpha = 0.05$.

For decreasing Δ, this advantage in power of the single-stage FDR design over the single-stage FWE design decreases, whereas the optimal two-stage design controlling the FDR still has favorable properties as compared to the two-stage FWE design. Hence, larger second-stage costs can be afforded to achieve the same power as the corresponding single-stage design. This may lead to a crossing of the two corresponding curves.

7.2 Impact of design misspecifications

Whereas costs are usually known a priori, the optimal designs depend on the unknown proportion π_0 and effect size Δ. Hence, the impact of design misspecifications in the planning phase is an important issue. In the following, again we consider the scenario $C = 20\,000$, $m_1 = 1000$ and $\alpha = 0.05$. It is assumed that the optimal r and γ_1 were planned for the situation where $\Delta = 0.75$, $\pi_0 = 0.99$ and $k = 1$. Figure 3 shows the differences between the power of the two-stage designs and the single-stage design as a function of the true π_0 and Δ for controlling the FDR and FWE rate. Positive values indicate superiority of the two-stage design. The example with a cost ratio $c_2 = 15$ (confer Wang *et al.*, 2006) is plotted for the pilot (first row of the

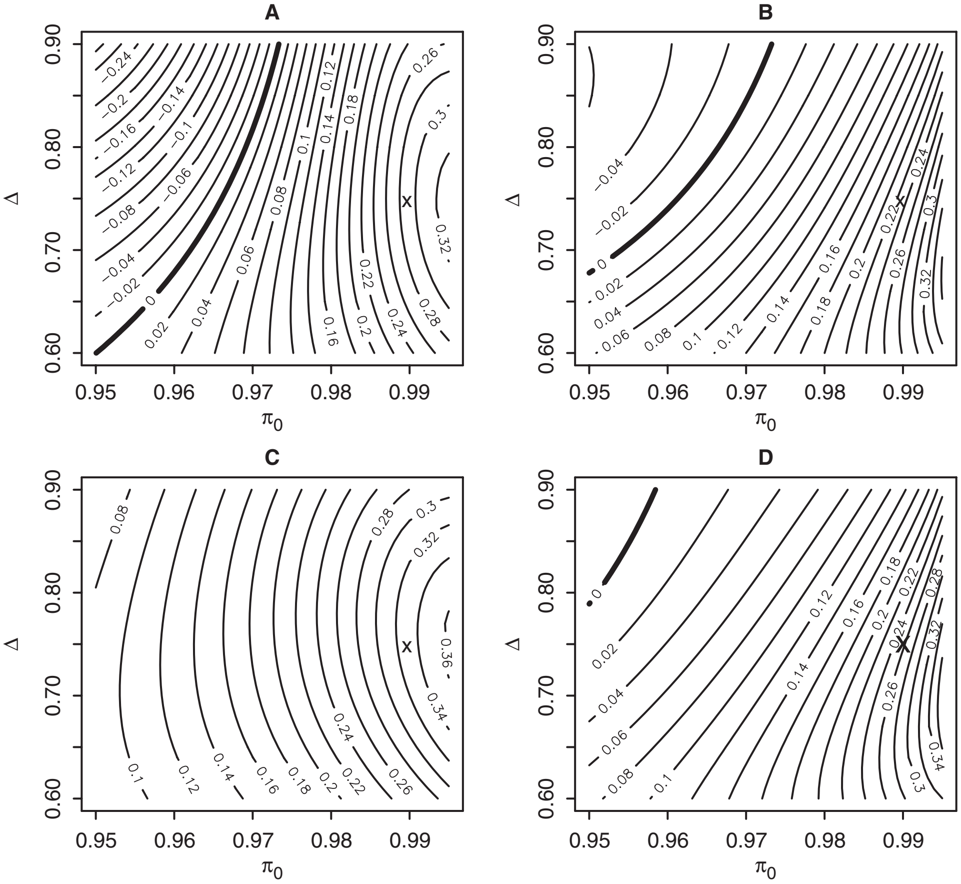

Fig. 3. Contour plots for the difference in power between the single-stage and the pilot design (first row) and the single-stage and the integrated design (second row) as a function of the true π_0 and Δ for controlling the FWE rate (first column) or the FDR (second column). Positive values indicate superiority of the two-stage design. Bold lines mark equality between the single-stage and the two-stage design. Asymptotically, optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$ (marked as cross, confer Table 1). $C = 20\,000$, $c_2 = 15$ and $m_1 = 1000$.

panels) and the integrated design (second row). Not surprising, the figures show that the integrated design is more robust against misspecifications of π_0 and Δ than the pilot design: it uses the whole data set from both stages for test decisions. The most robust design is the integrated design controlling the FWE (Fig. 3C). Here, in the parameter subspace, the two-stage integrated design shown is always noticeably better than the single-stage design. Controlling the FDR, the advantage of the single-stage design to adapt for π_0 results in smaller differences between the integrated two-stage design and the single-stage design (Fig. 3D): in the left upper corner, the single-stage design is outperforming the two-stage design. The pilot design controlling the FWE rate is more sensible with regard to the design misspecifications than the pilot design controlling the FDR. The design applies 'non-optimal' selection criteria and controlling the FWE rate no adaption to the correct parameters is possible in the second-stage sample (Fig. 3A): in the left upper corner, the power of the single-stage design may become substantially larger than the two-stage pilot design. Controlling the FDR adapting to the true parameters in the second-stage sample helps a little (Fig. 3B): there is only a slightly larger power of the single-stage design as compared to the two-stage pilot design in the left upper corner. Generally, a design optimal for a fraction of true null hypotheses which is larger than the true π_0 can lead to a considerable loss of power as compared to the corresponding single-stage design. However, if the true π_0 gets larger than the proportion used for planning and the true effect size Δ is close to the one used for planning generally the difference between two-stage designs and the single-stage design increases. Optimism in the planning phase with regard to the number of true alternatives may help to avoid a loss of power due to design misspecification. If the true effect size Δ gets larger than the one from the planning phase for values of π_0 close to the true one, the power of the two-stage and single-stage designs both approach 1 so that the differences in the contour plots decrease.

8 DISCUSSION

We have investigated two-stage designs in the situation that large numbers of null hypotheses are tested and only a small proportion of them are expected to be wrong. Moreover, it was assumed that there are constraints on total costs of the experiment. The first stage is used for screening out promising hypotheses which are then investigated further at the second stage. We focused on an important scenario in practice assuming that costs per measurement differ between stages: on the one hand, extra costs may arise when the same measurements have to be designed for a subset of hypotheses selected in an interim analysis and investigated at the second stage. On the other hand, the investigator from the very beginning may have the choice between a low-cost method and a high-cost method (which hopefully is more efficient in terms of the effect size under the alternatives). Given a large number of candidate hypotheses, we derived asymptotically optimal designs in terms of power using the simplifying assumptions of common alternatives (either controlling the FWE rate or the FDR).

We would like to summarize the results in the following way: if two different methods are available, depending on the ratios between costs and effect sizes it is preferable to run two-stage designs which apply either the low-cost or the high-cost method at both stages. Designs starting with the low-cost method and switching to the more expensive method in the interim analysis may only be advisable if there is lack of resources, so that first-stage sample size for the high-cost method would be too small. However, it has to be kept in mind that the best design depends on the relationship of the effect size and the cost ratios. Hence, in case of effect size misspecifications in the planning phase, the low–high method may actually be more powerful than the low–low or the high–high strategy. However, it seems natural to apply a design which is preferable under the parametric constellation considered in the planning phase. In the integrated design, the optimal way of combining more data from both stages arising from different measurement methods depends on the effect size ratio between stages, which introduces a further complication for appropriately designing such experiments applying different methods.

Two-stage screening designs are a very powerful tool even if we deal with equal effect sizes at the second stage, but the costs for designing the measurements for the selected hypotheses at the second stage are fairly high. Only severe design misspecifications in the planning phase may lead to a noticeable loss of power such that the single-stage design may become superior in power. With regard to the impact of design misspecification in the proportion of true alternatives, it seems to be preferable not to assume too small proportions in the planning phase. Integrated designs which use data from both stages for the final test decisions are more robust against design misspecifications.

With respect to deviations from the underlying assumption, we calculated optimal designs for the unknown variance case using the central and non-central t-distributions instead of the corresponding normal distributions. Again, assuming $\Delta = 0.75$, $c_2 = 5$ and 15 from Table 1, the optimal parameters for the pilot design controlling the FWE rate are $r = 0.722$, $\gamma_1 = 0.020$ and $r = 0.703$, $\gamma_1 = 0.007$, respectively, which are very close to those of the known variance case. The corresponding optimal power values for the unknown variance case drop to 0.681 and 0.473. For the control of the FDR, the corresponding optimal design parameters in the unknown variance case change to $r = 0.748$, $\gamma_1 = 0.026$ for $c_2 = 5$ and to $r = 0.757$, $\gamma_1 = 0.009$ for $c_2 = 15$. The optimal power decreases to 0.747 and 0.565, respectively. However, using the optimal parameters for the known variance case in the situation of unknown variances leads to virtually the same performance as using the optimal parameters from the unknown variance case. Note that in the unknown variance case, the decision which of the procedures (low–low, high–high or low–high) is preferable is more difficult because no common crossing point in costs as a function of c_2 between the three procedures exists. However, the region where the low–high procedure is preferable still remains small.

To investigate the impact of correlation, we assumed an autoregressive correlation structure among the hypotheses. The correlation between hypotheses i and j is given by $\rho^{|i-j|}$ for some $\rho \in (0, 1)$. The alternative hypotheses $(\Delta = 0.75)$ are randomly distributed among the hypotheses. For example the simulated power values (100 000 runs) for $c_2 = 5$ assuming a correlation of $\rho = 0.2$, 0.6 and 0.9 are 0.753, 0.749 and 0.728, respectively when controlling the FWE rate, and 0.802, 0.798 and 0.777, respectively when controlling the FDR (compare Table 1). Hence, the impact of correlation is small like in the case of constant costs in Zehetmayer *et al.* (2005). For the two-sided situation, we refer to their proposal to test a set of $2m_1$ one-sided hypotheses.

ACKNOWLEDGEMENTS

We thank Sonja Zehetmayer, Martin Posch and the three anonymous referees for constructive comments. This work was supported by the Austrian FWF-Fund no. P18698-n15.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bukszár,J. and Van den Oord,E. (2006) Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for Pearson's statistic. *Biometrics*, **62**, 1132–1137.

Lehmacher,W. and Wassmer,G. (1999) Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286–1290.

Ohashi,J. and Clark,A.G. (2005) Application of the stepwise focusing method to optimize the cost-effectiveness of genome-wide association studies with limited research budgets for genotyping and phenotyping. *Ann. Hum. Genet.*, **69**, 323–328.

R Development Core Team (2005) R: a language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria.

Satagopan,J.M. *et al.* (2002) Two-stage designs for gene-disease association studies. *Biometrics*, **58**, 163–170.

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

Satagopan,J.M. *et al.* (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, **60**, 589–597.

Skol,A.D. *et al.* (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.*, **38**, 209–213.

Storey,J.D. (2002) A direct approach to false discovery rate. *J. R. Stat. Soc. B*, **64**, 479–498.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B*, **66**, 187–205.

Van den Oord,E.J. and Sullivan,P.F. (2003) A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. *Hum. Hered.*, **56**, 188–199.

Wang,H. *et al.* (2006) Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol*, **30**, 356–368.

Zehetmayer,S. *et al.* (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, **21**, 3771–3777.