STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1565–1580
Published online 15 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3090

Adaptive designs: Looking for a needle in the haystack—A new challenge in medical research

Peter Bauer*,†

Section of Medical Statistics, Medical University of Vienna, Vienna, Austria

SUMMARY

The statistical principles of fully adaptive designs are outlined. The options of flexibility and the price to be paid in terms of statistical properties of the test procedures are discussed. It is stressed that controlled inference after major design modifications (changing hypotheses) will include a penalty: Intersections among all the hypotheses considered throughout the trial have to be rejected before testing individual hypotheses. Moreover, feasibility in terms of integrity and persuasiveness of the results achieved after adaptations based on unblinded data is considered as the crucial issue in practice. In the second part, sample size adaptive procedures are considered testing a large number of hypotheses under constraints on total sample size as in genetic studies. The advantage of sequential procedures is sketched for the example of two-stage designs with a pilot phase for screening promising hypotheses (markers) and controlling the false discovery rate. Finally, we turn to the clinical problem how to select markers and estimate a score from limited samples, e.g. for predicting the response to therapy of a future patient. The predictive ability of such scores will be rather poor when investigating a large number of hypotheses and truly large marker effects are lacking. An obvious dilemma will show up: More optimistic selection rules may be superior if in fact effective markers exist, but will produce more nuisance prediction if no effective markers exist compared with more cautious strategies, e.g. aiming at some control of type I error probabilities. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: adaptive designs; optimality; weighting; feasibility; large number of hypotheses; false discovery rate; pilot designs; prediction scores

1. INTRODUCTION

Adaptive designs in clinical trials have been used in the past to tackle various types of problems: e.g. in early phase dose–response clinical trials for an efficient estimation of doses associated with

*Correspondence to: Peter Bauer, Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.
†E-mail: peter.bauer@meduniwien.ac.at

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

Received 26 July 2007
Accepted 30 August 2007

a certain level of toxicity or effectiveness when using response-dependent dose escalation or, in later phases, when using response-dependent randomization, e.g. to reduce the number of patients treated with inferior treatments [1]. Over the last decade, a new type of adaptive design that allows flexible interim decisions has provoked an ongoing and intensive discussion among biometricians [2, 3]. Whereas for conventional designs in general the main features of the design have to be completely pre-specified, adaptive designs allow for changes of essential features of the design based on data from inside or outside the ongoing trial without compromising on the type I error rate. The striking property is that in principle there is no need to pre-specify the type and details of the potential adaptations in advance. The construction of tests in such adaptive designs relies on the very simple principles that open an enormous amount of flexibility. Clearly when extending conventional methods in such a radical way a price will have to be paid for the many options offered by adaptive designs.

In the following, a short introduction is given about the principles of constructing such adaptive tests. Then, the inferential problems resulting from these simple principles are discussed, most importantly that these tests violate the sufficiency principle. The discussion closes with the important issue of feasibility: How can we maintain integrity and persuasiveness of the results after flexible decisions have been performed based on unblinded data?

The second part of the paper deals with a problem with which biometricians are increasingly confronted. For example, in gene expression or gene association studies the experimenter deals with an extremely large number of markers (and associated hypotheses), but due to constraints on resources sample sizes are rather small. It is shown how, in the situation of constrained costs, adaptive sequential procedures can improve the power when promising hypotheses are screened at earlier stage with small sample sizes, to be investigated further at later stages with larger sample sizes. These procedures are ‘adaptive’ in the sense that sample sizes at later stages depend on the random number of screened hypotheses. The criterion used for selection is the false discovery rate (FDR) which is estimated from the data (thus leading to ‘adaptive’ decision boundaries). Finally, we push the problem further and consider prediction: What can we expect, if in samples of patients responding and not responding to a particular therapy markers are selected and used to construct a score for prediction of response in future patients. The predictive ability of such estimated scores is investigated by simulating their receiver operating characteristic (ROC) curves. Some general arguments will be given in the closing section.

2. FULLY FLEXIBLE DESIGNS

Currently, there is discussion relating to the concept ‘fully flexible’ designs where no fixed adaptation rule is given in advance as opposed to the so-called ‘planned flexible’ designs, where flexibility follows a strict predefined adaptation rule. Although in the latter case the type I error rate of a test based on the conventional (sufficient) statistics can be calculated in principle (and therefore can also be adjusted to control a given overall level α), in the first case no test exists which fully exploits the alpha level [4].

2.1. The combination test principle

To keep things simple we consider a two-stage design with a single interim analysis, the generalization to more than two stages being straightforward. We deal with a one-sided test of the mean

DOI: 10.1002/sim

of a normal distribution, variance known ($\sigma^2=1$), with the null hypothesis $H_0:\mu=0$ *versus* the alternative $H_1:\mu>0$. In a conventional group sequential two-stage design, the test statistic in the interim analysis is based on the mean of the n_1 observations at the first stage. In the final analysis, the test statistic is calculated from the overall mean after pooling the n_2 observations of the second stage with the n_1 observations of the first stage. In a fully flexible design, the interim analysis is based on the same test statistics as the group sequential test. However, in the final analysis instead of pooling the two samples from stages 1 and 2, the two separate test statistics from the disjoint samples are combined according to a predefined combination function [5, 6]. Suitable test statistics include the p-values p_1 and p_2 or z-scores z_1 and z_2 from the disjoint samples. Examples of combination functions considered in the literature are Fisher's product criterion p_1p_2 [6] or the linear combination of z-scores which is identical to the so-called inverse normal combination function $w_1z_1+w_2z_2=w_1\Phi^{-1}(1-p_1)+w_2\Phi^{-1}(1-p_2)$, $w_1^2+w_2^2=1$ [7, 8]. If the p-values are calculated from the stage-wise z-scores, the inverse normal combination functions in group sequential designs corresponds to the sufficient test statistics if the weights are pre-fixed as the square root of the fractions of the total sample size planned for the first and second stages, respectively, i.e. $w_1=\sqrt{n_1/(n_1+n_2)}$, $w_2=\sqrt{n_2/(n_1+n_2)}$. Both combination functions have been proposed for meta analysis when combining evidence from independent trials (e.g. [9]). Now, the crucial point is that under the null hypothesis H_0 the stage-wise test statistics under fairly general conditions [10] preserve their distributional properties even when adaptations have been performed in the interim analysis. In our example, such an adaptation may be an increase in the second-stage sample size from n_2 to $\widetilde{n}_2$: Whatever data from the first stage are used to determine the new sample size $\widetilde{n}_2$ under H_0, the second-stage p-value p_2 is still uniformly distributed and independent of the first-stage p-value p_1, and the stage-wise z-scores still follow independent standard normal distributions (if at the second stage independent samples are taken). Hence, exploiting these invariance properties, critical regions derived for the non-adaptive scenario can also be applied in the adaptive scenario. Moreover, when using the inverse normal combination function 'induced' by the pre-planned group-sequential design and applying the same sequential decision boundaries, the adaptive combination test is identical to the group sequential test when no adaptations are performed. This principle opens a wide field: It can be generalized to other types of test problems, to more than two stages and even to the situation of a 'continuous' recursive application of the combination test after each observation. The latter allows to perform adaptive interim analyses at any time point and to modify the number, the timing and the decision boundaries of future interim analysis [11]. The principle can also be stretched to cover adaptations beyond sample size reassessment, e.g. when a primary variable is modified in an interim analysis. In such a situation, different individual hypotheses (or sets of hypotheses) $H_0^{(1)}$ and $H_0^{(2)}$ are tested at the two stages, respectively. However, it has to be kept in mind that in this situation the combination test addresses only the intersection $H_0^{(1)}\cap H_0^{(2)}$. This may be sufficient in some scenarios, e.g. when establishing a dose–response relationship. However, in general the question will come up as to which of the null hypotheses considered at the different stages is wrong. In this situation methods of multiple testing have to be applied to provide controlled inference on the individual null hypotheses [12–14]. Adding such a multiple testing procedure has been considered to be a major complication for the interpretation of the results of adaptive designs. However, the common (and extensively used) practice of handling design modification in conventional designs by writing amendments to the protocol simply ignores the problem of possible changes of the null hypothesis. Adaptive designs acknowledge that design modifications may be required in practice. However,

DOI: 10.1002/sim

it should be clear that by starting with a large set of questions (hypotheses) at earlier stages of adaptive designs cannot be simply looked at as a tool for generating hypotheses that are then investigated in a confirmatory way at later stages. To obtain a controlled inference on the individual null hypotheses, e.g. by the closed testing principle, various intersection hypotheses including all the null hypotheses of previous stages have to be rejected before the individual null hypotheses resulting from the adaptations can be addressed. There seems to be a lack of awareness particularly in the applied field that correctly performed inference on individual hypotheses in adaptive designs asks for a penalty when essential features such as the hypotheses are changed.

2.2. *The conditional error function principle*

The second invariance principle behind fully flexible designs is based on the conditional error function. Assume that a conventional trial has been pre-planned for a test controlling a particular level α. If we look at the data at any time during the trial, the conditional error function is defined as the probability (under the null hypothesis) to perform a type I error later on, given the data up to the interim analysis [15]. Then we can replace the remainder of the design by any other design (sometimes called a 'secondary' design) which never leads to a larger conditional error than the original pre-planned design (given the data observed up to now), [15–17]. Such a design modification can never inflate the overall type I error rate of the original design. Clearly, this invariance principle can also be applied in a recursive way. It combines the advantages of planning a conventional design (e.g. a group sequential design) but allowing for deviations from this design if this is considered to be necessary. Hence, the spending functions for early rejections and futility stopping can be re-defined for the remainder of the trial, i.e. the number of interim analyses, sample sizes between the analyses and the decision rules can be adapted before the remainder of the design is performed. The method is of a disturbing generality, because the only restriction for design modifications is preservation of the conditional error. This even more clearly demonstrates that fully flexible designs are a larger class than group sequential designs, provided that the conditional error function can be calculated, which may become difficult if nuisance parameters are involved [18]. In principle, any design that includes planned flexibility with a fixed adaptation rule can be generalized to a fully flexible design if its conditional error function can be quantified. Consequently, adaptive designs are also a generalization of designs with pre-planned flexibility.

The combination test and the conditional error approach of achieving flexibility are closely related because the conditional error function induces a way of combining early data with later data, and the combination function induces a conditional error function [11, 19].

2.3. *Optimality*

It has been argued that adaptive tests can be uniformly improved by other tests using sufficient statistics [20, 21]. In the first of the two papers, it is assumed that a fixed adaptation rule is followed, which defines spending functions for early rejections and early acceptance and that interim analyses are of no costs. In the second paper, the risk function considered is a linear combination of the type I error rate, the type II error rate and the expected information. The Bayes risk in the form of an expectation of this risk function over a discrete *a priori* distribution on a finite grid in the parameter space is minimized. It is shown that each admissible decision rule is a Bayes rule with a solution based on the sufficient statistics. However, real life is more complex. The risk structure may and does change even during a trial. Biometricians with the experience of data safety and monitoring know that this happens often. For example, an unexpected safety issue may arise in a clinical trial.

DOI: 10.1002/sim

As a consequence, a larger sample size under the experimental therapy may be required to achieve a sufficiently precise estimate of the cost–benefit relationship. Now, the original weight in the *a priori*-defined risk function for the expected information may be irrelevant. The costs for additional sampling may be completely dominated by the need to get more information on a variable other than any pre-planned outcome variable. It has been shown that given the adaptation and the interim data we can again use conditionally optimal designs based on the sufficient statistics for the rest of the trial [22]. Originally one of the motivations behind adaptive design was to oppose the increasing tendency of writing amendments to conventional protocols whenever design adaptations seemed to be necessary (questioning if their impact on the type I error rate has always been fully understood). Strong arguments have been raised against the concept of mid-trial sample size reassessment in general [23]. Instead conventional group sequential designs powered at a small effect size are advocated. Such 'overpowered' designs lead to smaller expected sample sizes than adaptive tests with sample size reassessment based on conditional power arguments. This planning philosophy, however, will confront us with unusually large maximum sample sizes. A careful consideration of futility boundaries will be required to keep the sample size small also for parameters close to the null hypothesis. It is appropriate that sample size reassessment based on early results which are highly variable should not become standard in clinical trials [24]. However, it is generally agreed that the decision to modify the sample size may be based on various reasons, such as a change in the environment because a contender has turned up, because the market situation has changed, because one likes to switch from non-inferiority to the higher goal of superiority, because one has stopped ineffective or unsafe treatment arms and would like to reallocate the saved sample size to the remaining treatments, because of the need of additional information on safety or because of other reasons.

2.4. Weighting

When calculating the combination test statistic in the case of (sample size) adaptations, observations from different stages in general are weighted differently (due to the *a priori* definition of the combination function). Although there are also other tests in use which apply different weights for different observations, the fact remains that adaptive test violates the sufficiency principle, as has been realized by several authors. As a consequence, extreme situations for absurd test decisions can be constructed, so that the adaptive test rejects while the overall mean has the wrong sign [3, 15]. It has to be mentioned that the problem of such 'treatment × stage interactions' has been intensively discussed already in the beginning [5, 6]. One can identify the region in the sample space where it may happen that the adaptive design rejects but not the conventional fixed sample size. Here the conditional error of the adaptive test is larger than the conditional error of the conventional fixed sample size test [24]. The bad news is that in general this may happen if reasonable sample size reassessment rules are applied, e.g. increasing the sample size if a small effect has been observed. Applying suitable early rejection and futility boundaries may help to avoid such inconsistencies. We can also apply the marginally conservative 'dual test' which rejects only if both the adaptive and the conventional tests based on the sufficient statistics reject (either the fixed sample test size in the total sample [25] or the pre-planned group sequential test [26]). In addition, a report of the stage-wise test statistics will help in checking the internal consistencies of the data over the trial and may be taken as a progress in transparency compared with group sequential designs. There seems to be no generally sound alternative to adaptive tests when no fixed adaptation rule is given. For the conventional ('un-weighted') LR test the rejection region generally depends on the value

DOI: 10.1002/sim

of the parameter under the alternative and on the distribution of the sample size which is unknown when not adhering to a pre-defined adaptation rule. Conditioning on the sample size is not sound either since in many situations the sample size is highly informative [3]. Obviously frequentist inference runs into real troubles if the sample size is random and its distribution is not known in advance. One alternative approach is, for any outcome in the interim analysis, to determine the (sample size) adaptation rule for which the un-weighted test based on the sufficient statistics has the largest conditional type I error rate. Then the level of this conventional test is adjusted from α to $\alpha{*}(\leqslant\alpha)$ so that the actual level is still not exceeding the targeted value α, even if an experimenter would always choose an adaptation rule that produces the largest conditional type I error rate [15, 27, 28]. As a consequence, the level is not fully exploited by the adjusted conventional test whenever the experimenter deviates from the worst-case adaptation rule. Hence, the adjusted test can be uniformly improved by applying an adaptive test using the worst-case conditional error function. Even applying the dual adaptive test, which requires also a rejection of the conventional test at the level $\alpha\geqslant\alpha_*$, would uniformly improve the worst-case adjusted un-weighted test. This is at least a noteworthy property.

2.5. Adaptive designs are going beyond sample size reassessment

The weighting issue is particularly disturbing for the simple scenario of sample size reassessment. However, there are many other types of adaptations that might be of relevance in a clinical trial: Adding or skipping interim analyses, changing the error probability spending function, adaptations of test statistics (e.g. estimating scores to be used in the test statistics at subsequent stages), selecting subgroups, change or modification of the primary endpoint, discontinuing (or adding) treatment arms, changing the goal (e.g. switching between superiority and non-inferiority), changing the randomization ratio and others. I think that future merits of flexible designs will mainly arise from these more fundamental design changes. Theoretically, it may be possible that a complex decision procedure based on different aspects (efficacy, safety, patient burden and costs) could be pre-defined in a concise way. However, it will always be a hardly tractable task to derive the frequentist properties of such decision procedures considering the many parameters and assumptions behind such an exercise. It sounds like a contradiction, but quite a few researchers are appreciating the simplicity how fully flexible design achieves control of the family-wise error (FWE) rate in the strong sense by an application of the closed testing principle (considering the relatively small gain to be expected from fully modelling complicated decision procedures).

2.6. Estimation

To obtain ‘good’ estimates of the treatment effect from a clinical trial is an important task, e.g. for judging the risk–benefit relationship. However, estimation is already a demanding issue in conventional group sequential designs. In fully flexible designs, due to the lack of a pre-specified adaptation rule the sample space is not fully defined in advance and hence there are problems with quantities such as bias. There is an increasing amount of research on point and interval estimates in flexible designs including scenarios of treatment selection [7, 28–30], but much further work is necessary to understand the impact of flexibility on estimation.

2.7. Feasibility

Flexible designs require a careful planning and improved logistics to maintain integrity and persuasiveness of the results. The control of the information flow is crucial as plenty of un-blinded

DOI: 10.1002/sim

material may be needed in interim analyses to achieve good decisions. This seems to be the real challenge in practice, irrespective of the rather philosophical discussion whether the frequentist machinery offers satisfactory procedures to deal with the case of a random sample size which follows an unspecified distribution. (Clearly, a rather unsatisfactory alternative is just not to allow what we cannot handle in a nice and neat way.) Who decides when and which data can be looked at, who is entitled to suggest, who to enforce trial adaptations, who has access to un-blinded material? How can sufficient confidentiality be maintained to prevent bias resulting from the flow of un-blinded information? How are decisions communicated, how can study procedures be established which allow a rapid adaptations (e.g. when treatments are stopped after an adaptive interim analysis) and how the people outside the trial be convinced that no flaws and bias have been caused by this learning from experience type of approach used within a single trial? Related to the last question, there is an important issue of how to report the results of adaptive designs. Will it be possible in scientific journals to get enough space for a thorough description of the design, the motivations for and the type of adaptations actually performed? A recent review of the literature reveals a rather disappointing picture on the average quality of planning, analysing and reporting adaptive designs [31]. From our experience we know that applications of fully flexible designs can be handled [32], but it will need a rapid collection of clean data and an intensive involvement of statistical expertise under 'real-time' conditions. The latter may increase the fun or frustration of biometricians, depending on their understanding of the job. I believe that feasibility in practice is and will be the critical issue against a wide application of adaptive designs in clinical trials. Education of statisticians and applied researchers and guidelines, e.g. in the field of drug development, on planning, performing, analysing and reporting adaptive designs will be required. I believe that if investigators consider flexibility as an option in the planning phase but end with good reasons to perform a conventional trial, this would help to explore the problems in the design phase.

3. TESTING A LARGE NUMBER OF HYPOTHESES

Now a situation is considered which biometricians are confronted with more and more. In gene expression or gene association studies, researcher use to deal with a large number of markers, for example, when they are trying to identify markers that may help to discriminate between patients responding or not responding to a particular medical therapy. The pressure on statistical methods is high because such types of studies have been and are strongly promoted in the medical community with the promise to arrive at individualized therapies based on a prediction of the outcome from a patient's own genetic data. Due to the large number of hypotheses (which can go up to ten thousands) and the constraint on financial resources, the sample size per hypothesis is often rather small. When applying multiple test procedures to control the probability of false-positive decisions, the low sample size generally leads to a low power for identifying relevant markers. Therefore, alternative sequential screening procedures have been proposed by various authors. Early stages with a low sample size per hypothesis are used to identify promising candidates which at later stages are investigated with larger sample sizes. To demonstrate the method we again use a very simple model and consider a set of m_1 one-sided tests of the means of normal distributions with known variance ($\sigma^2=1$) with null and alternative hypotheses: $H_{0i}:\mu_i=0$ *versus* $H_{1i}:\mu_i>0,\ \ i=1,2,\ldots,m_1$. We assume independence of observations across hypotheses. The sample size is limited to a total of N. For the fixed sample size test we distribute N/m_1 observations equally over each of the m_1 hypotheses tests. For convenience we use the individual one-sided

p-values $p_1, p_2, \ldots, p_{m_1}$ for the test decisions. The critical regions for the individual tests are defined by $\{p_i \leqslant \gamma\}, i=1,\ldots,m_1$. We apply a multiple test that tries to control the FDR [33]. The FDR is defined by $\mathrm{FDR}=E(V/R|R>0)P(R>0)$, where V is the number of erroneously rejected null hypotheses and R the total number of rejections. Hence, the FDR is the expected proportion of erroneous rejections among all rejections, if no rejection occurred it is set to zero. It is a more liberal criterion than the FWE rate (except in the situation that all null hypotheses are true). It is particularly appealing in high-dimensional problems where the experimenter tends to focus strongly on the few significant results discovered. Then it is convenient to know that the expected fraction of false-positive results among those identified to be significant is controlled independently of how many and which hypotheses have been identified. It is worth mentioning that there is a certain correspondence to characterize the statistical properties of diagnostic tests in medicine by sensitivity and specificity on the one hand and positive and negative predictive value on the other hand. Whereas controlling the FWE is related to the concept of sensitivity and specificity, controlling the FDR is related to the concept of predictive values. Hence, it is not surprising that under some simple assumptions there is a bridging to Bayesian inference: The quantity $E(V/R|R>0)$, called the ‘positive FDR’, is equal to the posterior probability that a null hypothesis is true, given that it has been rejected [34].

There are different ways to achieve control of the FDR. Since V and R are both random one way [34] is to estimate the FDR from the sample, $\widehat{\mathrm{FDR}}_\lambda(\gamma)=\widehat{\pi}_0\gamma m_1/\max(\#\{p_i\leqslant\gamma\},1)$, where λ is a positive pre-defined constant, $0<\lambda<1$, e.g. $\lambda=0.5$, and $\widehat{\pi}_0=\#\{p_i>\lambda\}/[(1-\lambda)m_1]$ is an estimate of the proportion π_0 of true null hypotheses. Here, the number of observed p-values in the right tail of the distribution $(>\lambda)$ is simply compared with the corresponding number that would be expected if all null hypotheses were in fact true. Now, the critical boundary γ for the individual p-values can be determined by numerically searching for the supremum of γ such that the inequality $\widehat{\mathrm{FDR}}_\lambda(\gamma)\leqslant\alpha$ still holds in the given data set. Note that estimating the FDR (as a function of γ) is extremely simple because it requires only counting how many p-values in the sample fall below γ and how many lie above λ. Storey *et al.* [35] have shown that the expected value of $\widehat{\mathrm{FDR}}$ is an upper boundary for the true FDR if the p-values under the null hypotheses follow independent and uniform distributions. We later apply the simple assumption that the same alternative $\mu_i=\Delta$ holds whenever the individual null hypothesis is not true. Then we can use an asymptotic value γ (for large m_1) to calculate the per hypothesis power (here identical to the expected fraction of true rejections among the alternatives). We now assume that $m_1=6000$ null hypotheses are tested, and the fraction of true null hypotheses is $\pi_0=0.99$. Then the above procedure for an FDR of 0.05 in a single-stage design with a sample size of 8 per hypothesis $(N=8m_1=48\,000)$ for common alternatives with $\Delta=1$ achieves an asymptotic power of 0.186 and for $\Delta=0.5$ the power is as low as 0.0022 (determined by simulation since using an asymptotic critical boundary γ is not appropriate in this scenario). There are many different proposals for sequential procedures to improve the poor performance of single-stage designs, for references see [36, 37]. Here, I will discuss only some results for a simple example of such designs, the good old pilot design.

3.1. The pilot design

At the first, pilot stage, a proportion r of the total sample size N is distributed equally among the m_1 null hypotheses, leading to a sample size $n_1=rN/m_1$ per hypothesis. All (promising) hypotheses with a p-value $p_i^{(1)}\leqslant\gamma^{(1)}$ in the pilot phase, $H_{0i_1},\ldots,H_{0i_{m_2}}$, say, are carried on to the second, the main (confirmatory) stage.

DOI: 10.1002/sim

At the second stage, the remaining observations $(1-r)N$ are equally distributed among the m_2 hypotheses selected at the first stage. This leads to a sample size $n_2=(1-r)N/m_2$. A separate multiple testing procedure to control the FDR is applied to the second-stage data, now only referring to the reduced set of individual null hypotheses $H_{0i_1},\ldots,H_{0i_{m_2}}$.

This procedure is adaptive as the number of hypotheses m_2 and the sample size n_2 at the confirmatory part are random, depending on the outcome of the pilot phase. However, the design adheres to a fixed adaptation rule. Also, the critical boundary applied in the confirmatory main second stage is 'adaptive' in the sense that the critical boundary is estimated and depends on the data. I do not consider 'integrated designs' with an internal pilot phase, where the first-stage data are also used for inference in the final analysis (which do not noticeably improve the behaviour of the asymptotically optimal simple pilot approach but, by using all data collected on a hypothesis for the final test decision, is more robust against design misspecifications in the planning phase [36–38]).

Asymptotically optimal designs can be calculated by maximizing the power as a function of r and γ_1, given $N, m_1, \pi_{0,}\Delta$ and the FDR [36]. As an example taking again $N=48\,000, m_1=6000, \pi_0=0.99$ and $\mathrm{FDR}=0.05$, the optimal fraction of the total sample size to be used at the first pilot stage is $r=0.681$ for $\Delta=1$ and $r=0.535$ for $\Delta=0.5$, respectively. Hence, small sample sizes of $n_1=5.45$ and 4.28 (non-integer numbers from the asymptotic optimization) are applied at the first stage, respectively. The corresponding critical values for screening promising p-values in the pilot phase are $\gamma_1=0.112$ and 0.059, respectively. This means that, in case of existing strong effects, more hypotheses should be carried on to the confirmatory stage for which we then require lower sample sizes (n_2 values around 21) compared with the case with lower effects (n_2 values around 61). The overwhelming advantage of the simple adaptive screening procedures can be seen from the power values (calculated for the fixed optimal decision boundaries and confirmed by simulations): We obtain a power 0.847 for $\Delta=1$ and 0.255 for $\Delta=0.5$ compared with 0.186 and a value below 0.01 given above for the single-stage reference design. When looking at the expected sample size at the second stage, we realize that we should choose a design where the power for the test of the screened hypotheses at the second stage is very large: Whenever we have correctly screened an alternative, the best philosophy is to avoid any type II error in the confirmatory part. The method seems to work also in the unknown variance case and for correlated test statistics [36, 38]. It can also be extended to more than two stages [37]. It is not intended to discuss any details how this procedure may be improved. The main message is that adaptive multi-stage designs are a promising tool for handling such a non-appealing problem with large number of questions and a limited total sample size for the whole experiment. The good old-screening philosophy of taking small samples at early stages of experimentation to identify promising hypotheses and to apply larger sample sizes in the strongly reduced set of screened hypotheses at later stages has to be reconsidered for this type of a problem. The crucial issue is not the type of design (pilot or integrated design) and the number of stages of the type of error control, but it is important to get rid of the non-promising hypotheses as early as possible [37]. Note that in such a scenario even randomly selecting hypotheses for a single-stage design could improve on power [39].

3.2. *What to expect for prediction of clinical outcome*

Finally, we change to the more delicate problem of prediction scores that have to be constructed and estimated from a large pool of such genetic markers in limited samples. We will look at the statistical properties of such scores, e.g. in terms of how well they can predict the outcome of a medical therapy in future patients. Again, a simple scenario similar to those considered above is investigated in the last part of the paper. We assume that we have samples available of patients responding to a particular treatment ($n_{\mathrm{r}} = 50$) and of patients not responding to the treatment ($n_{\mathrm{nr}} = 50$). There are again $m = 6000$ markers for which responders and non-responders follow independent normal distributions with means μ_{r_i} and μ_{nr_i}, $i = 1, \ldots, m$, respectively (variance known $\sigma^2 = 1$). For $m\pi_0$ markers, the means are equal ($\mu_{\mathrm{r}_i} = \mu_{\mathrm{nr}_i}$, the null hypothesis is true), for the other $m_e = m(1-\pi_0)$ markers the same alternative $\mu_{\mathrm{r}_i} - \mu_{\mathrm{nr}_i} = \Delta$ holds for the average difference between responders and non-responders. These assumptions with respect to number of markers and sample size roughly mimic the existing clinical research work [40]. Without loss of generality let us assume that the effective markers are denoted by the random variables $Y_1, Y_2, \ldots, Y_{m_e}$, whereas $Y_{m_e+1}, \ldots, Y_{m_1}$ refer to the remaining non-effective markers. Clearly, in the highly symmetric and simple scenario assumed above, the best prediction of response for a future patient would be based on a linear function of his measurements $Y_1, Y_2, \ldots, Y_{m_e}$ for the m_e effective markers using equal weights: $\sum_{i=1}^{m_e} (\mu_{\mathrm{r}_i} - \mu_{\mathrm{nr}_i}) Y_i = \Delta \sum_{i=1}^{m_e} Y_i$. To assess the predictive ability of scores in the following, we will consider ROC curves resulting from varying threshold values for the score. At first, we pretend to know all the effective markers and the correct (equal weights) for the markers in the linear score. As a benchmark we would like to achieve a prediction such that the ROC curve of the corresponding score crosses the point which satisfies the condition: sensitivity = specificity = 0.9. To obtain such an ROC curve the constant effect size Δ required among the effective markers clearly depends on their number m_e. For $m_e = 1, 10$ and 60, the required effect sizes $\Delta(m_e)$ are 2.56, 0.81 and 0.33, respectively. These numbers show that if in fact few effective markers would exist which allow a good prediction, then these markers can easily be identified from responder non-responder studies applying small sample sizes. On the other hand, for many effective markers with only moderate effect sizes the identification of effective markers in such a large set of candidates will become difficult (as can be suspected from the arguments in the previous section).

To look into the problem in more details we now drop the assumption of knowing the relevant effective markers and assume that (1) the markers for the score have to be selected among a large set of candidates and (2) the coefficients of the selected markers in the score have to be estimated, both simultaneously from the samples of 50 responders and 50 non-responders. For demonstration we use two extreme selection procedures. (i) The first selects the significant markers by a fixed sample size multiple test procedure for the m_1 comparisons of the means between responders and non-responders estimating and controlling a particular FDR in the way described in the previous section. This ‘cautious’ strategy is a safeguard against the selection of markers (and the construction of a score) in case there is in fact no noticeable influence of any marker on the response to therapy. It accounts for parameter constellations close to the global null hypothesis. (ii) The second, the ‘optimistic’ or unprotected, strategy simply selects the k best markers (which show the largest mean difference between the sample of responders and non-responders). Here, the experimenter is convinced that effective markers must exist among such a large number of candidates. Basic principles are put forward that a genetic influence on a particular biological process should be present and, considering the broad range of candidates, should show up at least in a few of them. Sometimes the arguments are close to a proof by a disaster: If no effective markers would exist this would be a disaster for the theories, for the numerous research groups dealing with these kinds of problems and for funding policies.

For estimation of the score for both selection procedures we simply plug in the estimates of the mean differences between responders and non-responders of the k selected markers in the training sample: Exploiting the assumption of independent and normally distributed markers with a common known variance leads to the score $\widehat{D}=\sum_{j=1}^{k}(\widehat{\mu}_{\mathrm{r}_{i_j}}-\widehat{\mu}_{\mathrm{nr}_{i_j}})Y_{i_j}$, where $Y_{i_j}, j=1,\ldots,k$, are the measurements of the k selected markers of a future patient for which a prediction of response is performed. If $\widehat{D}>c$ we predict a response, otherwise a non-response. The predictive ability of such a score is assessed by the ROC curve where sensitivity (for response) is plotted against (1−specificity) as a function of c.

3.3. A simulation study

To obtain an impression of the statistical properties of data-driven scores, the two simple selection procedures combined with the simple estimation procedure for the score are investigated in a simulation study. In the samples of 50 responders and non-responders each, a total of $m_1=6000$ independently and normally distributed markers with known variance ($\sigma^2=1$) are determined. For the simulation the parameter constellations are varied in terms of the number of effective markers $m_e=10, 60$. The constant effect size Δ for the effective markers is determined as a function of m_e so that for known parameters the optimal achievable ROC curve passes through the benchmark point (0.9,0.9) in the (sensitivity, specificity) plane. This leads to $\Delta=0.81$ for $m_e=10$ and $\Delta=0.33$ for $m_e=60$. Simulations were also run under the global null hypothesis of no effective marker at all ($m_e=0$). For the cautious selection procedure, the targeted FDR $(=0.05, 0.20, 0.30, 0.50)$ for the optimistic procedure the number k $(=5, 10, 20, 30, 60, 80)$ of selected best markers has been varied. Figure 1 shows some results from an ongoing joint research project with Alexandra Goll. Under the simple assumption made above with $m_e=10, 60$ an FDR of 0.2 turns out to be a reasonable compromise to obtain a good selection in the cautious strategy just as the selection of the 20 best markers works quite well in both scenarios for the optimistic strategy. Values of 0.05 or lower for the FDR (as the significance levels usually applied in clinical trials) seem to be too restrictive for the selection of markers. The simulations show (at least in our simple model) that a better prediction is achieved by tolerating a certain proportion of nuisance markers to be included in the score. Including more nuisance markers also increases the chance of capturing effective markers. Figure 1 shows 100 simulated ROC curves resulting from selection procedures. The solid curve is the optimal benchmark curve through the point (0.9, 0.9) which could be achieved in the theoretical case that the effective markers and their effects are known. The dashed mean ROC curve (averaging over sensitivity for given specificity) is calculated from 1000 samples (including only samples where markers have been selected). The probability of selecting any marker is denoted by p_{select}. The general tendencies that can be seen from the results using $\mathrm{FDR}=0.2$ and $k=20$ in the two selection strategies, respectively, are similar for other choices of the selection parameters. In case of 10 markers with a common clear effect size of 0.81 (first row of the panels in Figure 1), the selection leads to an estimated score which, although being clearly away from the achievable optimum, still allows a reasonable prediction leading to ROC curves with points providing average sensitivities and specificities both above 0.8. If we select only the 10 best markers, the results of the optimistic strategy is better than for $k=20$ as shown in Figure 1. For $k=10$ (data not shown) the results are very similar to the cautious strategy with $\mathrm{FDR}=0.2$, in fact the optimistic strategy in this situation works with an FDR of about 0.23. Things get considerably worse for the case of 60 markers with rather small common effect of 0.33, for both election procedures (second row of the panels in Figure 1). The estimated ROC curves are very poor and variable. Even if exactly the 60 best markers are selected simulations show that there is no dramatic improvement in the predictive ability of the score. Note that the cautious strategy applying an $\mathrm{FDR}=0.2$ does not select any marker at all with an estimated probability of $1-\widehat{p}_{\mathrm{select}}=0.44$, whereas by definition

DOI: 10.1002/sim

the unprotected procedure always ends with a score. This has dramatic consequences under the global null hypothesis (third row of panels in Figure 1): Now the unprotected procedure leads to a completely non-informative score in all cases, whereas for the cautious procedure by definition the probability to end with a selection of markers and building a score is targeted at the FDR$=0.2$. This demonstrates the dilemma of the task we are faced with. It may be possible to improve the selection and estimation procedures (note, however, that we have already exploited the assumption of independence and constant variance across markers) but a contradiction will remain: Being cautious may help not to produce too many nuisance results if the postulated relationships do not exist. Being more optimistic and liberal may improve the results, particularly if the true state of nature is close to what the selection procedure is targeted at (e.g. if in our case k is close to the actual number of effective markers). However, the unprotected procedure is vulnerable for the situations of no existing effects. Here it will always produce nuisance results, and due to the large numbers of tests the magnitude of the observed best effects may mislead the experimenter. Scores proposed in the literature, e.g. to predict response to cancer therapy, seem to mirror these types of problems when validation results in external samples are reported [41].

4. DISCUSSION

Remembering the last 40 years in medical statistics and going back to the end of the sixties of the last century, things started in a rather unorganized and unregulated way. Guidelines and standards have been around for quite a time in other areas such as quality control. There was a tendency to apply rather simple methods while statisticians had to struggle with writing their own programs on insufficient hardware. Several issues considered to be crucial nowadays, such as multiplicity of statistical testing in clinical trials, were not generally agreed to be relevant. Important concepts such as interval estimation have been used rather rarely. As a reaction to this pioneering time, the following 30 years created a large amount of regulations and guidelines. This brought more and more rules into the game, e.g. in the development process of new medical treatments. The parties involved in these processes tended to adhere closely to the principles (more or less agreed by the statistical community as a paradigm to be presently applied). It is in the nature of 'rules' that experimenters hesitate to adopt methods that could be considered to break the rules, since they do not want to be blamed later when a trial failed for whatever reasons. In clinical trials, this led to the development of very rigid study protocols where all details of statistical planning and analysis are laid down *a priori* (which as I recollect led to a big step forward in the average quality of results and conclusions). This is laudable *per se*; however, real life is more complicated. Hence, whenever changes of a protocol seem to be necessary, it has become a common practice to achieve the required flexibility by writing amendments to the protocol. In the ethical committee of the Medical University of Vienna in 2005, on average we received more than two such amendments per protocol for drug trials (which are considered important enough to be reported to the ethical committee). Fully adaptive or flexible designs acknowledge that mid-trial design may become a necessary or at least useful option to improve an ongoing trial, particularly if unforeseen evidence from inside or outside the trial changes risks, utilities and costs assumed in the planning phase. The price to be paid for flexibility has been widely discussed. Among the methodological problems of frequentist inference in the situation that the sample space and the distribution of the sample size are not known in advance, the main problems will arise at the application side: Which measures can be taken to maintain the integrity and persuasiveness of results (considering the necessity of decisions during an ongoing trial based on un-blinded data and possible modifications of the hypotheses connected with adaptations)?

A growing area and a new challenge in medical statistics is how to deal with a large number of hypotheses in clinical trials, e.g. in gene expression or gene association studies. More and more clinical trials in drug development are accompanied by building up banks of blood and tissue

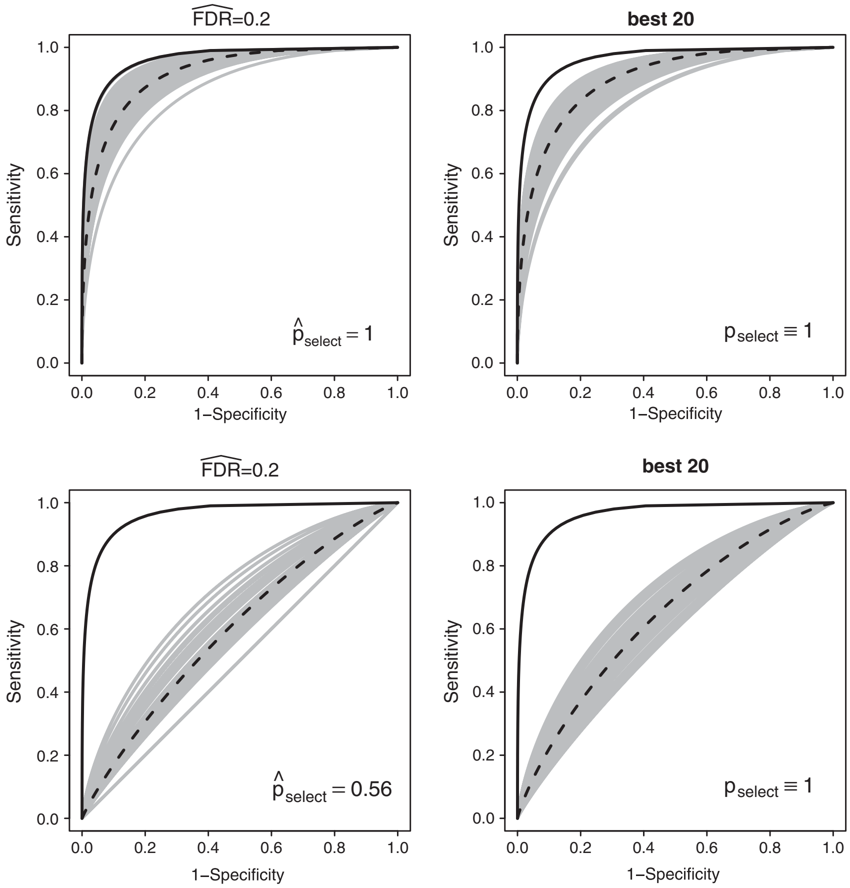

Figure 1. Simulated ROC curves for the prediction of the outcome in a future patient based on a score determined from samples of 50 patients responding and 50 patients not responding to a particular therapy, respectively. A total of 6000 markers are measured, which are assumed to be independent and normally distributed with known variance. The number of effective markers is varying between $m_e = 10$ (first row of the panels), $m_e = 60$ (second row) and $m_e = 0$ (last row). The three left panels show the results for the selection of markers based on a multiple test with FDR$=0.2$, the right panels from selecting the best observed 20 markers for the score. The solid black curve gives the optimal achievable benchmark ROC curve, where the common effect size $\Delta(m_e)$ among the m_e effective markers is fixed such that the optimal benchmark curve crosses through the point $(0.9, 0.9)$. The probability to select any markers is denoted by p_{select}. Hundred simulated ROC curves are shown in grey colour, whereas the dotted black curve is an average over 1000 simulated ROC curves, averaging only over samples where 'effective' markers have been selected at all.

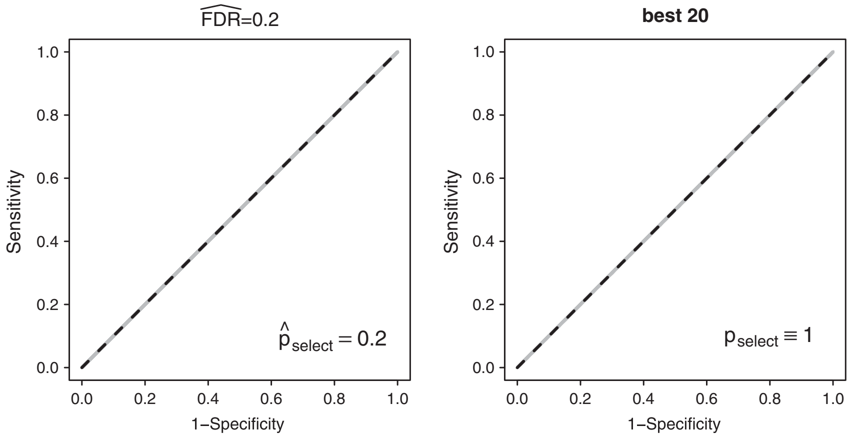

Figure 1. *Continued.*

samples to be later used for genetic analyses. This creates various problems of how to inform the patient and get consent. As a member of ethical committees for decades I have hardly ever observed investigators telling the patients that they are looking for a needle in the haystack. One of the promises in this area is individualization of therapy based on the prediction of a patient's outcome from individual genetic information. Limitation of resources, however, is a serious problem in such a situation. We deal with the wrong asymptotic, an increasing number of hypotheses and limited total costs (which lead to small number of samples per hypothesis). I have sketched that simple classical (and adaptive) screening-type designs may substantially improve the chances to identify relevant markers. New concepts of error control, such as the false discovery rate (FDR) bridging to Bayesian methods of inference, have been established and seem set to become a standard in this area. In the recent years, a lot of research has been performed for such designs. They have become practically relevant since the technical equipments for individually designing measurement devices have become available. Due to the clear superiority of methods based on flexible medical devices, statisticians should strongly ask for their further development instead of only dealing with methods based on standard devices already available.

At last based on simple models and methods, it has been demonstrated that it will be hard to meet the promises when statistical methods are used to perform prediction for individual patients. The large number of candidate variables creates a problem for selection of markers and estimation of suitable predictors from limited samples. Although the results achieved by the two simple philosophies for selecting and estimating prediction scores may be improved by more sophisticated methods, the basic trade-off between taking care not to produce too many nuisance findings and optimism leading to a more liberal selection of markers will remain. This should simply remind us that, as elsewhere in life, there is no such thing as a free lunch also in medical research.

ACKNOWLEDGEMENTS

The author would gratefully acknowledge the continuous support of W. Brannath, F. König, M. Posch and S. Zehetmayer. He is particularly indebted to A. Goll for supplying some results from a joint research project supported by the Austrian FWF, grant P18698-N15, which will be part of her Ph.D. thesis. The author also thank the associate editor for his suggestions to improve the manuscript.

REFERENCES

1. Flurnoy N, Rosenberger WF (eds). Special issue on adaptive designs in clinical trials. *Journal of Statistical Planning and Inference* 2006; **136**:1747–1955.
2. Brunner E, Schuhmacher M (eds). Adaptive designs in clinical trials. *Biometrical Journal* 2006; **48**:491–737.
3. Burman CF, Sonesson C. Are flexible designs sound? *Biometrics* 2006; **62**:664–669; 681–684.
4. Proschan MA. Discussion of the paper: Burman CF, Sonesson C. Are flexible designs sound. *Biometrics* 2006; **62**:674–676.
5. Bauer P. Multistage testing with adaptive designs; with discussion. *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130–148.
6. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
7. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.
8. Cui L, Hung HMJ, Wang S. Modification of sample size in group sequential trials. *Biometrics* 1999; **55**:321–324.
9. Hedges LV, Olkin I. *Statistical Methods of Meta-analysis*. Academic Press: New York, London, 1985.
10. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 2002; **97**:1034–1041.
11. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
12. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
13. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**:261–277.
14. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
15. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
16. Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
17. Müller H-H, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
18. Posch M, Timmesfeld N, König F, Müller H-H. Conditional rejection probabilities of Student's *t*-test, (non)-stochastic curtailment, and design adaptations. *Biometrical Journal* 2004; **46**:389–403.
19. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.
20. Tsiatis AA, Metha C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
21. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.
22. Brannath W, Bauer P, Posch M. On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference* 2006; **136**:1956–1961.
23. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971–993.
24. Bauer P, König F. The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine* 2006; **25**:23–36.
25. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics in Medicine* 2003; **22**:953–969.
26. Denne JS. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; **20**:2645–2660.
27. Wassmer G. *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien*. Verlag Alexander Mönch: Köln, 1999.
28. Brannath W, Koenig F, Bauer P. Estimation in flexible two stage designs. *Statistics in Medicine* 2006; **25**: 3366–3381.
29. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **30**:3697–3714.

DOI: 10.1002/sim

30. Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. 2007, to appear.
31. Bauer P, Einfalt J. Application of adaptive designs—a review. *Biometrical Journal* 2006; **48**:493–506.
32. Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis G, Linssen G, Tebbe U, Schroder R, Tiemann R, Maching T, Neuhaus KL. The $\mathrm{Na}^+/\mathrm{H}^+$ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction—results of the evaluation of the safety and cardioprotective effects ofeniporide in acute myocardial infarction (ESCAMI) trial. *Journal of the American College of Cardiology* 2001; **38**:1644–1650.
33. 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.
34. Storey JD. A direct approach to false discovery rate. *Journal of the Royal Statistical Society*, *Series B* 2002; **64**:479–498.
35. 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.
36. Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* 2005; **21**:3771–3777.
37. Zehetmayer S, Bauer P, Posch M. Optimized multi-stage designs for a large number of hypotheses. 2007, submitted.
38. Goll A, Bauer P. Two-stage designs applying methods differing in costs. *Bioinformatics* 2007; **23**:1519–1526.
39. 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.
40. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation study. *The Lancet* 2005; **365**:488–492.
41. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrys for cancer outcomes and correlates: an empirical assessment. *The Lancet* 2003; **362**:1439–1444.

DOI: 10.1002/sim