# ADAPTIVE SEAMLESS DESIGNS: SELECTION AND PROSPECTIVE TESTING OF HYPOTHESES

Christopher Jennison
Department of Mathematical Sciences,
University of Bath, Bath BA2 7AY, U. K.
*email:* cj@maths.bath.ac.uk
and
Bruce W. Turnbull
Department of Statistical Science,
227 Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, U. S. A.
*email:*bwt2@cornell.edu

July 21, 2007

**SUMMARY**

There is a current trend towards clinical protocols which involve an initial "selection" phase followed by a hypothesis testing phase. The selection phase may involve a choice between competing treatments or different dose levels of a drug, between different target populations, between different endpoints, or between a superiority and a non-inferiority hypothesis. Clearly there can be benefits in elapsed time and economy in organizational effort if both phases can be designed up front as one experiment, with little downtime between phases. Adaptive designs have been proposed as a way to handle these selection/testing problems. They offer flexibility and allow final inferences to depend on data from both phases, while maintaining control of overall false positive rates. We review and critique the methods, give worked examples and discuss the efficiency of adaptive designs relative to more conventional procedures. Where gains are possible using the adaptive approach, a variety of logistical, operational, data handling and other practical difficulties remain to be overcome if adaptive, seamless designs are to be effectively implemented.

*Key words:* Adaptive designs; Clinical trials; Closure principle; Combination tests; Early data review; Enrichment designs; Flexible design; Group sequential tests; Interim analysis; Meta-analysis; Multiple testing; Non-inferiority testing; Nuisance parameters; Phase II/III designs; Sample size re-estimation; Seamless designs; Targeted designs; Two-stage procedure; Variance spending.

# 1 Introduction

In this paper we provide a review, examples and commentary on adaptive designs for clinical trials. This article is based on parts of a tutorial and short course that were given at the 62nd Deming Conference on Applied Statistics, held in Atlantic City NJ, December 4–8, 2006.

The topic of adaptive design for clinical trials is very broad and can include many aspects — Chow and Chang (2007). Here we concentrate on how interim (unblinded) estimates of primary outcome variables may be used to suggest modifications in the design of successive phases of a confirmatory trial. These may be changes in the sample size, design effect size, power curve, or treatments to be studied; they may be restrictions on the eligible patient population, or modifications of the hypotheses to be tested. Of course the drug development process has always consisted of a series of studies where the results of one trial influence the design of the succeeding study until the final confirmatory study (or studies) have been concluded. The difference now is (i) the studies are considered as phases of a single trial with little or no delay between them and (ii) all the data collected are combined to test the hypothesis (or hypotheses) ultimately decided upon. This latter is in contrast to the current conventional process in which primary inferences are based solely upon the data from the final confirmatory trial(s) and the design is strictly specified in the protocol. At first, this possibility would seem like "voodoo magic" to a frequentist statistician and open to all kinds of selection bias. However, statistical procedures have been developed in the past ten to fifteen years that allow data driven design modifications while preserving the overall type I error. As has been pointed out, the flexibility afforded by these procedures comes at a cost: there is a loss of efficiency (Jennison and Turnbull, 2003, 2006a, b) and some procedures can lead to anomalous results in certain extreme situations (Burman and Sonesson, 2006). There is still opportunity for abuse (Fleming, 2006) and regulatory authorities have been justifiably wary (EMEA, 2006, Koch, 2006, Hung, O'Neill, *et al,* 2006, Hung, Wang *et al,* 2006). Some of the perceived potential for abuse can be moderated by the presence of a "firewall", such as a DSMB or CRO, so that the sponsor remains blinded to outcome data to the extent that this is possible; see the discussion in Gallo (2006) and Jennison and Turnbull (2006c, p. 295).

# 2 Combination tests

## 2.1 The combination hypothesis

The key ingredient in most (although not all) adaptive procedures is the combination test. This is an idea borrowed from the methodology of meta-analysis. Suppose our experiment consists of $K$ stages. In stage $k$ ($1 \leq k \leq K$), we denote by $\theta_k$ the treatment effect for the endpoint, patient population, etc., used in this stage and we test

$$H_{0k}\colon\ \theta_k \leq 0 \quad \text{versus} \quad H_{1k}\colon\ \theta_k > 0.$$

The combination test is a test of the combined null hypothesis

$$H_0 \ = \ \bigcap_{k=1}^{K} H_{0k}$$

against the alternative:

$$H_A\colon \text{ At least one } H_{0k} \text{ is false.}$$

## 2.2 Interpreting rejection of the combination hypothesis

In some cases, all hypotheses $H_{0k}$ are identical and the interpretation of rejection of $H_0$ is clear. However in an adaptively designed study, any changes in treatment definition, primary endpoint and/or patient eligibility criteria, etc., imply that the $H_{0k}$ are *not* identical. When the $H_{0k}$ differ, rejection of $H_0$ does *not* imply rejection of any one particular $H_{0k}$. Typically, interest lies in the final version of the treatment, patient population and primary endpoint and its specific associated null hypothesis, $H_{0K}$. Then, special methods are needed to be able to test this $H_{0K}$ in order to account for other hypotheses having been tested "along the way" and the dependence of the chosen $H_{0K}$ on previously observed responses.

We will use $P$-values (or $Z$-values) to assess evidence for or against $H_0$. Let $P_k$ be the $P$-value from stage $k$ and $Z_k = \Phi^{-1}(1 - P_k)$ the associated Z-value.

## 2.3 Combining $P$-values

Becker (1994) and Goutis *et al.* (1996) review the many different methods that have been proposed for combining $P$-values. The two most popular methods are

1. **Inverse $\chi^2$** (R. A. Fisher 1932)

$$\text{Reject } H_0 \text{ if } \quad -2\log(P_1 \ldots P_K) \; > \; \chi^2_{2K}(\alpha),$$

where $\chi^2_{2K}(\alpha)$ is the upper $\alpha$ tail point of the $\chi^2_{2K}$ distribution.

This method has been preferred in adaptive designs by Bauer and Köhne (1994) and colleagues.

2. **Weighted inverse normal** (Mosteller and Bush, 1954)

$$\text{Reject } H_0 \text{ if } \quad w_1 Z_1 + \ldots + w_K Z_K \; > \; z(\alpha),$$

where $z(\alpha)$ is the upper $\alpha$ tail point of the standard normal distribution, $w_1, \ldots, w_K$ are pre-specified and $\sum w_i^2 = 1$.

For adaptive designs this method has been used, explicitly or implicitly, by L. Fisher (1998), Cui, Hung and Wang (1999), Lehmacher and Wassmer (1999) and Denne (2001).

Another test is the "maximum rule" where $H_0$ is rejected if $\max\{P_1, \ldots, P_K\} \; < \; \alpha^{1/K}$. An example with $K = 2$ is the "two pivotal trial rule". If we require both $P_1 < 0.025$ and $P_2 < 0.025$, then we are operating with an overall $\alpha = 0.000625$. Other combination rules are possible, for example, the test proposed by Proschan and Hunsberger (1995) in the case $K = 2$ can be cast as a combination test — Proschan (2003). For a review of combination tests and their relation to methods of meta-analysis, see Jennison and Turnbull (2005).

The key point is that even if the studies are designed dependently, conditional on previously observed $P$-values, $P_k$ is still $U(0,1)$ if $\theta_k = 0$. As this distribution does not depend on the previous $P_i$s, this is also true unconditionally so $P_1, \ldots, P_K$ are statistically independent if $\theta_1 = \ldots = \theta_K = 0$. Similarly, $Z_1, \ldots, Z_K$ are independent standard normal variates in this case. Of course, the $P_k$s (or $Z_k$s) are *not* independent if $H_0$ is not true. Since we have defined composite null hypotheses, $H_{0k}$: $\theta_k \leq 0$, we should also comment on the general case where the set of $\theta_k$s lies in $H_0 = \cap H_{0k}$ but some are strictly less than zero. Here, it can be shown that the sequence of $P_k$s is jointly stochastically larger than a set of independent $U(0,1)$ variates and, hence, tests which attain a type I error rate $\alpha$ when $\theta_1 = \ldots = \theta_K = 0$ will have lower type I error probability elsewhere in $H_0$. For $P_k$ to be $U(0,1)$, we have assumed the test statistic to be continuous (or approximately so) but actually a weaker condition than stochastic independence is all that is required. The combination test will be a conservative level $\alpha$ test if the distribution of each $P_k$ is conditionally sub-uniform, which will typically be the case for discrete outcome variables — Bauer and Kieser (1999, p. 1835). We are also making some measurability assumptions about the nature of the dependence — Liu, Proschan and Pledger (2002).

# 3 Application 1: Sample size modification.

Many papers in the literature on adaptive designs focus on this application. This is the most straightforward problem as there is just one parameter of interest, $\theta_1 = \ldots = \theta_K = \theta$, say, and the null hypothesis is unchanged throughout the trial so all the $H_{0k}$ are identical. At an interim stage, the outcome data are inspected with a view to modifying the sample size.

## 3.1 Sample size modification based on interim estimates of nuisance parameters

The sample size needed to satisfy a power requirement often depends on an unknown nuisance parameter. Examples include: unknown variance ($\sigma^2$) for a normal response; unknown control response rate for binary outcomes; unknown baseline event rate and censoring rate for survival data.

*Internal pilots.* Wittes and Brittain (1990) and Birkett and Day (1994) suggest use of an "internal pilot study". Let $\phi$ denote a nuisance parameter and suppose the fixed sample size required to meet type I error and power requirements under a given value of this parameter is $n(\phi)$. From a pre-study estimate, $\phi_0$, we calculate an initial planned sample size of $n(\phi_0)$. At an interim stage, we compute a new estimate $\hat{\phi}_1$ from the data obtained so far. The planned sample size $n(\phi_0)$ is then modified to a new target sample size of $n(\hat{\phi}_1)$. Variations on this are possible, such as only permitting an increase over the original target sample size. Upon conclusion of the trial, the usual fixed sample size test procedure is employed. The problem is that the random variation in $\hat{\phi}_1$ implies that the standard test statistic does not have its usual distribution. For example, in the case of normal responses, the final $t$-statistic does not have the usual Student's $t$-distribution based on the sample size of $n(\hat{\phi}_1)$. This causes the final test to be somewhat liberal, with type I error rate slightly exceeding the specified value of $\alpha$. The intuitive reasoning for this is as follows. If $\sigma^2$ is underestimated at the interim stage, $n(\hat{\phi}_1)$ will be low and the final estimate of $\sigma^2$ will have less chance to "recover" to be closer to the true value. On the other hand, if $\sigma^2$ is overestimated at the interim stage, $n(\hat{\phi}_1)$ will be large and the additional data will give a more accurate estimate of $\sigma^2$. Thus, the final estimate of $\sigma^2$ is biased downwards and results appear more significant than they should. A similar inflation of type I error occurs in studies involving binary endpoints. However, numerical investigations have shown this inflation to be modest in most cases. Typically the increase is of the order of 10%, e.g., from $\alpha = 0.05$ to 0.055. Kieser and Friede (2000) were able to quantify the maximum excess of the type I error rate over the nominal $\alpha$ and this knowledge can

be applied to calibrate the final test when using a sample size re-estimation rule, so as to guarantee that the type I error probability does not exceed $\alpha$.

For normal responses, Denne and Jennison (1999) proposed a more accurate internal pilot procedure based on Stein's (1945) two-stage test. Stein's procedure guarantees exact control of error rates, but the final $t$-statistic uses the estimate $\hat{\sigma}^2$ from the pilot stage only. Since the analysis is not based on a sufficient statistic, it may suffer from inefficiency or lack of credibility. The approximation of Denne and Jennison (1999) follows Stein's approach but uses the usual fixed sample $t$-statistic and adjusts the degrees of freedom in applying the significance test. The adjustment is somewhat *ad hoc*, but their numerical examples show type I and II error rates are more accurately controlled.

In a comparative study with normal responses, estimating the variance would typically require knowledge of the two treatment arm means at the interim stage. This unblinding of the interim estimate of the mean treatment difference is often considered undesirable. Methods have been proposed to obtain an interim estimate of $\sigma^2$ without unblinding — Gould and Shih (1992), Zucker *et al.* (1999), Friede and Kieser (2001, 2002).

*Information monitoring.* Information monitoring is a natural extension of the internal pilot approach to incorporate group sequential testing. Because estimated treatment effects are used to decide when early stopping is appropriate, blinding of monitors to the estimated effect is not an issue.

Lan and DeMets (1983) presented group sequential tests which "spend" type I error as a function of observed information. In a $K$-stage design, at the end of the $k$th stage ($1 \leq k \leq K$) a standardized test statistic based on the current estimate $\hat{\theta}_k$ is compared with critical values derived from an error spending boundary. The critical values depend on the current information in the accumulated data, typically given by $\mathcal{I}_k = \{\mathrm{Var}(\hat{\theta}_k)\}^{-1}$ which of course depends on the accumulated sample size. A maximum information level is set to guarantee a specified type II error requirement. When the quantity $\mathcal{I}_k = \mathcal{I}_k(\phi)$ depends also on an unknown nuisance parameter $\phi$, Mehta and Tsiatis (2001) and Tsiatis (2006) suggest using the same procedure but replacing $\mathcal{I}_k(\phi)$ by $\mathcal{I}_k(\hat{\phi})$ at each stage, where $\hat{\phi}_k$ is the current estimate of $\phi$. The overall target sample size is updated based on the current estimate $\hat{\phi}_k$ so that the maximum information will be achieved. When the sample sizes are large, type I and II error rates are controlled approximately. However this control may not be achieved so accurately in small or moderate sample sizes. There are typically several types of

approximation going on here. In the case of normal response with unknown variance, $t$-statistics are being monitored but the error spending boundary is computed using the joint distribution of $Z$-statistics. Furthermore, only estimates of the observed information are available. The correlation between $Z$-statistics at analyses $k_1$ and $k_2$ ($k_1 < k_2$) is $\sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})}$ and it is natural to approximate this by $\sqrt{(\hat{\mathcal{I}}_{k_1}/\hat{\mathcal{I}}_{k_2})}$; however, it is preferable to write the correlation as $\sqrt{(n_{k_1}/n_{k_2})}$, where $n_{k_1}$ and $n_{k_2}$ are the sample sizes at the two analyses, since this removes the ratio of two different estimates of $\sigma^2$. Surprisingly perhaps, simulation studies show that even with moderate to large sample sizes the estimated information may actually decrease from one stage to the next, due to a high increase in the estimate of $\sigma^2$. A pragmatic solution is simply to omit any analysis at which such a decrease occurs and to make a special adjustment to spend all remaining type I error probability if this happens at the final analysis. Finally, we note that re-estimating the target sample size based on $\hat{\sigma}_k^2$ produces a downwards bias in the final estimate of $\sigma^2$ and inflation of the type I error rate for the same reasons as explained above for the Wittes and Brittain (1990) internal pilot procedure. In fact, repeated re-estimation of sample size exacerbates the problem since it enhances the effect of stopping when the current estimate of $\sigma^2$ is unusually low. In our simulations we have seen the type I error probability increase from $\alpha = 0.05$ to 0.06 or more for a $K = 5$ stage study with a target degrees of freedom equal to 50. For more accurate approximations, Denne and Jennison (2000, Sec. 3) suggest a $K$-stage generalization of their two-stage procedure (Denne and Jennison, 1999).

*Bauer and Köhne's (1994) adaptive procedure.* The combination test principle of Section 2 can be used to construct a two-stage test of $H_0$ with exact level $\alpha$, yet permit an arbitrary modification of the planned sample size in the second stage. Consider the case of comparing two treatments with normal responses. From the first stage, estimates $\hat{\theta}_1$ and $\hat{\sigma}_1^2$ are obtained from $n_1$ observations per treatment. The $t$-statistic $t_1$ for testing $H_0$ vs $\theta > 0$ is converted to a $P$-value

$$P_1 = Pr_{\theta=0}\{T_{2n_1-2} > t_1\}$$

or $Z$-value $Z_1 = \Phi^{-1}(1-P_1)$. The sample size for stage 2 may be calculated using $\hat{\sigma}_1^2$ as an estimate of variance — for example, following the method of Wittes and Brittain (1990). Posch and Bauer (2000) propose modifying sample size to attain conditional power at the original alternative, again assuming variance is equal to the estimate $\hat{\sigma}_1^2$.

In stage 2, estimates $\hat{\theta}_2$ and $\hat{\sigma}_2^2$ are obtained from $n_2$ observations per treatment. The $t$-statistic

$t_2$ for testing $H_0$ vs $\theta > 0$ based on stage 2 data is converted to a $P$-value

$$P_2 = Pr_{\theta=0}\{T_{2n_2-2} > t_2\}$$

or $Z$-value $Z_2 = \Phi^{-1}(1 - P_2)$. In the final analysis the null hypothesis $\theta \leq 0$ is tested using one of the combination methods described in Section 2, such as the inverse $\chi^2$ test or an inverse normal test with pre-assigned weights. We note that the test statistic in these cases is *not* a function of a sufficient statistic. Friede and Kieser (2001) compare this design with that of Wittes and Brittain (1990) with respect to power and sample size. Timmesfeld *et al.* (2007) consider adaptive re-designs which may increase (but not decrease) sample size and show how a test based on the final standard $t$-statistic can be constructed to meet a type I error condition exactly. Their method is also applicable in multi-stage procedures. Although the construction of an exact $t$-test is an impressive feat, the power advantage of the test proposed by Timmesfeld *et al.* (2007) over combination tests using non-sufficient statistics appears to be slight.

Lehmacher and Wassmer (1999) propose a generalization of the adaptive, weighted inverse normal test which facilitates exact $K$-stage tests when an exact $P$-value can be obtained from each group of data. Pre-assigned weights $w_1, \ldots, w_K$ are defined, then for $k = 1, \ldots, K$, test statistics are

$$Z(k) \;=\; (w_1 Z_1 + \ldots + w_k Z_k) \,/\, (w_1^2 + \ldots + w_k^2)^{1/2},$$

where the $Z_k$ are $Z$-statistics based on data in each group. The joint distribution of the sequence $Z(1), \ldots, Z(K)$ has the standard form seen in group sequential tests and a boundary with type I error rate $\alpha$ can be created in the usual way. For normal responses with unknown variance, the process for obtaining $Z_k$ from group $k$ data of $n_k$ observations per treatment is as follows: first calculate the $t$-statistic $t_k$ for testing $H_0$ vs $\theta > 0$, then convert this to the $P$-value $P_k = Pr_{\theta=0}\{T_{2n_k-2} > t_k\}$, and finally, obtain the Z-value $Z_k = \Phi^{-1}(1 - P_k)$.

### 3.2 Sample size modification based on interim estimates of efficacy parameters

When sample size is modified in response to an interim estimate of the treatment effect, it is typically increased because results are less promising than anticipated and conditional power is low. In this case a combination test using $P$-values from the different stages can be used to protect the type I error rate despite the adaptations employed. This flexibility will inevitably lead to a testing procedure which does not obey the sufficiency principle. Our investigations of rules for modifying sample size based on conditional power under the estimated effect size, an idea that appears to

have intuitive appeal, show that the resulting procedures are less efficient than conventional group sequential tests with the same type I and II error probabilities — Jennison and Turnbull (2003, 2006a, b). See also Fleming (2006) for a critical commentary on an adaptive trial design and the inefficiency arising from unequal weighting of observations. Burman and Sonesson (2006) point out how lack of sufficiency and lack of invariance in the treatment of exchangeable observations can lead to anomalous results. Admittedly, their examples concern extreme situations but they show just how far credibility could be undermined.

Adaptation ought to be unnecessary if the trial objectives and operating characteristics of the statistical design have been discussed and understood in advance. Thought experiments can be conducted to try to predict any eventualities that may occur during the course of the trial, and how they might be addressed. For example, discussion of the changes investigators might wish to make to sample size on seeing a particular interim estimate of the effect size could indicate the power they would really wish to achieve if this were the true effect size. It may seem an attractive option to circumvent this process and start the trial knowing that necessary design changes, such as extending the trial, can be implemented later using, say, the variance spending methodology of Fisher (1998). However, postponing key decisions can be costly in terms of efficiency (Jennison and Turnbull, 2006b).

We recommend that in a variety of situations, conventional non-adaptive group sequential tests should still be the first methodology of choice. These tests have been well studied and designs have been optimized for a variety of criteria. For example, Jennison and Turnbull (2006d) show how to design conventional group sequential boundaries when the sponsor team have differing views about the likely effect size and wish to attain good power across a range of values without excessive sample sizes under the highest effect sizes. With error spending (Lan and DeMets, 1983) and information monitoring (Mehta and Tsiatis, 2001), conventional group sequential tests already possess a great degree of flexibility. Incorporating adaptive choice of group sizes in pre-planned group sequential designs based on sufficient statistics, as proposed by Schmitz (1993), can lead to additional gains in efficiency but these gains are small and unlikely to be worth the administrative complications — see Jennison and Turnbull (2006a, d).

Of course, it is always possible that truly unexpected events may occur. In that case, modifications to the trial design can be implemented within a group sequential design following the approach of Cui *et al.* (1999) or the more general method based on preserving conditional type I error described by Müller and Schäfer (2001). Perhaps a statement that such adaptations might be

considered in exceptional circumstances would be a worthwhile "escape clause" to include in the protocol document.

# 4 Testing multiple hypotheses

In the following applications we will be testing multiple hypotheses. These may arise when considering multiple treatments, multiple endpoints, multiple patient population subsets, or both a superiority and a non-inferiority hypothesis. Even if there is only a single hypothesis, $H_{0K}$, to be tested at the final analysis, if the hypotheses $H_{0k}$ have been changing along the way we need to account for the selection bias caused by presence of other hypotheses which might have been selected as the final $H_{0K}$.

Suppose there are $\ell$ null hypotheses, $H_i$: $\theta_i \leq 0$ for $i = 1, \ldots, \ell$. A multiple testing procedure must decide which of these are to be rejected and which not. A procedure's ***familywise error rate*** under a set of values $(\theta_1, \ldots, \theta_\ell)$ is

$$Pr\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\} \;=\; Pr\{\text{Reject any true } H_i\}.$$

The familywise error rate is controlled *strongly* at level $\alpha$ if this error rate is at most $\alpha$ for all possible combinations of $\theta_i$ values. Then

$$Pr\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_\ell).$$

Using such a procedure, the probability of choosing to focus on any particular parameter, $\theta_{i^*}$ say, and then falsely claiming significance for the null hypothesis $H_{i^*}$ is at most $\alpha$.

There are a number of procedures that provide strong control; see, for example, Hochberg and Tamhane (1987) and Westfall and Young (1993). Here we concentrate on the *closed testing procedures* of Marcus *et al.* (1976) which provide strong control by combining level $\alpha$ tests of each $H_i$ and of intersections of these hypotheses.

We proceed as follows. For each subset $I$ of $\{1, \ldots, \ell\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Suppose we are able to construct a level $\alpha$ test of each intersection hypothesis $H_I$, i.e., a test which rejects $H_I$ with probability at most $\alpha$ whenever all hypotheses specified in $H_I$ are true. Special methods are required to test an intersection hypothesis $H_I = \cap_{i \in I} H_i$. For instance, tests of the individual $H_i$s could be combined using a Bonferroni adjustment for the multiplicity of hypotheses.

Often a less conservative method is available — for example, Dunnett (1955), Hochberg (1988) and Hommel (1988). If test statistics are positively associated, as is often the case, then the method of Simes (1986) gives a level $\alpha$ test of $H_I$ — Sarkar and Chang (1997), Sarkar (1998). In specific applications, there may be relationships between hypotheses. For example, the null hypotheses of no treatment effect in several small patient sub-groups imply there is no treatment effect in a larger combined sub-group.

The closed testing procedure of Marcus *et al.* (1976) can be stated succinctly:

> The hypothesis $H_j$: $\theta_j \leq 0$ is rejected overall if, and only if, $H_I$ is rejected for *every* set $I$ containing index $j$.

The proof that this procedure provides strong control of familywise error rate is immediate:

> Let $\tilde{I}$ be the set of all true hypotheses $H_i$. For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected. Since $H_{\tilde{I}}$ is true, $Pr\{\text{Reject } H_{\tilde{I}}\} = \alpha$ and, thus, the probability of a familywise error is no greater than $\alpha$.

We shall need to deal with combination tests of a hypothesis or an intersection hypothesis using data from the different stages between re-design points. At the final analysis, applying the closed testing procedure will require computation of various intersection hypothesis test statistics, each one being the combination of the corresponding test statistics from each stage. We review these methods by means of some examples. We will see that the closed testing approach is essential to avoid inflation of type I error probability by data-dependent hypothesis generation brought about by possible re-design rules. Principal references for this material include Bauer and Köhne (1994), Bauer and Kieser (1999), Bretz *et al.* (2006) and Schmidli *et al.* (2006). In some of the examples, we will also examine how the same problems might be handled using alternative, more "conventional" methods. In these examples, for pedagogical reasons (and for notational ease) we will look only at two-stage procedures, so there is just a single re-design point. This is the most common case. Note, however, that either or both stages could themselves have a sequential or group sequential design. In any case, one way to construct multi-stage adaptive designs is by recursive use of a two-stage design — Brannath, Posch and Bauer (2002). We start by considering the problem of treatment selection and testing.

# 5 Application 2: Treatment selection and seamless Phase IIb/III transition

## 5.1 Adaptive designs

There is an extensive literature on selection and ranking procedures dating back more than 50 years — see Bechhofer, Santner and Goldsman (1995). Group sequential methods for multi-armed trials have been surveyed in Jennison and Turnbull (2000, Chap. 16). However, here we desire a very specific structure: a treatment selection phase followed by a testing phase. We start in stage 1 (Phase IIb) with $\ell$ competing treatments (or dose levels) and a control. At the end of this stage, one or possibly several treatments are selected and carried forward into stage 2 (Phase III) along with a control arm. Usually Phase III is planned after the results of Phase IIb are received and there is a hiatus between the two phases. In a *seamless* design, the two stages are jointly planned "up front" in a single protocol, with no hiatus. A committee, preferably independent of the sponsor, makes the selection decision at the end of Phase IIb based on accumulated safety and efficacy outcomes, according to pre-specified guidelines. Some flexibility may be permitted, but if there is too much flexibility or if there is major deviation from selection guidelines, this may necessitate new approvals from institutional review boards (IRBs) and changes in patient consent forms, and the ensuing delay could negate the benefits of the seamless design. In an *adaptive* seamless design, data from both phases are combined, using methods described in Section 2, in such a way that the selection bias is avoided and the familywise type I error is controlled at the specified level $\alpha$. This is the flexible approach espoused by Bauer and Köhne (1994), Posch *et al.* (2005), Schmidli *et al.* (2006) and others. There are, however, other more traditional methods available, which we will also discuss.

Let $\theta_i, i = 1, \ldots, \ell$, denote the true effect size of treatment (or dose level) $i$ versus the control treatment. The two-stage design proceeds as follows:

*Stage 1 (Phase IIb)*

Observe estimated treatment effects $\hat{\theta}_{1,i}$, $i = 1, \ldots, \ell$.

Select a treatment $i^*$ to go forward to Phase III. Treatment $i^*$ will have a high estimate $\hat{\theta}_{1,i^*}$ (not necessarily the highest) and a good safety profile.

*Stage 2 (Phase III)*

Test treatment $i^*$ against control.

Formally, in order to reject $H_{i^*}$: $\theta_{i^*} \leq 0$, we need to reject each intersection hypothesis $H_I$ with $i^* \in I$ at level $\alpha$, based on combined Phase IIb and Phase III data. Here, $H_I = \cap_{i \in I} H_i$ states that $\theta_i \leq 0$ for all $i \in I$. Intuitively, treatment $i^*$ is chosen for the good results observed at this dose in Phase IIb and we must adjust for this selection effect when adding the Phase IIb data on dose level $i^*$ to the final analysis after Phase III. Under a global null hypothesis of no treatment effect for any treatment, the Phase IIb data on treatment $i^*$ should be viewed as *possibly the best results out of $\ell$ ineffective treatments*, rather than typical results for a single, pre-specified treatment.

Thus, there are two ingredients needed to apply the closure principle to combination tests:

(a) Testing an intersection hypothesis.

(b) Combining data from the two stages, Phase IIb and Phase III.

We tackle problem (b) first using a combination test as described in Section 2. We denote the $P$-value for testing $H_I$ in Phase IIb by $P_{1,I}$ and the $P$-value for testing $H_I$ in Phase III by $P_{2,I}$. Correspondingly, we define $Z_{1,I} = \Phi^{-1}(1 - P_{1,I})$ and $Z_{2,I} = \Phi^{-1}(1 - P_{2,I})$.

Using the inverse $\chi^{\mathbf{2}}$ method, we reject $H_I$ if

$$-\log(P_{1,I}\, P_{2,I}) \;>\; \frac{1}{2}\,\chi^2_4(\alpha). \tag{1}$$

Alternatively, using the weighted inverse normal method, we reject $H_I$ if

$$w_1 Z_{1,I} + w_2 Z_{2,I} \;>\; z(\alpha). \tag{2}$$

Here $w_1$ and $w_2$ are pre-specified weights with $w_1^2 + w_2^2 = 1$. If there is a common sample size, $m_1$ say, per treatment arm in Phase IIb and an anticipated sample size $m_2$ in Phase III, then an obvious choice is $w_i = \sqrt{(m_i/m)}$, $i = 1, 2$, where $m = m_1 + m_2$.

Turning to problem (a), consider first testing $H_I$ in Phase IIb. Suppose we calculate a $P$-value, $P_{1,i}$, for each $H_i$: $\theta_i \leq 0$. Using the Bonferroni inequality, the overall $P$-value for testing $H_I$ is $m \times \min_{i \in I} P_{1,i}$, where $m$ is the number of indices in $I$. Schmidli *et al.* (2006) propose using Simes' (1986) modification of the Bonferroni inequality which can be described as follows. Let $P_{1,(j)}$, $j = 1, \ldots, m$, denote the $m$ $P$-values in increasing order. Then the $P$-value for testing $H_I$ is

$$P_{1,I} \;=\; \min_{j=1,\ldots,m} (m\, P_{1,(j)}/j).$$

If treatment $i^*$ has the highest $\hat{\theta}_{1,i}$ and smallest $P$-value of all $k$ treatments, then $P_{1,(1)} = P_{1,i^*}$ in any set $I$ containing $i^*$. The term $m\, P_{1,(j)}/j$ with $j = 1$ becomes $m\, P_{1,i^*}$, the usual "Bonferroni

adjusted" version of $P_{1,i^*}$. Simes' method allows other low $P$-values to reduce the overall result: if a second treatment performs well, $P_{1,(2)}/2$ may be smaller than $P_{1,i^*}$, reducing $P_{1,I}$. We shall give an example of Simes' calculation shortly.

Testing $H_I$ in Phase III is simpler. In order to reject $H_{i^*}$: $\theta_{i^*} \leq 0$, we need to reject each $H_I$ with $i^* \in I$. Only treatment $i^*$ is studied in Phase IIb (along with the control), so a test of such an $H_I$ using Phase IIb data is based on $\hat{\theta}_{2,i^*}$ — and there is just one such test. Hence, all $H_I$ of interest have a common $P$-value in Phase III, namely $P_{2,I} = P_{2,i^*}$. Using either combination test, inspection of (1) or (2) shows that the key statistic from Phase IIb is:

$$\max_I P_{1,I} \quad \text{over sets } I \text{ containing } i^*.$$

Finally, we note that it is possible to construct unbiased point and interval estimates of treatment effect upon conclusion of the procedure; see Posch *et al.* (2005).

### 5.2 Alternative two-stage selection/testing designs

There have been several more classical proposals for addressing this problem. These include Thall, Simon and Ellenberg (1988), Schaid, Wieand and Therneau (1990) and Sampson and Sill (2005). Also, Stallard and Todd (2003) present a $K$-stage group sequential procedure where at stage 1 the treatment $i^*$ with the highest value of $\hat{\theta}_{1,i}$ is selected to be carried forward to stages 2 to $K$. For two stages, their procedure turns out to be equivalent to that of Thall, Simon and Ellenberg (1988). None of these tests are presented in terms of the closure principle, but they can be interpreted in that framework. We shall focus first on the approach of Thall, Simon and Ellenberg (TSE), which can be described as follows.

*Stage 1 (Phase IIb)*

Take $m_1$ observations per treatment and control.

Denote the estimated effect of treatment $i$ against control by $\hat{\theta}_{1,i}$ and let the maximum of these be $\hat{\theta}_{1,i^*}$.

If $\hat{\theta}_{1,i^*} < C_1$, stop and accept $H_0$: $\theta_1 = \ldots = \theta_\ell = 0$, abandoning the trial as futile.

If $\hat{\theta}_{1,i^*} \geq C_1$, select treatment $i^*$ and proceed to Phase III.

*Stage 2 (Phase III)*

Take $m_2$ observations on treatment $i^*$ and the control.

Combine data in $T_{i^*} = (m_1\,\hat{\theta}_{1,i^*} + m_2\,\hat{\theta}_{2,i^*})/(m_1 + m_2)$.

If $T_{i^*} < C_2$, accept $H_0$.

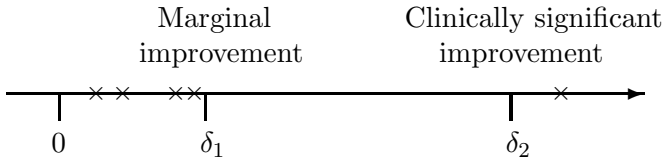
If $T_{i^*} \geq C_2$, reject $H_0$ and conclude $\theta_{i^*} > 0$.

Type I error and power requirements impose conditions on the values of $m_1$, $m_2$, $C_1$ and $C_2$. In defining the type I error rate, treatment $i^*$ is said to be "chosen" if it is selected at the end of stage 1 and $H_0$ is rejected in favor of $\theta_{i^*} > 0$ in the final analysis. TSE define the type I error rate as

$$Pr\{\text{Any experimental treatment is "chosen"}\}$$

under $H_0$: $\theta_1 = \ldots = \theta_\ell = 0$.

The power of the procedure is defined as $Pr_{\boldsymbol{\theta}}\{\text{An acceptable choice is made}\}$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\ell)$. Any treatment with $\theta_i \geq \delta_2$ is said to be "acceptable" for a specified $\delta_2 > 0$, termed the clinically significant improvement. Clearly power depends on the full vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\ell)$. Specifying also a marginal improvement $\delta_1$ where $0 < \delta_1 < \delta_2$, TSE set their power condition at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, defined by $\theta_1 = \ldots = \theta_{\ell-1} = \delta_1$ and $\theta_\ell = \delta_2$. They show this is the least favorable configuration (i.e., has the smallest power) for cases where at least one treatment is "acceptable" and no $\theta_i$ lies in the interval $(\delta_1, \delta_2)$ — see the figure below.




Numerical integration under $H_0$ and $\boldsymbol{\theta}^*$ enables parameters $m_1$, $m_2$, $C_1$ and $C_2$ to be found satisfying type I error and power conditions. Tests minimizing expected sample size averaged over these two cases can then be found by searching feasible parameter combinations. It is notable that in carrying out this minimization, TSE make an overall assessment of costs and benefits in dividing sample size between Phases IIb and III.

Although the TSE method was proposed for testing the simple null hypothesis $\theta_1 = \ldots = \theta_\ell = 0$, Jennison and Turnbull (2006e) show that it protects the familywise error rate at level $\alpha$ when testing the $\ell$ hypotheses $H_i$: $\theta_i \leq 0$, $i = 1, \ldots, \ell$. The TSE method can also be written as a closed testing procedure and this allows a less conservative approach when a treatment $i^*$ is selected with $\hat{\theta}_{1,i^*}$ below the highest $\hat{\theta}_{1,i}$.

Jennison and Turnbull (2006e) note that studies of efficiency gains from combining Phase IIb and Phase III have been somewhat limited up to now. Some comparisons have been made of the

total sample size in (a) separate Phase IIb and Phase III trials *versus* (b) a combined design with Phase IIb data used at the end of Phase III. Bretz *et al.* (2006) consider examples where the sample size per treatment in Phase IIb is equal to the sample size per treatment in Phase III. In this case, the combined study saves 30% of the total sample size for selecting one of $K = 2$ treatments and testing this against the control. However, such a large overall sample size in Phase IIb is unusual and in such a situation the case could be made for counting the Phase IIb trial as a supporting study (filling the role of one of the two confirmatory trials normally required). Todd and Stallard (2005) present an example where sample size per treatment is 25 in Phase IIb and 1400 in Phase III, so savings can be at most 2% of the total sample size!

### 5.3 A pedagogic example

To help clarify the ideas in this section, let us consider an example. Suppose four treatments for asthmatic patients are to be compared against a control treatment of current best practice. The endpoint is an asthma quality of life score (AQLS) at six weeks. We assume that responses are approximately normal with standard deviation $\sigma/\sqrt{2}$, where $\sigma = 5$ units.

Suppose that in Phase IIb, 100 observations are taken for each of the four treatments and a control group. Treatment $i^* = 4$ is selected for testing in Phase III, where a further 500 observations are taken on that treatment and the control treatment. The results are summarized in Table 1.

The question we examine is whether treatment 4 should be recommended at significance level $\alpha = 0.025$. According to the frequentist paradigm, the answer will depend on the design that was specified. We consider three possible types of design that could have led to the results in Table 1.

1. **Conventional.** Here we suppose separate Phase IIb and Phase III trials were conducted with the final decision to be made on the Phase III data alone.

2. **Bauer and Köhne.** Here, we assume the protocol specified that Phase IIb and Phase III results would be combined using the procedure of Bauer and Köhne (BK) with (a) inverse $\chi^2$ or (b) weighted inverse normal methods, the particular form of combination test being specified in advance. In either case, a closed testing procedure is to be applied using Simes' method to test intersection hypotheses.

3. **Thall, Simon and Ellenberg.** In this case, we assume the design described in Section 5.2 was used with $C_1 = 0$, so the trial would have been abandoned if the highest experimental mean in Phase IIb was lower than the control mean. The final critical value is set as $C_2 = 0.449$

Table 1: Observed $P$-values and $Z$-values in the pedagogic example. On the strength of its Phase IIb performance, treatment $i^* = 4$ was selected to go forward to Phase III.

***Phase IIb results***

| | Control | Trt 1 | Trt 2 | Trt 3 | Trt 4 |
|---|---|---|---|---|---|
| $n$ | 100 | 100 | 100 | 100 | 100 |
| $P$ (1-sided) | | 0.20 | 0.04 | 0.05 | 0.03 |
| $Z$ | | 0.84 | 1.75 | 1.64 | 1.88 |

***Phase III results***

| | Control | Trt 4 |
|---|---|---|
| $n$ | 500 | 500 |
| $P$ (1-sided) | | 0.04 |
| $Z$ | | 1.75 |

which maintains the overall type I error probability of 0.025 with allowance for the selection of treatment $i^*$ in Phase IIb. (*Note:* This is equivalent to a two-stage version of the Stallard and Todd (2003) procedure with a particular futility boundary.)

**Design 1.** The analysis under the Conventional protocol, is straightforward. Since the Phase III $P$-value of 0.04 is greater than 0.025, the null hypothesis is not rejected and the result of the trial is negative.

**Design 2.** In a Bauer and Köhne design, we must obtain (adjusted) $P$-values from each stage and combine them.

*Stage 1*

Using the closure principle and Simes' method to test each intersection hypothesis, we find $P_1 = \max_{I:4\in I}\{P_{1,I}\} = 0.075$.

(For sets $I$ containing $i^* = 4$, namely $\{1,2,3,4\}$, $\{1,2,4\}$, $\{1,3,4\}$, $\{2,3,4\}$, $\{1,4\}$, $\{2,4\}$,

$\{3,4\}$ and $\{4\}$, the maximum $P_{1,I}$ comes from $I=\{1,3,4\}$ and equals $3\,P_{1,3}/2 = 0.075$.)

*Stage 2*

$P_2 = 0.04$.

*(a) Combining $P_1$ and $P_2$ by the inverse $\chi^2$ method*

The combination statistic is $-2\log(0.075\times 0.04) = 11.6$. Since $P = Pr\{\chi^2_4 > 11.6\}$ $= 0.0204$, treatment 4 is found effective against control at overall level $\alpha = 0.025$.

*(b) Combining $P_1$ and $P_2$ by the inverse normal method*

The combination statistic is

$$\sqrt{100/600}\cdot\Phi^{-1}(1-0.075)+\sqrt{500/600}\cdot\Phi^{-1}(1-0.04) = 2.19.$$

Since $P = Pr\{Z > 2.19\} = 0.0144$, treatment 4 is found effective against control at overall level $\alpha = 0.025$.

So, both methods of combining $P_1$ and $P_2$ produce a positive outcome for the trial.

**Design 3.** In the TSE design, we have

*Stage 1*

$Z_{1,i^*} = 1.88, \quad \hat\theta_{1,i^*} = \sigma\cdot Z_{1,i^*}/\sqrt{n_1} = 0.940.$

*Stage 2*

$Z_2 = 1.75, \quad \hat\theta_{2,i^*} = \sigma\cdot Z_2/\sqrt{n_2} = 0.391.$

Combining these,

$$T_{i^*} \;=\; \frac{100\,\hat\theta_{1,i^*} + 500\,\hat\theta_{2,i^*}}{600} \;=\; 0.483.$$

Since $T_{i^*} > C_2 = 0.449$, treatment 4 is found to be effective against the control at an overall significance level less than 0.025. Thus, the trial has a positive outcome and a recommendation can be made in support of treatment 4.

Here, $T_{i^*}$ is a sufficient statistic for comparing treatment $i^* = 4$ against the control based on data from both stages and the critical value $C_2$ satisfies the requirement

$$Pr\{\hat\theta_{1,i^*} > C_1 = 0 \text{ and } \hat\theta_{2,i^*} > C_2 | H_0\} = 0.025.$$

The associated $Z$-statistic is

$$Z = \sqrt{\frac{100}{600}} \cdot Z_{1,i^*} + \sqrt{\frac{500}{600}} \cdot Z_2 = \frac{\sqrt{600}}{\sigma} \cdot T_{i^*} = 2.365$$

and this exceeds the equivalent critical value of $\sqrt{600/\sigma} \times 0.449 = 2.20$.

Now we have seen how the data would be analyzed in each case, we can compare properties of the three designs. Note that each design has at most two stages. In the first stage, Phase IIb, there are to be 100 subjects on each of the four competing treatments and in the control group, making 500 in all. The program may be ended as futile if no treatments look promising. Otherwise, the best treatment will be chosen for comparison with the control in the second stage, Phase III, where there will be 500 subjects on the selected treatment and 500 controls. At the end, a decision will be made either to accept the selected treatment or to reject it for insufficient improvement over the control. The overall probability of accepting an experimental treatment when all four treatments are equal to the control is to be no greater than $\alpha = 0.025$.

We consider the designs described above:

(1) Conventional

(2a) BK with inverse $\chi^2$ combination test

(2b) BK with inverse normal combination test

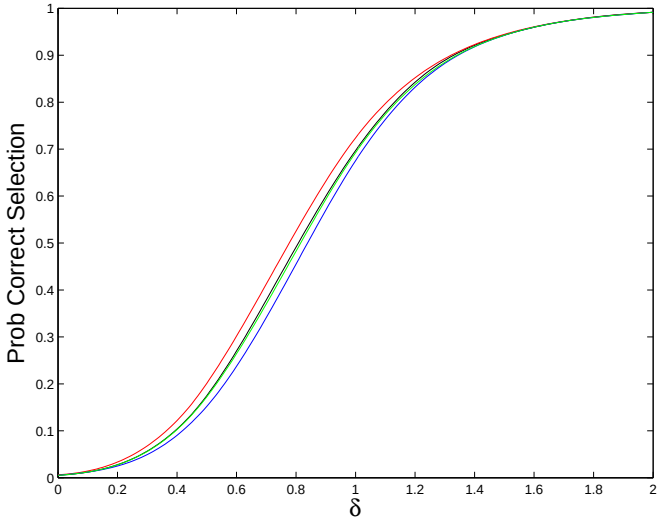
(3) TSE design.

Since the expected sample sizes are the same for all three strategies, the key comparison is between their power functions. Power depends on the full vector of treatment effects $(\theta_1, \ldots, \theta_4)$ in a complex way. We have made a simple start to exploring the power functions of the above designs by calculating power when three treatment effects are zero and the fourth is positive, so $\theta = (0, 0, 0, \delta)$ with $\delta > 0$. For (1), the Conventional design, power is given by

$$Pr\{\text{Treatment 4 is selected in Phase IIb}\} \times Pr\{H_{0,4}\text{: } \theta_4 \leq 0 \text{ is rejected after Phase III}\}.$$

The first factor is the probability the observed mean of treatment 4 exceeds that of treatments 1 to 3 and the control at the end of Phase IIb. The second factor is simply the power of a two-sample $Z$-test based on $n_2 = 500$ observations per sample. Both terms can be computed directly. Power for (2a) and (2b), the two BK designs, was obtained by simulation using 500,000 replications. For (3), the TSE design, power was computed using Equation (3) of Thall, Simon and Ellenberg (1988) or, equivalently, from the formulae given by Stallard and Todd (2003, Sec. 2) with $K = 2$. These power calculations were checked by simulations.

Figure 1: Power functions of four 2-stage selection/testing procedures. The top curve is for the TSE procedure; the middle curves, which are hardly distinguishable, are for the Conventional and BK (inverse normal) procedures; the bottom curve is for the BK (inverse $\chi^2$) procedure.




Power curves for the four designs are shown in Figure 1. It can be seen from these results that, in this example, the TSE procedure is uniformly more powerful than the other designs, but not by much. The BK procedures have added flexibility and this might be regarded as compensation for the slight reductions in power. Note, though, that modifying Phase III sample size in the light of Phase IIb data could raise inefficiency concerns similar to those described in Section 3. Combining the two phases can have other significant benefits if this removes the delay in passing from Phase IIb to Phase III. Of course, the Conventional and TSE designs can also be applied "seamlessly" and gain this advantage too.

It should be noted that in this comparison, the TSE design has a subtle, small advantage as it is constructed to have type I error rate exactly equal to 0.025. In contrast, because of the the futility boundary in Phase IIb, the Conventional design and the two BK designs are conservative, with type I error probabilities 0.0200, 0.0212 and 0.0206, respectively, when $\theta_1 = \ldots = \theta_4 = 0$. These three procedures can be adjusted to have type I error rate equal to 0.025: for the Conventional design, the critical value for $Z_2$ is reduced from 1.96 to 1.86, corresponding to a significance level of $P = 0.031$; in the BK (inverse $\chi^2$) procedure, we change the significance level for $P_1P_2$ from 0.025 to 0.030, and in the BK (inverse normal) test we decrease the critical value for the combination $Z$-statistic from 1.96 to 1.86. This adjustment for a futility boundary has been termed "buying

back alpha". When the conservative procedures are adjusted as just described, their power curves become closer to that of the TSE procedure, but remain in the same order.

# 6 Application 3: Target population selection in "enrichment" designs

Suppose we are interested in testing a particular experimental treatment versus a control. In stage 1, subjects are drawn from a general population, $\Omega_1$ say, and assigned at random to treatment and control groups. We define $\theta_1$ to be the treatment effect in $\Omega_1$. However, the population $\Omega_1$ is heterogeneous and we can think of $\theta_1$ as a weighted average of effects taken over the whole population. We also specify $\ell-1$ sub-populations of $\Omega_1$. These sub-populations, denoted $\Omega_2, \ldots, \Omega_\ell$, are not necessarily disjoint or nested, although they could be. The case of sub-populations based on genomic markers is considered by Freidlin and Simon (2005) and Temple (2005). We denote the treatment effects in these populations by $\theta_2, \ldots, \theta_\ell$, respectively. At the end of stage 1, based on the observed outcomes, one population, $\Omega_{i^*}$, is chosen and all subjects randomized between the treatment and control in stage 2 are taken from $\Omega_{i^*}$. The selected population could be $\Omega_1$ or any of the sub-populations. The choice of $i^*$ will reflect a trade-off. A highly targeted sub-population where the treatment is most effective could lead to a positive result with a smaller sample size; otherwise, a highly effective treatment in a small sub-population could be masked by dilution with responses from the remainder of the population where there is little or no effect. On the other hand, if a positive effect is present across the broad population, there are benefits for public health in demonstrating this and associated commercial advantages from more general labelling. It follows that $\Omega_{i^*}$ will not necessarily be the population with the highest observed treatment effect at the end of stage 1. The choice will be based both on the magnitudes of the observed treatment effects and on various medical and economic considerations. Thus, a "flexible" procedure is needed. As in the previous application, we must account for a selection bias if we are to combine stage 1 and stage 2 results to test the data generated hypothesis $H_{i^*}$: $\theta_{i^*} \leq 0$.

The procedure is analogous to that of Section 5. For each population $i$, the null hypothesis of no treatment effect, $H_i : \theta_i \leq 0$, is to be tested against the alternative $\theta_i > 0$. We shall follow a closed testing procedure in order to control familywise type I error. This involves testing intersection hypotheses of the form $H_I = \cap_{i \in I} H_i$, implying that $\theta_i \leq 0$ for all $i \in I$. For a given $H_I$, we perform a test in each stage of the trial and then combine the test statistics ($P$-values) across stages. At the end of stage 2 we can test $H_{i^*}$ and, additionally, any $H_j$ such that $\Omega_j$ is a subset of $\Omega_{i^*}$. In

particular if $i^* = 1$ and the full general population is selected, then it will be possible to test all of $H_1, \ldots, H_\ell$. It is easiest to describe the procedure by means of a simple example, from which the general method should be clear.

**Example.**

In our illustrative example, sub-populations are:

1. The entire population
2. Men only
3. Men over 50
4. Men who are smokers

Within a stage, each intersection hypothesis will be tested by combining $P$-values from individual hypotheses using Simes' (1986) method. We then combine $P$-values from the two stages by a weighted inverse normal rule, giving

$$Z(p_1, p_2) = w_1 \, \Phi^{-1}(1 - p_1) + w_2 \, \Phi^{-1}(1 - p_2),$$

where $w_1 = w_2 = \sqrt{0.5}$.

The elementary null hypotheses are $H_i$: $\theta_i \leq 0$ for $i = 1, \ldots, 4$. In stage 1, the individual test of $H_i$ is performed using the estimate $\hat{\theta}_{1,i}$ from subjects within the relevant sub-population, giving $P$-value $P_{1,i}$.

In stage 2, it may only be possible to test some of the $H_i$ using stage 2 data, for example, if recruitment is restricted to "Men only", we can test $H_2$, $H_3$ and $H_4$ but not $H_1$, since $\theta_1$ is a weighted average of effects on both men and women. Thus, we obtain stage 2 $P$-values $P_{2,i}$ for some hypotheses $H_i$ but not for others.

*Using the closure principle.* In order to reject $H_{i^*}$: $\theta_{i^*} \leq 0$ in the overall procedure with familywise error probability $\alpha$, we need to reject each intersection hypothesis $H_I$ for which $i^* \in I$ at level $\alpha$, based on combined stage 1 and stage 2 data. The need for adjustment when testing multiple hypotheses is clear. Even when some of these hypotheses are taken out of consideration by focusing on a sub-population $\Omega_{i^*}$ in stage 2, the choice of this sub-population is data driven and subject to selection bias. The familywise test will automatically take care of these effects of multiplicity and sub-population selection.

*Testing an intersection hypothesis $H_I$: $\theta_i \leq 0$ for all $i \in I$.* As in Section 5, we will need (a) to test an intersection hypothesis and (b) to combine data from the two stages. We first tackle

problem (b) using the weighted inverse normal combination test described above. Letting $P_{1,I}$ and $P_{2,I}$ denote $P$-values for testing $H_I$ from stages 1 and 2, we calculate

$$Z(P_{1,I}, P_{2,I}) = w_1\, \Phi^{-1}(1 - P_{1,I}) + w_2\, \Phi^{-1}(1 - P_{2,I}),$$

using the specified weights $w_1 = w_2 = \sqrt{0.5}$. Then, we reject $H_I$ if

$$Z(P_{1,I}, P_{2,I}) \;>\; \Phi^{-1}(1 - \alpha).$$

Now consider the problem (a) of testing the intersection hypothesis $H_I$ within each stage. In stage 1, we can calculate a $P$-value, $P_{1,i}$, for each $H_i$: $\theta_i \leq 0$. Using the Bonferroni inequality, the overall $P$-value for testing $H_I$ would be $m$ times the minimum $P_{1,i}$ over $i \in I$, where $m$ is the number of indices in $I$. However, as in Section 5, we follow the recommendation of Schmidli *et al.* (2006) and use Simes' (1986) modification of the Bonferroni inequality. Recall that in Simes' method, if $I$ has $m$ elements, the $P$-value for testing $H_I$ is

$$P_{1,I} \;=\; \min_{j=1,\ldots,m} (m\, P_{1,(j)}/j),$$

where $P_{1,(j)}$, $j = 1, \ldots, m$, denote the $m$ $P$-values in increasing order.

For testing an intersection hypothesis in stage 2, we have $P$-values, $P_{2,i}$, for *some* of the $H_i$: $\theta_i \leq 0$, depending on the section of the population from which recruitment took place in this stage. Let $I'$ denote the set of indices $i \in I$ for which we have a $P$-value $P_{2,i}$ and suppose there are $m'$ such indices.

We can apply Simes' method on the reduced set $I'$, as long as it is non-empty, yielding the $P$-value for testing $H_I$

$$P_{2,I} \;=\; \min_{j=1,\ldots,m'} (m'\, P_{2,(j)}/j),$$

where $P_{2,(j)}$, $j = 1, \ldots, m'$, are the $m'$ available $P$-values arranged in increasing order.

Suppose, in our example, that recruitment in stage 2 is restricted to "Men only" and we observe the results shown in Table 2. Had the study continued to recruit from the full population in stage 2, a $P$-value could have been calculated for each sub-population and all combination tests would be feasible. Because recruitment is restricted, elementary tests are only possible for sub-populations which are contained completely in the new recruitment pool. Since the sub-populations "Men over 50" and "Men who are smokers" are still sampled fully, we can test $H_3$ and $H_4$ as well as $H_2$. We cannot test $H_1$ since $\theta_1$ is a weighted average of effects on both men and women and stage 2 provides no information about women. Consequently, we can test $H_2$, $H_3$ and $H_4$ at the

Table 2: Observed $P$-values in the example of sub-population selection. After stage 1, it is decided to restrict recruitment in stage 2 to "Men only".

***Stage 1 results***

| | |
|---|---|
| Full population: | $P_{1,1} = 0.20$ |
| All men: | $P_{1,2} = 0.10$ |
| Men over 50 years: | $P_{1,3} = 0.03$ |
| Men who smoke: | $P_{1,4} = 0.03$ |

***Stage 2 results***

| | |
|---|---|
| All men: | $P_{2,2} = 0.11$ |
| Men over 50 years: | $P_{2,3} = 0.08$ |
| Men who smoke: | $P_{2,4} = 0.03$ |

global level. As an example, in testing $H_2$, the relevant sets $I$ all contain $i = 2$, so there is at least one element in the reduced set $I'$.

*Adjusted combined $P$-values.* The $P$-values for testing hypotheses after combining data from the two stages are displayed in Table 3. In order to reject $H_i$ at global level $\alpha = 0.025$, each intersection hypothesis $H_I$ with $i \in I$ must be rejected at this level. The adjusted $P$-value, $\tilde{P}_i$, for testing $H_i$ with protection of familywise error rate is the maximum $P$-value (combined over stages 1 and 2) for all $H_I$ with $i \in I$. Thus, for testing $H_2$ concerning the sub-population of "All men", we have

$$\tilde{P}_2 \;=\; \max\{P_2, P_{12}, P_{23}, P_{24}, P_{123}, P_{124}, P_{234}, P_{1234}\} \;= 0.072;$$

in testing $H_3$ for "Men over 50 years", we obtain

$$\tilde{P}_3 \;=\; \max\{P_3, \ldots, P_{1234}\} \;=\; 0.035;$$

and the test of $H_4$ for "Men who smoke" has

$$\tilde{P}_4 \;=\; \max\{P_4, \ldots, P_{1234}\} \;=\; 0.020.$$

We can therefore reject $H_4$, but not $H_2$ or $H_3$, at global significance level $\alpha = 0.025$ and quote an adjusted $P$-value of $\tilde{P}_4 = 0.020$.

Table 3: Tests of intersection hypotheses

| | $P$-values | | Combined | |
|---|---|---|---|---|
| | Stage 1 | Stage 2 | Z | P |
| $H_1$ | .20 | — | — | — |
| $H_2$ | .10 | .11 | 1.77 | .038 |
| $H_3$ | .03 | .08 | 2.32 | .010 |
| $H_4$ | .03 | .03 | 2.66 | .004 |
| $H_{12}$ | .20 | .11* | 1.46 | .072 |
| $H_{13}$ | .06 | .08* | 2.09 | .018 |
| $H_{14}$ | .06 | .03* | 2.43 | .008 |
| $H_{23}$ | .06 | .11 | 1.97 | .025 |
| $H_{24}$ | .06 | .06 | 2.20 | .014 |
| $H_{34}$ | .03 | .06 | 2.43 | .008 |
| $H_{123}$ | .09 | .11* | 1.82 | .035 |
| $H_{124}$ | .09 | .06* | 2.05 | .020 |
| $H_{134}$ | .045 | .06* | 2.30 | .011 |
| $H_{234}$ | .045 | .09 | 2.15 | .016 |
| $H_{1234}$ | .06 | .09* | 2.05 | .020 |

* Value for $P_{2,\{1\}\cup I}$ in stage 2 equals $P_{2,I}$ for $I \subseteq \{2,3,4\}$.

Note that the weights $w_1$ and $w_2$ need to be specified *a priori*. The choice $w_1 = w_2$ is appropriate if it is intended that both stages should have equal sample size. An advantage of this adaptive method is that the stage 2 sample size can be modified in light of the stage 1 data. However, the final analyses are not then based on sufficient statistics and so some of the criticisms raised in Section 3 obtain.

# 7 Application 4: Switching endpoints

We consider a trial with a specified primary response, endpoint 1, and null hypothesis $H_1$: $\theta_1 \leq 0$ to be tested against $\theta_1 > 0$, where $\theta_1$ is the treatment effect on endpoint 1. However, investigators recognise they may wish to switch to an alternative outcome, endpoint 2, in stage 2: in this case they will want to test the null hypothesis $H_2$: $\theta_2 \leq 0$, where $\theta_2$ is the treatment effect on endpoint 2. To allow for this eventuality, the study is designed in two stages and a combination rule is stipulated for aggregating $P$-values from the two stages. It is necessary to define a procedure following the closure principle in order to protect the familywise error rate while considering both $H_1$ and $H_2$.

Let $P_{i,j}$ denote the $P$-value for testing $H_j$ from stage $i$ data, for $i$ and $j$ equal to 1 and 2. If

data on an endpoint are only recorded when it is the "designated" endpoint for a stage, some of the $P_{i,j}$ will not be available. We define the following tests of individual null hypotheses for use in the overall procedure.

*To test $H_1$:*

Apply the specified combination test to $P_{1,1}$ and $P_{2,1}$.

*To test $H_{12} = H_1 \cap H_2$:*

If endpoint 1 is retained in stage 2, apply the combination test to $P_{1,1}$ and $P_{2,1}$.

If a switch is made to endpoint 2, apply the combination test to $P_{1,1}$ and $P_{2,2}$.

*To test $H_2$:*

Use $P_{2,2}$ only, rejecting $H_2$ at level $\alpha$ if $P_{2,2} \leq \alpha$.

If the original endpoint is retained throughout the study, rejection of $H_1$ overall requires both individual hypotheses $H_1$ and $H_{12}$ to be rejected. However, in this case the test of $H_{12}$ is the same as that of $H_1$ so the result is just as it would have been had the option of switching to endpoint 2 not been considered. We have deliberately chosen to define the test of $H_{12}$ this way, using $P_{1,1}$ from stage 1 even if data on endpoint 2 are available in this stage, in order for the overall test of $H_1$ to have this property.

If the switch is made to endpoint 2, overall rejection of $H_2$ needs both $H_2$ and $H_{12}$ to be rejected. The individual test of $H_2$ is the test that would arise if the second stage were a self-contained study conducted to test $H_2$. The closure principle imposes the additional requirement that the combination test applied to $P_{1,1}$ and $P_{2,2}$ should lead to rejection of $H_{12}$. One might question whether it might not be preferable to start a new trial after stage 1, defining endpoint 2 as the primary response variable, so that only the test of $H_2$ need be considered. However, staying with the original trial could save time by avoiding the waiting period while a new study is organized and approved. Moreover, credibility might suffer if a sponsor were to stop a number of trials early, discarding unfavorable results to start new trials with slightly different endpoints.

The above procedure can be extended to allow a switch to a different new endpoint. Suppose, for example, other events in stage 1 could have led to use of endpoint 3 with associated null hypothesis $H_3$. A closed testing procedure should include $H_3$ in the set of null hypotheses. If endpoint 1 is retained, rejection of $H_1$ overall requires all the individual hypotheses $H_1$, $H_{12}$, $H_{13}$ and $H_{123}$ to

be rejected. Applying the combination test to $P_{1,1}$ and $P_{2,1}$ gives a valid test for each of these hypotheses and, with this definition, there is no change to the overall requirement for rejecting $H_1$. Under a switch to endpoint 2 after stage 1, global rejection of $H_2$ requires each of $H_2$, $H_{12}$, $H_{23}$ and $H_{123}$ to be rejected. In this case, we retain the previous definitions for tests of $H_2$ and $H_{12}$, we use $P_{2,2}$ to test $H_{23}$, and we define the test of $H_{123}$ to be the combination test applied to $P_{1,1}$ and $P_{2,2}$. Since the tests of $H_{23}$ and $H_{123}$ replicate those of $H_2$ and $H_{12}$, the addition of endpoint 3 does not change the requirements that must be met in order to reject $H_2$. If the switch is made to endpoint 3, we test both $H_{13}$ and $H_{123}$ by the combination test applied to $P_{1,1}$ and $P_{2,3}$, and we test both $H_3$ and $H_{23}$ through the single $P$-value $P_{2,3}$.

It is not difficult to check that this approach can accommodate other potential endpoints too, treating these in the same way as the above extension to endpoint 3. The criteria for overall rejection of endpoints 1, 2 or 3 will not be affected by these additional endpoints. It follows that the precise definition of a new endpoint can actually be left until the end of stage 1 and information emerging from this first stage can be used in making this definition.

This flexibility to consider a variety of possible new endpoints contrasts with the requirements for changing to a sub-population seen in Section 6. There, it was crucial to define the set of potential sub-populations at the outset and the degree of "adjustment" to $P$-values for having this option increased with the number of potential sub-populations. The reason for this qualitative difference is that in the "switching" procedure there is a precedence of hypotheses, with $H_1$ taking priority: in particular, the $P$-value $P_{1,1}$ is defined as the data summary from stage 1 to be used in all tests of intersection hypotheses involving $H_1$.

If there is a clear restriction to just two endpoints and data are available on both in the first stage, it may be reasonable to use an alternative definition for the test of $H_2$, making this a combination test of $P_{1,2}$ and $P_{2,2}$. This must, of course, be specified in the study protocol as the method that will be used. Consider, for example, a study where the primary endpoint is the change in a clinical measurement over a 6 month period following the start of treatment. Investigators may declare a second endpoint to be the change in this measurement after 4 months. It is necessary to keep the test of $H_{12}$ as we have defined above in order to (a) leave the overall test of $H_1$ the same and (b) maintain validity in the test $H_{12}$. One cannot, for example, use $P_{1,2}$ in place of $P_{1,1}$ in the test of $H_{12}$ only when a switch is made to endpoint 2, as the decision to make this switch will have been motivated by stage 1 data on endpoint 2. If three or more endpoints are considered, it is technically possible to define an overall procedure that makes use of stage 1 data on an all

endpoints, but we would not necessarily recommend this: with endpoints 2 and 3 available, a test of $H_{23}$ using stage 1 data will have to combine $P_{1,2}$ and $P_{1,3}$ by, for example, Simes' test and the procedure becomes more complex as further endpoints are added.

# 8 Application 5: Selecting between superiority and non-inferiority hypotheses

Let $\theta$ denote the treatment effect when testing a single experimental treatment versus an active control with respect to a single primary endpoint of interest. We consider the situation where, although it would be preferable to demonstrate superiority of the experimental treatment, it is also of value to demonstrate non-inferiority. Let $0 = \delta_1 < \delta_2 < \ldots < \delta_\ell$ be a sequence of pre-specified non-inferiority margins, as discussed by Hung and Wang (2004). We are interested in the collection of hypothesis tests $\mathcal{T}_1, \ldots, \mathcal{T}_\ell$, where in $\mathcal{T}_i$, we test:

$$H_i : \theta \leq -\delta_i \quad \text{versus} \quad \theta > -\delta_i.$$

Rejection of $H_1$ implies a positive result for superiority, whereas rejection of $H_i$ for a value of $i$ between 2 and $\ell$ implies non-inferiority with margin $\delta_i$. Often $\ell = 2$ and there is only one non-inferiority margin of interest.

Consider testing $H_i$ using the closed testing procedure that controls the familywise error $\alpha$. In order to reject $H_i$ we must reject intersection hypotheses $H_I$ for every subset $I$ with $i \in I$. Because of the nesting of the hypotheses implied by $0 = \delta_1 < \delta_2 < \ldots < \delta_\ell$, we have $H_I = H_{\max(I)}$ where $\max(I) = \max\{j : j \in I\}$. Also, because of the nested hypotheses, we have $P_1 \geq P_2 \geq \ldots \geq P_\ell$. Thus $P_I = P_{\max(I)}$ and $\max\{P_I : i \in I\} = P_i$, which is achieved when $I = \{i\}$. This means that for each $i = 1, \ldots, \ell$, we may reject $H_i$ if $P_i \leq \alpha$ and no adjustment is needed for multiplicity. Nor does it matter in which order the hypotheses are tested. Of course, this is a well-known result — see, for example, Morikawa and Yoshida (1995) and Dunnett and Gent (1996).

Consider now a two-stage or multi-stage adaptive design. The $P$-values for each of the elementary hypothesis can be combined across stages using the methods described in Section 2. Each hypothesis can be tested separately using combination $P$-values and there is no need to adjust for the multiplicity of hypotheses. This problem has been considered by a number of authors including Wang *et al.* (2001), Brannath *et al.* (2003), Shih *et al.* (2004) and Koyama *et al.* (2005). The possible advantages of an adaptive multistage design are described by Wang *et al.* (2001). They consider the case $\ell = 2$ where there is only one non-inferiority margin. They point out that

often the non-inferiority margin is smaller than the effect size at which power for the superiority test is specified, and so the target sample size for demonstrating non-inferiority with power when $\theta = 0$ can be much larger than that needed to demonstrate superiority. As data accrue, it can become apparent which goal is the most relevant and the sample size may be adjusted accordingly. Wang *et al.* (2001) give an example where there are impressive savings in expected sample size. However, the procedures are not based on sufficient statistics and so some of the criticisms raised in Section 3 obtain. It should be possible to design a conventional group sequential test that is planned for the larger non-inferiority trial, but with an aggressive boundary so that the trial can be terminated early if the results justify a superiority declaration. This would follow the spirit of the approach recommended by Jennison and Turnbull (2006d) and bear similarities to the designs proposed by Koyama *et al.* (2005).

# 9 Discussion

There has been lively debate about the potential advantages and disadvantages of adaptive designs. Methodological criticisms have concentrated on issues of inefficiency and failure to satisfy the sufficiency principle — Tsiatis and Mehta (2003), Jennison and Turnbull (2003, 2006a, b), Fleming (2006), Burman and Sonesson (2006). Some of the practical problems of implementation are considered in Gallo *et al.* (2006) and the accompanying discussion, Gallo (2006), Gould (2006) and Fleming (2006).

In practice, adaptive designs have been utilized with some success and continue to be applied in new trials. If the logistics can be put in place to implement "seamless" designs, the elimination of delays between different stages of a trial will be a major advance. Adaptive methods promise the additional flexibility to refine the target population or primary endpoint during the course of a trial. It is clear that the current interest in adaptive designs will continue and there is a keen audience for further developments in the methodology and reports of its application.

## REFERENCES


Bauer P. and Kieser M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833–1848.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041. Correction *Biometrics* **52**, (1996), 380.

Bechhofer, R.E., Santner, T.J. and Goldsman, D.M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. Wiley, New York.

Becker, B.J. (1994). Combining significance levels. In *The Handbook of Research Synthesis*, Eds. Cooper, H. and Hedges, L.V. Russell Sage Foundation, New York, Chap. 15, pages 215–230.

Birkett, M.A. and Day, S.J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455–2463.

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *J. American Statistical Association* **97**, 236–244.

Brannath, W., Bauer, P., Maurer, W. and Posch, M. (2003). Sequential tests for noninferiority and superiority. *Biometrics* **59**, 106–114.

Bretz, F., Schmidli, H., König, F., Racine, A. and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts (with discussion). *Biometrical Journal* **48**, 623–634.

Burman, C-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* **62**, 664–669.

Chow, S-C. and Chang, M. (2007). *Adaptive Design Methods in Clinical Trials.* Chapman & Hall/CRC, Boca Raton, Florida.

Cui, L., Hung, H.M.J. and Wang, S-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857.

Denne, J.S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645–2660.

Denne, J.S. and Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* **18**, 1575–1585.

Denne, J.S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika* **87**, 125–134.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. American Statistical Association* **50**, 1096–1121.

Dunnett, C.W. and Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine* **15**, 1729–1738.

EMEA (2006). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. (Draft 23 March 2006). EMEA (European Medicines Agency) Committee for Medicinal Products for Human Use (CHMP). http://www.emea.eu.int/pdfs/human/ewp/245902en.pdf

Fisher, L.D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–1562.

Fisher, R.A. (1932). *Statistical Methods for Research Workers, 4th Ed.* Oliver and Boyd, London.

Fleming, T.R. (2006). Standard versus adaptive monitoring procedures: a commentary. *Statistics in Medicine* **25**, 3305–3312.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11**, 7872–7878.

Friede, T. and Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine* **20**, 3861–3873.

Friede, T. and Kieser, M. (2002). On the inappropriateness of an EM based procedure for blinded sample size re-estimation. *Statistics in Medicine* **21**, 165–176.

Gallo, P. (2006). Operational challenges in adaptive design implementation. *Pharmaceutical Statistics* **5**, 119–124.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M. and Pinheiro, J. (2006). Adaptive designs in clinical drug development — an executive summary of the PhRMA Working Group (with discussion). *J. Biopharmaceutical Statistics* **16**, 275–312.

Gould, A.L. (2006). How practical are adaptive designs likely to be for confirmatory trials? *Biometrical Journal* **48**, 644–649.

Gould, A.L. and Shih, W.J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics: Theory and Methods* **21**, 2833–2853.

Goutis, C., Casella, G., and Wells, M.T. (1996). Assessing evidence in multiple hypotheses. *J. American Statistical Association* **91**, 1268–1277.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

Hung, H.M.J., O'Neill, R.T., Wang S.J. and Lawrence J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* **48**, 565–573.

Hung, H.M.J. and Wang, S.J. (2004). Multiple testing of noninferiority hypotheses in active controlled trials. *J. Biopharmaceutical Statistics* **14**, 327–335.

Hung, H.M.J., Wang, S.J. and O'Neill R.T. (2006). Methodological issues with adaptation of clinical trial design. *Pharmaceutical Statistics* **5**, 99–107.

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

Jennison, C. and Turnbull, B.W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **23**, 971–993.

Jennison, C. and Turnbull, B.W. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *J. Biopharmaceutical Statistics* **15**, 537–558.

Jennison, C. and Turnbull, B.W. (2006a). Adaptive and nonadaptive group sequential tests. *Biometrika* **93**, 1–21.

Jennison, C. and Turnbull, B.W. (2006b). Discussion of "Are flexible designs sound?" *Biometrics* **62**, 670–673

Jennison, C. and Turnbull, B.W. (2006c). Discussion of "Executive summary of the PhRMA Working Group on adaptive designs in clinical drug development." *J. Biopharmaceutical Statistics* **16**, 293–298.

Jennison, C. and Turnbull, B.W. (2006d). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* **35**, 917–932.

Jennison, C. and Turnbull, B.W. (2006e). Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. *Biometrical Journal* **48**, 650–655.

Kieser, M. and Friede, T. (2000). Recalculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901–911.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal* **48**, 574–585.

Koyama, T., Sampson, A.R. and Gleser, L.J. (2005). A framework for two-stage adaptive procedures to simultaneously test non-inferiority and superiority. *Statistics in Medicine* **24**, 2439–2456.

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

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

Liu, Q., Proschan, M.A. and Pledger, G.W. (2002). A unified theory of two-stage adaptive designs. *J. American Statistical Association* **97**, 1034–1041.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Mehta, C. and Tsiatis, A.A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* **35**, 1095–1112.

Mosteller, F. and Bush, R.R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology, Vol.1* Ed. G. Lindzey, Addison-Wesley, Cambridge, MA, pages 289–334.

Morikawa, T., Yoshida, M. (1995). A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *J. Biopharmaceutical Statistics* **5**, 297–306.

Müller, H-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **57**, 886–891.

Posch, M. and Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics* **56**, 1170–1176.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C. and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* **24**, 3697–3714.

Proschan, M.A. (2003). The geometry of two-stage tests. *Statistica Sinica* **13**, 163–177.

Proschan, M.A. and Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

Sampson, A.R. and Sill, M.W. (2005). Drop-the-losers design: Normal case (with discussion). *Biometrical Journal* **47**, 257–281.

Sarkar, S. K. (1998). Some probability inequalities for ordered $\text{MTP}_2$ random variables: A proof of the Simes conjecture. *The Annals of Statistics* **26**, 494–504.

Sarkar, S. K. and Chang, C-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. American Statistical Association* **92**, 1601–1608.

Schaid, D.J., Wieand, S. and Therneau, T.M. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* **77**, 507–513.

Schmidli, H., Bretz, F., Racine, A. and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biometrical Journal* **48**, 635–643.

Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures.* Lecture Notes in Statistics, 79, Springer-Verlag: New York.

Shih, W.J., Quan, H. and Li, G. (2004). Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* **23**, 2781–2798.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22**, 689–703.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.

Temple, R.J. (2005). Enrichment designs: Efficiency in development of cancer treatments. *J. Clinical Oncology* **23**, 4838–4839.

Thall, P.F., Simon, R. and Ellenberg, S.S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.

Timmesfeld, N., Schäfer, H. and Müller, H-H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine* **26**, 2449–2464.

Todd, S. and Stallard, N. (2005). A New Clinical Trial Design Combining Phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal* **39**, 109–118.

Tsiatis, A.A. (2006). Information-based monitoring of clinical trials. *Statistics in Medicine* **25**, 3236–3244.

Tsiatis, A.A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–78.

Wang, S-J, Hung, H.M.J., Tsong, Y. and Cui, L. (2001). Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* **20**, 1903–1912.

Westfall, P.H. and Young, S.S, (1993). *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Wittes J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.

Zucker, D.M., Wittes, J.T., Schabenberger, O. and Brittain, E. (1999) Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine* **18**, 3493–3509.