# META-ANALYSES AND ADAPTIVE GROUP SEQUENTIAL DESIGNS IN THE CLINICAL DEVELOPMENT PROCESS

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:* turnbull@orie.cornell.edu

October 30, 2005

# SUMMARY

The clinical development process can be viewed as a succession of trials, possibly overlapping in calendar time. The design of each trial may be influenced by results from previous studies and other currently proceeding trials as well as by external information. Results from all these trials must be considered together in order to assess the efficacy and safety of the proposed new treatment. Meta-analysis techniques provide a formal way of combining the information. We examine how such methods can be used in combining results: (1) from a collection of separate studies, (2) from a sequence of studies in an organized development program, (3) from stages within a single study using a (possibly adaptive) group sequential design. We present two examples. The first concerns the combining of results from a Phase IIb trial using several dose levels or treatment arms with those of the Phase III trial comparing the treatment selected in Phase IIb against a control. This enables a "seamless transition" from Phase IIb to Phase III. The second example examines the use of combination tests to analyze data from an adaptive group sequential trial.

*Key words:* Adaptive designs; Clinical trials; Early data review; Flexible design; Group sequential tests; Interim analysis; Meta-analysis; Overviews; Sample size re-estimation; Systematic reviews; Two-stage procedure; Variance spending.

# 1 Introduction

The development process for a new drug often takes as long as ten to fifteen years and can cost over US$750 million. Hence any acceleration of the process, however slight, to identify beneficial or problematic drugs early can lead to large savings in economic cost and/or benefits for human health. The importance of this fact in the area of HIV prevention trials, for example, has been emphasized recently by Padian (2004). The process consists of many stages from preclinical laboratory and animal studies through to the various phases of clinical trials leading to review by the FDA, or other regulatory body, of the New Drug Application. These stages are carried out successively but typically may overlap in calendar time — see, for example, the CDER Handbook (`http://www.fda.gov/cder/handbook/develop.htm`). At each stage, a decision must be made whether to abandon or continue the process. This decision will be based on all available data from preceding and current stages as well as any pertinent external information. If the decision is to continue, then the next study is designed (patient population, drug dose, etc.) with the help of results from the previous stages. This information is combined both in a formal and informal way. If the studies are separate and independent, then various *meta-analysis* methods have been proposed to provide a quantitative way to combine the results — see, for example, Hedges and Olkin (1985), Sutton et al. (2000), Egger et al. (2001) or Whitehead (2002). As we shall see, certain meta-analysis methods based on P-values remain valid even when the design of one study may depend on results from other studies in the series.

Near the end of the process, the confirmatory Phase III clinical trial is typically a large scale multi-center trial involving a large number of patients. The trial lasts several years and it is now becoming more customary to have interim monitoring performed by a data safety and monitoring board (DSMB) — see, for example, the guidance in ICH E9 (FDA, 1998). A formal group sequential design is often included in the trial protocol or in the DSMB charter. Over time, information, be it blinded or unblinded, internal or external to the trial, accumulates and this may naturally indicate worthwhile changes to be made to the trial design. A common change is sample size re-estimation based on a better knowledge of a key nuisance parameter such as a response variance or a baseline event rate. Such modifications to the design are often approved. In fact, this possibility and the consequent re-estimation rule can be pre-specified in the design protocol. However, other design modifications have been more problematic. Examples include changes in dosing levels or spacing, changes in patient population, changes from a superiority to a non-inferiority objective, and, in particular, changes in the designated target effect size to be detected with the pre-specified power (or equivalently a change in this stated power). As an example, Cui, Hung and Wang (1999) described a trial for prevention of myocardial infarction where unexpected interim results revealed a lack of power at a relevant alternative. Partly because of concerns about statistical validity, the sponsor withdrew the request for a protocol modification to increase the sample size and the trial eventually ended with a negative conclusion. Recently there has been a lot of activity in the topic of so-called *adaptive* group sequential designs. Basic papers on this subject include Bauer and Köhne (1994), Proschan and Hunsberger (1995), Fisher (1998), Cui et al. (1999), Lehmacher and Wassmer (1999), Denne (2001) and Müller and Schäfer (2001). At the time of writing, we have found over 100 papers in the literature on this subject. These proposals allow design modifications at interim looks, planned or unplanned, while preserving the overall Type I error rate $\alpha$. However this flexibility can come at the cost of statistical efficiency — see, for example, Jennison and Turnbull (2003), Tsiatis and Mehta (2003) and Posch, Bauer and Brannath (2003).

In this paper we examine how essentially the same "meta-analysis" methods can be used in combining results: (1) from a collection of separate studies, (2) from a sequence of studies in an organized development program, (3) from stages within a single study using a (possibly adaptive) group sequential design. In the next section we review methods for testing a null hypothesis based on several studies — so-called "combination tests". The studies may not be independent. In Section 3, we examine the properties of one of these methods, the weighted inverse normal combination test, in more detail. In Section 4, we use graphical methods to compare various combination tests. Sections 5 and 6 present two examples. The first concerns the combining of results from a Phase IIb trial using several dose levels or treatment arms with those of the Phase III trial which compares the dose or treatment selected after Phase IIb with a control. This enables a "seamless transition" from Phase IIb to Phase III. The second example examines the use of combination tests to analyze data from an adaptive group sequential trial.

## 2 Testing a null hypothesis based on several studies

We start by supposing there is a null hypothesis of no treatment effect that has been studied in $K(\geq 2)$ separate and independent experiments. Suppose that, for $1 \leq k \leq K$, the treatment effect in the $k$th experiment can be quantified by a real parameter $\theta_k$ such that the null hypothesis of no treatment effect is represented by $H_{0k}$: $\theta_k = 0$. We consider a one-sided alternative hypothesis $H_{Ak}$: $\theta_k > 0$. When combining studies, directions of deviations from each $H_{0k}$ are important and hence use of a two-sided alternative is not very useful. The overall hypothesis of no treatment effect is the intersection

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

which implies $\theta_k = 0$ for each $k = 1, \ldots, K$. The alternative hypothesis $H_A$ is that at least **one** $\theta_k > 0$. We are seeking a P-value for testing the overall null hypothesis, $H_0$, versus $H_A$.

Goutis, Casella and Wells (1996) give a very general discussion on the use of P-values as a means to assess evidence from multiple studies. In our situation, let $P_k$ denote the P-value from testing $H_{0k}$ in the $k$th study ($1 \leq k \leq K$). We shall assume $P_k$ has the uniform distribution $U(0,1)$ under $H_{0k}$. (This is generally the case, but may only be approximately true if the responses are discrete or if $H_{0k}$ is composite — see Robins, van der Vaart and Ventura (2000).)

Becker (1994) has surveyed methodology for combining significance levels and, in her Table 15.1, she lists some eighteen different proposals. Here we shall consider just a few that are commonly used in the social science literature and are natural to use in the clinical trials context. The first such method is an inverse chi-square or “sum of the logs” method proposed by R. A. Fisher (1932). Here an $\alpha$-level tests rejects $H_0$ if

$$-2\log(P_1 P_2 \ldots P_K) \;>\; \chi^2_{2K}(\alpha), \tag{1}$$

where $\chi^2_{2K}(\alpha)$ denotes the upper $\alpha$ percentage point of a chi-squared distribution with $2K$ degrees of freedom. This follows because under $H_0$ the $\{P_k\}$ are independent $U(0,1)$ and so $-\log P_k \sim$ Exp(1) $\sim \frac{1}{2}\chi^2_2$. An overall P-value is defined by finding that value of $\alpha$ for which equality is obtained in (1).

The most commonly used method in the social sciences is the inverse normal method (Stouffer et al., 1949). Define $Z_k = \Phi^{-1}(1 - P_k)$ for $k = 1, \ldots, K$. Then under $H_0$, each $Z_k \sim N(0,1)$ and an $\alpha$-level test rejects $H_0$ if

$$\frac{1}{\sqrt{K}}\,(Z_1 + \ldots + Z_K) \;>\; z(\alpha), \tag{2}$$

where $z(\alpha)$ denotes the upper $\alpha$ percentage point of a standard normal N(0,1) distribution. A generalization proposed by Mosteller and Bush (1954) is the weighted inverse normal method whereby $H_0$ is rejected at level $\alpha$ if

$$w_1 Z_1 + \ldots + w_K Z_K \;>\; z(\alpha), \tag{3}$$

where fixed weights $\{w_k\}$ are chosen to satisfy $\sum_{k=1}^{K} w_k^2 \;=\; 1$. Clearly the left hand side of (3) has a standard normal distribution under $H_0$.

In fact, Mosteller and Bush (1954) proposed differential weighting of the Z-values based on sample sizes of the component studies. Hedges, Cooper and Bushman (1992) have pointed out that the Z-values and P-values "are already weighted" in some sense and so such adjustment may not be needed. However the rule (3) with weights based on sample sizes does have one very desirable property which we shall discuss in the next section.

Not included in Becker's (1994, Table 15.1) list, is the "maximum" combination test. This test does appear in the list of Goutis et al. (1996, Table 1). Here the $\alpha$-level test rejects $H_0$ if

$$\max{(P_1, \ldots, P_K)} \;<\; \alpha^{1/K}. \tag{4}$$

The FDA's "two pivotal trial" rule (FDA, 1995) can be considered a special case of this maximum method in which $K = 2$ and both trials need to be significant at the level 0.025 (one-sided) in order for $H_0$ to be rejected. Formally, as noted by *e.g.* Li and Huque (2003, p.623), this joint requirement for the two trials is a strict one as implies an overall significance level for the combination test of $\alpha = 0.025^2 = 0.000625$. However, as pointed out by the referee, the object of having two pivotal studies is not to require such a small significance level but rather to ensure replication of a finding that is not subject to systematic bias or fraud. Nevertheless, such extreme P-values *are* of interest. Fleming and Richardson (2004, p.668) quote instances where the FDA has viewed evidence of treatment effect with one-sided P-values in the range 0.0005–0.005 from just a single trial to be "compelling" and perhaps obviating the need for a second trial. (Of course, final approval will depend on a number of other factors, internal and external to the study.)

Note that for any of the combination tests we have described, an overall P-value is defined by finding that value of $\alpha$ for which equality is obtained in, respectively, (1), (2), (3) or (4).

Now, meta-analyses such as we have described are typically performed in order to obtain an over-view of a collection of studies found in the literature. We shall consider a different application: using these methods to combine data formally from a series of studies making up a drug development program. In following a program of studies, it is important to state at the outset which method will be used to combine the results from the studies. Of course, there is a a similar caveat when embarking on a meta-analysis of a collection of published studies — otherwise, one could try a number of the many different methods and pick the one giving the lowest combined P-value! In a program of research, the results of previous studies help determine the nature of the next study, including its sample size, and so the studies cannot be considered independent. (The same can be true for a collection of published studies but this feature is typically ignored!) This lack of independence means there is a danger, under some of the combination rules, that this can lead to inflation of the Type I error rate of the combination test.

Consider first the inverse $\chi^2$ test (1). Note that the $\{P_k\}$ are *statistically independent* under $H_0$ even if the results from one study are used to design another. This is because each has a $U(0,1)$ conditional distribution. (To be rigorous, we need to be able to apply Theorem 2 of Liu, Proschan and Pledger (2002). To invoke their theorem we note that, in any practical application, their measurability conditions will hold; and we have assumed that the P-values are continuous and uniformly U[0,1] distributed under $H_0$.) Hence (1) remains an $\alpha$-level test and the corresponding P-value remains valid. However, the $\{P_k\}$ are not necessarily statistically independent under $H_A$.

Combination tests can be used recursively. Suppose initially it is decided that there will be $K = 2$ studies and rule (1) will be employed. After observing the results and the P-value $P_1$ of the first study, it is decided to run two further studies rather than one, resulting in observed P-values $Q_1$ and $Q_2$, say. Then $P_2$ is defined as the solution to $-2\log(Q_1Q_2) = \chi^2_4(P_2)$. Applying the original rule, $P_1$ and $P_2$ are combined using (1) with $K = 2$. Note this is not the same as using (1) with $K = 3$ and P-values $P_1$, $Q_1$ and $Q_2$ and the latter approach would not give a valid $\alpha$-level test. This illustrates the importance of pre-specifying $K$ and the combination rule for $P_1, \ldots, P_K$. There is an obvious generalization to $K \geq 2$ of this recursive strategy.

Now we turn to the inverse normal method (2). Under $H_0$, each $Z_k \sim N(0,1)$ and, by the same reasoning as we used above for the P-values, the $\{Z_k\}$ are statistically independent — even if the results from one study are used to design another. To illustrate recursive use of (2), suppose that $K = 2$ is pre-specified but after observing $Z_1$ it is decided to run $M(Z_1)$ additional studies instead of one, resulting in Z-values $\tilde{Z}_1, \ldots, \tilde{Z}_{M(Z_1)}$. Let $Z_2 = (\tilde{Z}_1 + \ldots + \tilde{Z}_{M(Z_1)})/\sqrt{M(Z_1)}$, then, by first conditioning on $Z_1$, it is evident that $(Z_1 + Z_2)/\sqrt{2}$ is still standard normal under $H_0$. However, if we incorrectly ignore the dependence of $M$ on $Z_1$ and naively apply (2) with $K = M$, the resulting statistic is $(Z_1 + \tilde{Z}_1 + \ldots + \tilde{Z}_{M(Z_1)})/\sqrt{1 + M(Z_1)}$ and this is not standard normal under $H_0$. If $M(Z_1) = 1$ as originally specified, the two statistics coincide.

For fixed $K$ and weights $w_1, \ldots, w_K$ with $\sum w_k^2 = 1$, the weighted inverse normal test (3) retains level $\alpha$ if the results from one study are used to design another, for example, the sample size of study $k$ may be influenced by the previously observed $Z_1, \ldots, Z_{k-1}$. Suppose, as in the previous paragraph, that $K = 2$ and after observing $Z_1$ it is decided to replace the second study by $M(Z_1)$ studies resulting in Z-values $\tilde{Z}_1, \ldots, \tilde{Z}_{M(Z_1)}$. Suppose also that, before commencing the $M(Z_1)$ studies, weights $\tilde{w}_1, \ldots, \tilde{w}_{M(Z_1)}$ depending on $Z_1$ are specified and these satisfy $\tilde{w}_1^2 + \ldots + \tilde{w}_{M(Z_1)}^2 = w_2^2$. The test statistic (3) is still $w_1 Z_1 + w_2 Z_2$, where now $w_2 Z_2 = \tilde{w}_1 \tilde{Z}_1 + \ldots + \tilde{w}_{M(Z_1)} \tilde{Z}_{M(Z_1)}$. This idea can be applied recursively and one can use such a recursive construction to show the test (3) retains level $\alpha$ if, for $k = 2, \ldots, K$, the weights $w_k = w_k(Z_1, \ldots, Z_{k-1})$ are allowed to depend on previous outcomes (technically, they should be measurable functions — Liu, Proschan and Pledger (2002)). Moreover, the number of terms, $K$, can itself be response-dependent. The freedom to define the number of terms and their associated weights "adaptively" in this way provides a significant extension to the method. L. Fisher (1998) has termed this approach *variance spending* and proposed it as a method for designing adaptive group sequential trials: each choice of a weight $w_k$ "spends" an amount $w_k^2$ of the variance of final Z-statistic and the study terminates when this total reaches 1. The conditions on how weights in adaptive designs can be defined are crucial if the specified Type I error probability is to be achieved. If $K = 2$, the condition $w_1^2 + w_2^2 = 1$ implies there is no freedom to change the originally specified $w_1$ and $w_2$ after observing $Z_1$. Proschan and Hunsberger (1995) present a cautionary example of a two-stage adaptive design which breaks this rule since the second group size depends on $Z_1$ and weights $w_1$ and $w_2$ are set proportional to the square roots of the group sizes: for certain sampling rules, the Type I error of the test given by (3) is in excess of $2\alpha$!

The development of Fisher's (1998) variance spending designs illustrates how methods for combined inference based on several separate studies can be applied to combine data from separate stages of a single study. A multi-stage study in which the design of each stage can depend on results of previous stages is termed an "adaptive" or "flexible" procedure. Use of the inverse $\chi^2$ test (1) in this context was first advocated by Bauer (1989) and Bauer and Köhne (1994), whereas an adaptive version of the weighted inverse normal test (3) was first proposed by Fisher (1998).

## 3 Combining $n$ paired differences

Consider first a simple example of a balanced parallel arm trial with $2n$ patients treated in $K$ different centers. In center $k$, $1 \leq k \leq K$, immediate continuous responses $\{X_{A1k}, X_{A2k}, \ldots\}$ and $\{X_{B1k}, X_{B2k}, \ldots\}$ are available on treatment arms A and B, respectively. The responses are assumed to be independent from normal distributions $N(\mu_A + \nu_k, \sigma^2)$ and $N(\mu_B + \nu_k, \sigma^2)$. We shall take the common variance $\sigma^2$ to be known and, without loss of generality, equal to $\frac{1}{2}$. Hence the differences $X_{ik} = X_{Aik} - X_{Bik}$ are independent and identically distributed (i. i. d.) as $N(\theta, 1)$ where $\theta = \mu_A - \mu_B$ is the parameter of interest. The goal is to test the null hypothesis $\theta = 0$ versus the one-sided alternative $\theta > 0$ with pre-specified Type I error rate $\alpha$. Although this seems a very simple and restrictive example, by embedding the partial sums of the paired differences into a Brownian motion and working with information rather than sample size, the techniques for this prototype example can be applied in a wide variety of realistic applications. These include other designs, such as crossover or longitudinal trials, and other endpoints, for example, binary or survival data, fitted to generalized linear models or Cox's proportional hazards regression model; for details see Jennison and Turnbull (2000, Chap. 3).

Suppose we have a fixed total of $n$ paired differences $X_{ik}$, $1 \leq i \leq m_k, 1 \leq k \leq K$, where there are $m_1$ pairs in the first center, $m_2$ in the second and so on, and $n = \sum_{k=1}^{K} m_k$. The Z-statistic from center $k$ is given by

$$Z_k = \frac{1}{\sqrt{m_k}} \left( X_{1k} + \ldots + X_{m_k k} \right), \quad 1 \leq k \leq K.$$

A test of the hypothesis $H_0$: $\theta = 0$ could be carried out by combining P-values of tests based on the data from each of the $K$ individual centers using one of the methods described in Section 2. Alternatively since the $\{X_{ik}\}$ are i. i. d., the hypothesis $H_0$ can be tested using the overall Z-statistic for the total sample,

$$Z = \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i=1}^{m_k} X_{ik}. \tag{5}$$

In general these two tests will lead to different results. However, they will concur if we use the weighted inverse normal combination statistic (3) with weights $w_k = \sqrt{m_k/n}$. In this case, the value of $Z$ from (5) is identical to that in (3). This means we obtain the same test of $H_0$, whether we consider this as one study or as a combination of $K$ component sub-studies — as long as we use the inverse normal combination statistic with weights proportional to the square roots of the sub-study sample sizes. This property does not obtain for other combination methods. An important property of $Z$ here is that it is a function of the sufficient statistic $\sum_{k=1}^{K} \sum_{i=1}^{m_k} X_{ik}$, weighting each paired difference equally. This means that the procedure is statistically efficient.

Notice also that this combination test (and this one only) is invariant to an arbitrary partition of the data. Suppose we wish to use data from this multi-center study to play the role of "two pivotal studies" to meet the FDA's rule referred to in Section 2. Suppose also it is agreed that it will suffice to achieve some stated overall significance level when results from the two sub-studies are combined in a weighted inverse normal test with weights proportional to the square roots of the sub-study sample sizes. We might choose to take the first $K/2$ centers to make up Study 1 and the remainder to make up Study 2. Alternatively, one could take the odd numbered centers for Study 1 and the even ones for Study 2. Whatever arbitrary way the observations are partitioned into components, the Z-statistic (3) remains unchanged when using the inverse normal combination test with weights proportional to the square roots of sub-study sample sizes. This should be reassuring to a reviewer who might otherwise be concerned there could be bias caused by a selectively chosen partition of the data.

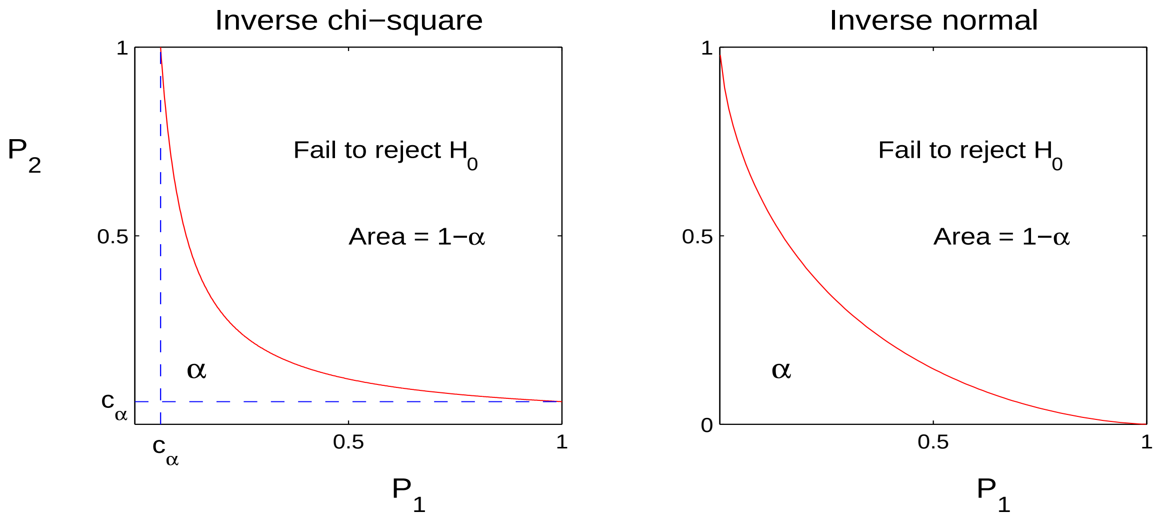
Note that, as argued in Section 2, this weighted combination rule will not maintain the Type I error rate $\alpha$ if the sample size in one sub-study can depend on results of others. Some other combination method must be used, such as an inverse normal test with fixed weights or the inverse $\chi^2$ test. But we then lose the efficiency benefits resulting from use of a sufficient statistic as well as the comfort of having a test that is invariant to different ways of partitioning of the data into sub-studies. From this, we see there is a tension between having the flexibility to design trials adaptively and being able to use the most natural test statistic for the inference — all driven by the requirement to maintain a fixed overall Type I error probability.

## 4 A graphical description of combination tests

We consider the combination tests of Section 2 for the case $K = 2$. We choose this case for simplicity but similar results can be envisioned for $K \geq 3$. Proschan (2003) has also exploited a graphical representation of two-stage tests in order to show their connection to positive quadrant tests.

We shall think of the two studies as ordered with Study 1 starting before Study 2, although

Figure 1: Rejection regions on probability (P) scale for two combination tests



their conduct can overlap chronologically. Recall that under $H_0$, the P-values $P_1$ and $P_2$ have independent $U(0,1)$ distributions. The range of $(P_1, P_2)$ is the unit square $[0,1] \times [0,1]$ and an $\alpha$ level test rejects $H_0$ if $(P_1, P_2) \in \mathcal{C}$, where the critical region $\mathcal{C}$ can be any subset of the unit square with area $\alpha$. The combination tests vary only with respect to the choice of this subset.

Our first example, Fisher's inverse $\chi^2$ test, rejects $H_0$ if

$$P_1 \cdot P_2 \;\leq\; c_\alpha \;=\; \exp(-\frac{1}{2} \cdot \chi_4^2(\alpha)). \tag{6}$$

The weighted inverse normal test rejects $H_0$ if

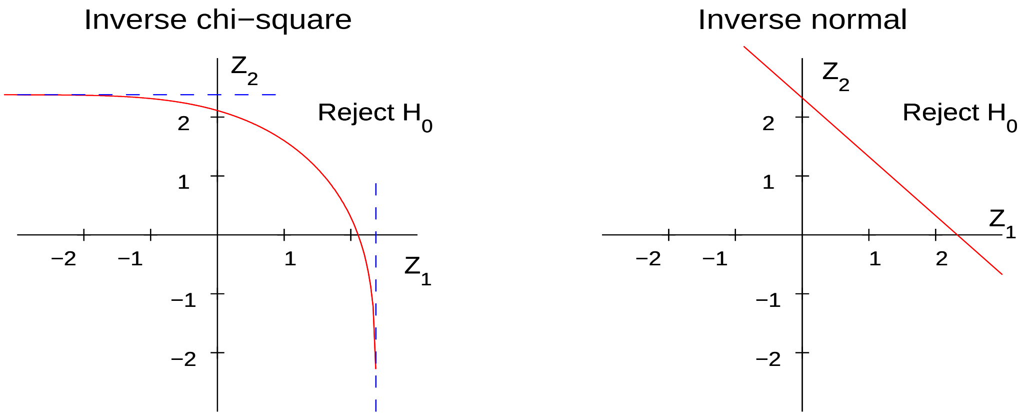
$$\sqrt{r}\; \Phi^{-1}(1 - P_1) + \sqrt{1-r}\; \Phi^{-1}(1 - P_2) \;\geq\; z(\alpha), \tag{7}$$

where $r = w_1^2 = 1 - w_2^2$ and, thus, $0 < r < 1$. These rejection regions are illustrated in Figure 1. The figure for the inverse normal test shows the unweighted case, for which $r = 0.5$.

Note that the inverse $\chi^2$ test will always reject $H_0$ if either $P_1$ **or** $P_2 \leq c_\alpha$. Thus, if $P_1 \leq c_\alpha$, there is no need to examine $P_2$ or even conduct the second study — rejection of $H_0$ is inevitable. No such curtailment is possible with the inverse normal test.

We can transform the axes in Figure 1 to a probit scale by plotting $(Z_1, Z_2)$ where $Z_k = \Phi^{-1}(1 - P_k)$ for $k = 1$ and 2. On this scale the critical regions for our two combination tests are

Figure 2: Rejection regions on probit (Z) scale for two combination tests



shown in Figure 2. Note that the boundary becomes linear for the inverse normal combination test. Here again, the inverse $\chi^2$ test will always reject $H_0$ if either $Z_1$ **or** $Z_2 \geq \Phi^{-1}(1 - c_\alpha)$.

Let us define the boundary of the critical region of a combination test like those in Figure 1 by the equation $P_2 = \tilde{A}(P_1)$. For the inverse $\chi^2$ test, we have

$$\tilde{A}(p_1) = \begin{cases} c_\alpha / p_1 & \text{for } c_\alpha < p_1 \leq 1, \\ 1 & \text{for } 0 \leq p_1 \leq c_\alpha. \end{cases}$$

For the inverse normal combination test, we have

$$\tilde{A}(p_1) = \Phi\left(\frac{\sqrt{r}\,\Phi^{-1}(1 - p_1) - z(\alpha)}{\sqrt{1 - r}}\right).$$

Note that, because the critical region has area $\alpha$, the function $\tilde{A}$ must satisfy

$$\int_0^1 \tilde{A}(p_1)\, dp_1 = \alpha. \tag{8}$$

Any non-increasing function $\tilde{A}(p)$ taking values on $[0, 1]$ and satisfying (8) is termed a *conditional (Type I) error function.* The hypothesis $H_0$ is rejected if and only if the second study P-value, $P_2$, is less than $\tilde{A}(P_1)$. In other words, the hypothesis test for the second study is carried out at significance level $\tilde{A}(P_1)$. The condition (8) ensures that the unconditional Type I error is $\alpha$, as required. The design of the second study does not have to be decided in advance of the results of the first study as long as significance level $\tilde{A}(P_1)$ is used. This connection with the conditional error probability has been pointed out by Posch and Bauer (1999), Wassmer (2000, Sec. 3.1.3) and Proschan (2003) among others.

There are of course many other choices for a conditional error function $\tilde{A}(P_1)$. Proschan and Hunsberger (1995) proposed a "circular" conditional error function given by

$$\tilde{A}(p_1) = \begin{cases} 1 & \text{for } p_1 \le c, \\ 1 - \Phi(\surd\{(\Phi^{-1}(1-c))^2 - (\Phi^{-1}(1-p_1))^2\}) & \text{for } c < p_1 < d, \\ 0 & \text{for } p_1 \ge d, \end{cases}$$

where $c$ and $d$ are constants chosen to satisfy (8). Using this conditional error function, the second study is not needed if either $P_1 < c$, for then $H_0$ is inevitably rejected (significance), or $P_1 > d$, in which case it is impossible for $H_0$ to be rejected and one may stop for futility if such a $P_1$ is observed.

It is possible to modify the the inverse $\chi^2$ and normal combination tests (6) and (7) so that they too have these curtailment features. Bauer and Köhne (1994) modify the inverse $\chi^2$ rule to allow stopping for futility after only the first study. This occurs if $P_1 > d$ and the increased probability of accepting $H_0$ is balanced by stopping to reject $H_0$ whenever $P_1 < c$ for a value of $c > c_\alpha$. The conditional error function is

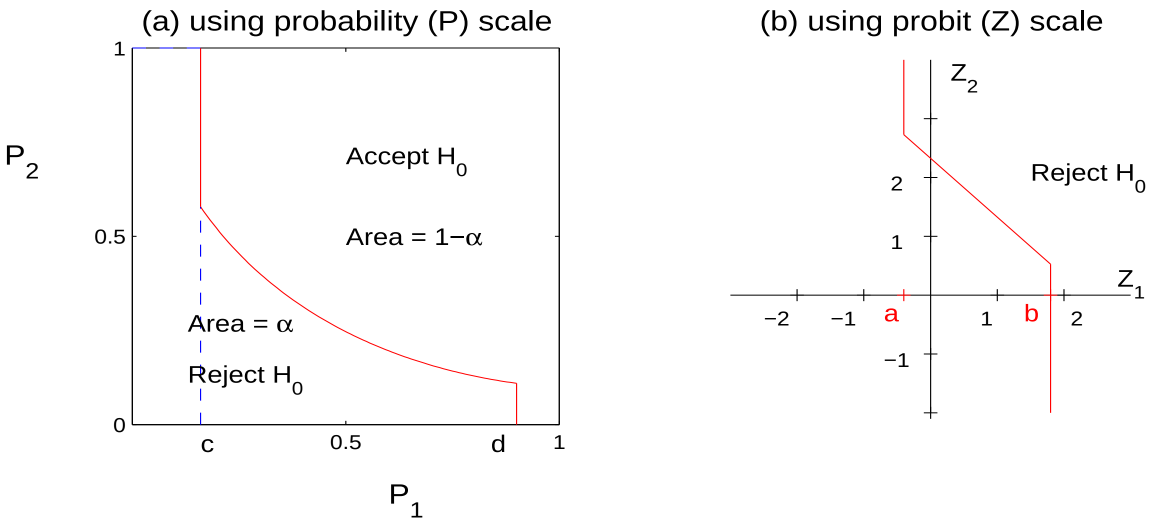
$$\tilde{A}(p_1) = \begin{cases} 1 & \text{for } p_1 \le c, \\ c_\alpha / p_1 & \text{for } c < p_1 < d, \\ 0 & \text{for } p_1 \ge d. \end{cases}$$

Here $c_\alpha = \exp(-\frac{1}{2} \cdot \chi^2_4(\alpha))$ as before, $c_\alpha < c < \alpha$ and $\alpha < d < 1$. The constants $c$ and $d$ must satisfy $\alpha = c + c_\alpha \cdot (\log d - \log c)$ to ensure the overall significance level (the area under the $\tilde{A}$ curve) is still $\alpha$. Table 1 of Bauer and Köhne (1994) gives pairs $(c, d)$ meeting these conditions for selected values of $\alpha = 0.01$, 0.025, 0.05 and 0.1.

The analogous modification of the weighted inverse normal test that allows both possibilities of rejection or acceptance of $H_0$ after only considering $P_1$ has conditional error function

$$\tilde{A}(p_1) = \begin{cases} 1 & \text{for } p_1 \le c, \\ \Phi\left(\frac{\sqrt{r}\,\Phi^{-1}(1-p_1) - z(\alpha)}{\sqrt{1-r}}\right) & \text{for } c < p_1 < d, \\ 0 & \text{for } p_1 \ge d. \end{cases}$$

Figure 3: Truncated rejection regions for inverse normal test



To ensure that the area under the $\tilde{A}$ curve equals the specified $\alpha$, the constants $c$ and $d$ must be chosen so that

$$\alpha = c + \int_c^d \Phi\left(\frac{\sqrt{r}\,\Phi^{-1}(1-p) - z(\alpha)}{\sqrt{1-r}}\right)\,\mathrm{d}p.$$

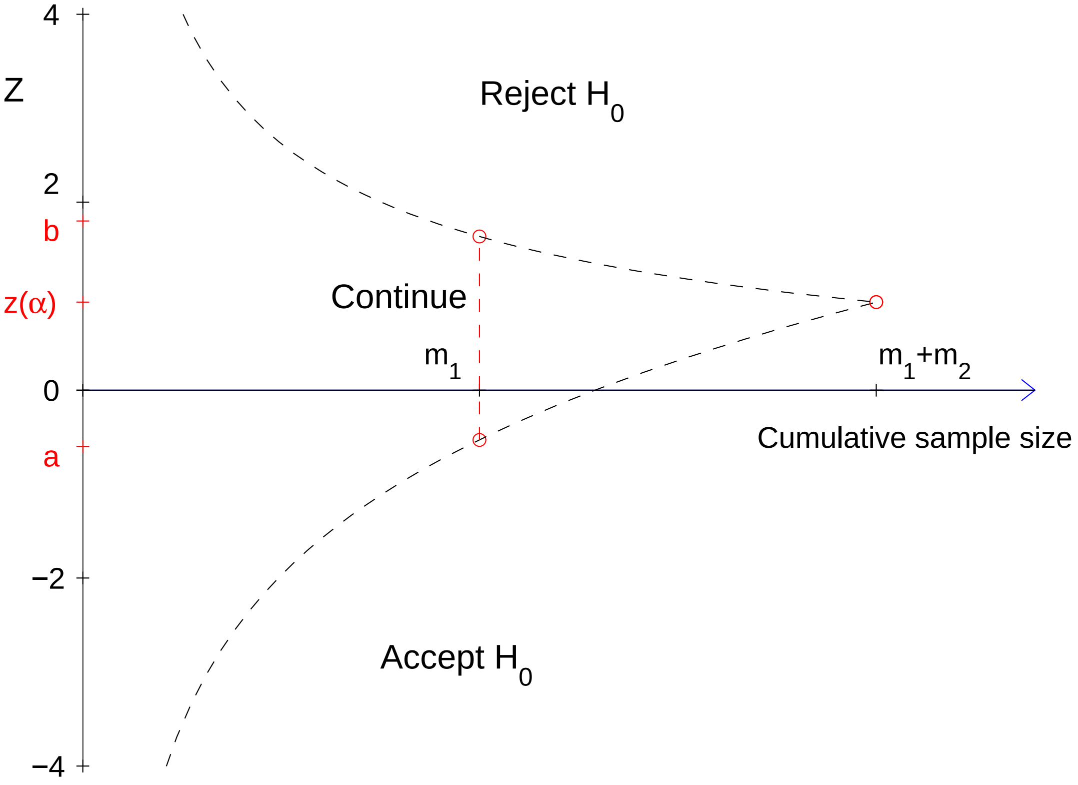
The rejection regions on the probability (P) scale and on the probit (Z) scale are shown, respectively, in the left and right panels of Figure 3. In the right panel, the critical values $a$ and $b$ are defined by

$$a = \Phi^{-1}(1-d) \quad \text{and} \quad b = \Phi^{-1}(1-c). \tag{9}$$

Hence, with only the results ($P_1$ or $Z_1$) of the first study at hand, if $Z_1 \geq b$ we may reject $H_0$ or if $Z_1 \leq a$ we can immediately declare the results as non-significant and abandon the second study as futile. Only if $a < Z_1 < b$ do we need to look at the results of the second study and then we reject $H_0$ when $\sqrt{r}\,Z_1 + \sqrt{1-r}\,Z_2 \geq z(\alpha)$.

We now return to the situation of Section 3 in which the data could be viewed as i. i. d. normally distributed. We take $K = 2$ but instead of viewing the data as arising from two studies or from two centers, we consider the data as coming from two successive stages of a single group sequential trial. Suppose also that we agree to use the weighted inverse normal test with weights proportional to the square roots of the group sizes. In the notation of the current section, this implies $r = m_1/n$ and

Figure 4: Boundary for a 2-stage group sequential test



$1-r = m_2/n$. With curtailment, the trial stops early at the first stage if $Z_1 \leq a$ or $Z_1 \geq b$, otherwise it continues to the second stage where $H_0$ is rejected only if $Z = (\sqrt{m_1}\, Z_1 + \sqrt{m_2}\, Z_2)/\sqrt{n} \geq z(\alpha)$. As noted in Section 3, for this weighting the final $Z$-value coincides with the overall $Z$-statistic for the total sample given by (5). This means that the rejection plan of Figure 3(b) can be displayed in the alternative fashion shown in Figure 4. Here the horizontal axis represents cumulative sample size ($m_1$ or $n = m_1 + m_2$) and the vertical axis, labelled $Z$, is the standardized $Z$-statistic, $Z_1$ at the first stage or $Z$ at the second stage, based on the cumulative sample at each point.

It can be seen that Figure 4 is the more familiar representation of a discrete boundary for a conventional group sequential test, in this case with $K = 2$ stages. For a general group sequential test, the critical value for $Z$ at the second stage need not equal $z(\alpha)$ — the requirement of all boundary points taken together is that the total probability under $H_0$ of rejecting $H_0$ should equal $\alpha$. Of course, for $K \geq 3$, representations such as those in Figures 1, 2 and 3 are more awkward because the rejection regions are subsets in $K$-dimensional space. On the other hand, if we are in the situation of Section 3, the discrete boundary representation as shown in Figure 4 is easily generalized to $K \geq 3$; see, for example, the cover diagram (also Figure 4.2) of Jennison and Turnbull (2000). The correspondence between the two representations exists only if the weighted inverse normal combination method is used with weights proportional to the square roots of the group sizes. This combination rule has, therefore, special significance as it produces tests in which decisions are functions of the sufficient statistics at each stage and, if the critical values are chosen correctly, tests can coincide with the optimal group sequential tests of Eales and Jennison (1992) and Barber and Jennison (2002).

It is a standard assumption in group sequential designs that group sizes are not allowed to depend on the results of previous stages. If group sizes do have such dependence, the overall Type I error rate $\alpha$ of a standard group sequential test may no longer be maintained, just as we saw for combination tests in Section 3. It is, however, possible to generalize group sequential designs to allow a pre-specified dependence of group sizes on observed data, selecting critical values to give overall Type I error probability $\alpha$. Such adaptive group sequential designs were proposed by Schmitz (1993) and Jennison and Turnbull (2004a, 2004b) have studied the efficiency gains they can provide relative to non-adaptive designs — finding the gains to be slight and unlikely to justify the extra complexities of the adaptive designs. Another reason for wishing to modify a study is to respond to new developments, internal or external to the study, which alter the investigators' objectives. Such adaptation can be achieved without affecting the Type I error probability, the key requirement being that the conditional Type I error probability should be maintained under the modified design; see Denne (2001) and Müller and Schäfer (2001). Thus, the function $\tilde{A}(p_1)$ plays a key role but now it is defined implicitly by the original design, as stated in the trial protocol. We shall discuss the use of such unplanned mid-course adaptations in the example of Section 6.

## 5 Example 1: Seamless transition between Phase IIb and Phase III trials

A phase IIb trial is used in some drug development programs to choose one of several dose levels while also comparing this set of treatments with an active control or placebo. The dose selected will be studied in Phase III: it may be chosen on the basis of the best response on the primary endpoint or, particularly for cancer treatments, it may be the highest dose that can be tolerated without serious toxicity. Before embarking on the Phase III study, investigators have the chance to modify details of the treatment and, if appropriate, re-define the primary response.

It is highly desirable to avoid delay between the end of a Phase IIb trial and the start of the subsequent Phase III trial. One way to streamline the process is to present both trials in a combined protocol. It may also be permissible to combine evidence of treatment benefit from the Phase IIb and Phase III trials. We shall illustrate how methods described earlier in this paper can be used in this context.

Consider a study of a new treatment for cholesterol reduction in which 3 dose levels of the new treatment are to be compared against a placebo control in Phase IIb. The study will enlist a total of 100 patients, the response for each being the reduction in serum cholesterol over an 8 week period. A trend test will be used to test for a treatment effect, linear in dose level. The null hypothesis at this stage is that there is no treatment effect at any dose, and this is tested against the alternative of a linear increase in effect with dose level, the one-sided P-value of this test being converted into a Z-value, $Z_1$. Although the whole investigation may be abandoned if no treatment effect is seen at any dose, this decision is likely to depend on the effects observed at each dose level as well as the linear trend statistic, $Z_1$. We shall not, therefore, attempt to formalize the decision to abandon the process after Phase IIb but we note that assuming continuation on to Phase III for all values of $Z_1$ introduces some conservatism into the overall Type I error rate.

The natural choice of dose for evaluation in the Phase III trial is the level producing the greatest reduction in serum cholesterol in Phase IIb but investigators are free to consider other aspects of treatment or response in making this choice. We suppose the Phase III trial is run as a two-sample comparison, again taking reduction in serum cholesterol over 8 weeks as the primary endpoint. The null hypothesis for this trial is that there is no difference between the selected treatment and the placebo control, tested against the alternative of a beneficial treatment effect at this dose. For now, suppose this trial has a fixed sample size of 200 subjects and a one-sided $t$-test is carried out giving P-value $P_2$, which converts to the Z-value $Z_2$.

Data from the two phases are combined in the overall test statistic

$$Z \; = \; w_1 Z_1 + w_2 Z_2$$

where the positive weights $w_1$ and $w_2$ are specified in the joint protocol, before the Phase IIb trial starts, and satisfy $w_1^2 + w_2^2 = 1$. Overall, the null hypothesis tested is the intersection of null hypotheses for the two stages, which in this case is the null hypothesis for the Phase IIb trial, $H_0$: no treatment effect at any dose level. If the overall Type I error probability is set at $\alpha$, then $H_0$ is rejected if and only if $Z > z(\alpha)$.

Note that if the overall null hypothesis is rejected, there is still work to do in arguing that the dose tested in Phase III, which was selected in a data-dependent way, delivers a beneficial effect. Ideally, one would like to be able to point to evidence in the Phase IIb data of a linear dose-response so that, under simple model assumptions, positive contributions from other dose levels to $Z_1$, and hence to the combined statistic $Z$, lend support to an effect at the selected dose.

Within this framework, there is considerable flexibility.

1. Changes can be made to the definition of treatment or response for the Phase III trial. There may, for example, be minor changes to the way the treatment is administered. In defining the response, analysis of Phase IIb data could suggest that reduction in serum cholesterol occurs over a different timescale and the 8 week period may be replaced by a longer or shorter interval.

2. It is possible to add a formal rule that the whole sequence of studies stops with acceptance of $H_0$ if $Z_1 < a$ after the Phase IIb trial. Then, the threshold for $Z$ to be significant can be re-calculated to give overall Type I error rate equal to $\alpha$. This is an example of a test allowing curtailment at stage 1, as discussed in Section 4, but in this case only early stopping for futility is permitted. A drawback of this approach is that, since this early stopping is only for a negative outcome, the critical value for $Z$ will now be lower than $z(\alpha)$ and it may be difficult to make the case that a $Z$-statistic below $z(\alpha)$ represents significant evidence at level $\alpha$. Furthermore, there is no guarantee that a formal rule will be adhered to: if the Phase IIb data show a non-linear dose response, investigators may wish to continue to Phase III on the strength of good results at one dose level, despite a low trend statistic, $Z_1$.

3. If the response variance observed in Phase IIb data differs substantially from that anticipated, the sample size of 200 for Phase III may be adjusted to maintain desired power at a specific effect size.

4. The Phase III trial may be monitored group sequentially; see Jennison and Turnbull (2000, Sec. 4.4) for a description of group sequential, one-sided $t$-tests. One could simply run the Phase III trial according to a group sequential plan with Type I error rate 0.05, say, then the overall conclusion would be to reject $H_0$ if the one-sided P-value on termination, $P_2$, corresponds to a $Z$-statistic, $Z_2$, satisfying

$$w_1 Z_1 + w_2 Z_2 > z(\alpha).$$

For computation of P-values upon termination, see Jennison and Turnbull (2000, Sec. 8.4). For $H_0$ to be rejected, we need

$$Z_2 \; > \; \frac{z(\alpha) - w_1 Z_1}{w_2}$$

i.e.,

$$P_2 \; < p(Z_1) \; = \; 1 - \Phi^{-1}\left\{\frac{z(\alpha) - w_1 Z_1}{w_2}\right\}.$$

Given that the output of Phase III will be used in this way in reaching the overall conclusion, it makes sense to conduct the Phase III trial as a group sequential test with Type I error probability $p(Z_1)$. Then, the one-sided P-value on termination will be less than $p(Z_1)$ if and only if this test rejects its null hypothesis. In this formulation, the outcome of the Phase III trial has been made to agree with the overall conclusion. Efficiency will be served by creating a group sequential stopping rule which leads to as early a decision as possible in the Phase III trial to accept or reject the null hypothesis. This is an example of recursive design, as discussed in Section 2

5. Ideally, accrual should switch straight from Phase IIb to Phase III so that all eligible patients can be entered into one or other study as they present themselves for treatment. There is a difficulty in doing this since it will take at least 8 weeks from the end of accrual before all the Phase IIb responses are known. Rather than allow a hiatus in accrual, we propose making a choice of the treatment for the Phase III trial using the phase IIb data available when the the Phase IIb recruitment period ends. Follow-up data will continue to come in and this will be included in the calculation of $Z_1$ in the usual way. The delay in learning the value of $Z_1$ is another reason against having a formal rule to stop for futility if $Z_1 < a$. Group sequential monitoring of Phase III, as described in item 4, is still feasible. The key point is to start Phase III accrual promptly, keeping study monitors blind to any Phase III data until

the Phase IIb follow-up is completed and a group sequential stopping rule for Phase III has been decided. Given the complexity of this scheme, it is desirable to specify a template of the Phase III design as fully as possible in the overall protocol. For example, the type of group sequential test and frequency of monitoring could be stipulated, ready for the required Type I error rate, $p(Z_1)$, to be incorporated as soon as it is known.

It is clear that this approach is only practical if accrual periods in both Phase IIb and Phase III trials are long in comparison to the response time. If not, too few of the Phase IIb responses will be available to make a good choice of treatment for Phase III, or the first interim analysis in Phase III will be later than desired due to the delay in completing the analysis of Phase IIb data — which has to be done before the Phase III group sequential design can be finalized. However, where the approach is feasible, the continuity between the two phases offers an opportunity to make significant gains.

# 6 Example 2: Mid-course adaptation within a Phase III trial

## 6.1 Initial formulation

We consider a Phase III trial designed to achieve Type I error rate $\alpha$ and power $1-\beta$ when the true treatment effect is equal to $\delta$. This could, quite possibly, be part of a sequence of studies, as described in Section 5. The Phase III trial could be a fixed sample size study or it could be planned with a group sequential design. There is now a range of methods in the literature that allow investigators to re-consider the trial design at an interim point and, if desired, modify the remaining part of the trial. In a group sequential trial, it is natural to do this at one of the planned interim analyses. In a fixed sample study, reasons may still arise to consider a design change at an intermediate point, possibly following an unplanned inspection of the data gathered thus far. We shall relate some of the proposals for adaptive re-design to the meta-analysis formulation, discuss what such adaptive modifications can achieve, and comment on the advantages and disadvantages of their use in various situations.

As a specific example, we shall consider a two treatment comparison with responses $X_{Ai} \sim N(\mu_A, \sigma^2)$ on treatment A and $X_{Bi} \sim N(\mu_B, \sigma^2)$ on treatment B, where the treatment effect is taken to be $\theta = \mu_A - \mu_B$. A fixed sample size study needs a sample size of

$$n_f \;=\; \{z(\alpha) + z(\beta)\}^2 2\sigma^2/\delta^2$$

per treatment arm to meet the Type I error rate and power requirements described above.

In a group sequential version of this design, analyses may be performed after groups of $g$ observations per treatment. At the $k$th analysis, the Z-statistic based on all the data collected thus far is calculated and the study stops to accept $H_0$ if $Z_k < a_k$ or to reject $H_0$ if $Z_k > b_k$. The maximum number of groups $K$ is fixed and the group size $g$ and critical values $(a_k, b_k)$, $k = 1, \ldots, K$, are chosen to ensure the required Type I error rate and power at $\theta = \delta$ are attained. The maximum sample size per treatment arm, $gK$, is equal to $Rn_f$ where $R > 1$ is termed the "inflation factor" and its value depends on the form of testing boundary and maximum number of analyses, $K$; see, for example, Jennison and Turnbull (2000).

## 6.2 Connections between adaptive designs and meta-analysis

Bauer (1989) and Bauer and Köhne (1994) propose a method for modifying the design when a study is already in progress. An interim point at which modifications can be made is specified in the study protocol, as is a rule for combining the P-values, $P_1$ and $P_2$, calculated from data obtained in the two parts of the study, either side of this interim point. These authors use the inverse $\chi^2$ rule but any other rule is allowable, as long as it is pre-specified. This method of combining two sets of data has a clear interpretation as a meta-analysis of the two sub-studies. Since the P-value $P_2$ typically has a $U(0,1)$ null distribution, conditional on the outcome of the first part of the study (see Section 2), the combination rule remains valid after the re-design. This approach provides flexibility for investigators to modify aspects of the study at the interim re-design point: the treatment may be refined, dosage may be adjusted, even the response variable may be altered. Such changes parallel the evolution of a treatment seen over the course of a sequence of separate studies but allowing this to happen within a single trial offers the prospect of more rapid treatment development.

A more challenging problem is to preserve the Type I error rate when making mid-course design changes if this eventuality was not considered in the original plan. Consider first the case of the fixed sample study described above with $n_f = 100$ and suppose there is reason to modify the design after observing $m_1$ subjects on each treatment. Define the Z-statistics from current and future data

$$Z_1 \;=\; \frac{\sum_1^{m_1}(X_{Ai} - X_{Bi})}{\sqrt{m_1 2\sigma^2}} \quad \text{and} \quad Z_2 \;=\; \frac{\sum_{m_1+1}^{100}(X_{Ai} - X_{Bi})}{\sqrt{m_2 2\sigma^2}},$$

where $m_2 = 100 - m_1$. In the final analysis originally planned, the Z-statistic is

$$Z \;=\; \frac{\sum_1^{100}(X_{Ai} - X_{Bi})}{\sqrt{100 \cdot 2\sigma^2}} \;=\; \frac{\sqrt{m_1}\; Z_1 \,+\, \sqrt{100 - m_1}\; Z_2}{\sqrt{100}} \tag{10}$$

and $H_0$ is rejected if $Z > z(\alpha)$. The P-values for the two parts of the data are $P_1 = 1 - \Phi^{-1}(Z_1)$ and $P_2 = 1 - \Phi^{-1}(Z_2)$ and, in terms of these P-values, the final decision is to reject $H_0$ if

$$\frac{\sqrt{m_1}\;\Phi(1 - P_1) \,+\, \sqrt{100 - m_1}\;\Phi(1 - P_2)}{\sqrt{100}} \;>\; z(\alpha). \tag{11}$$

Thus, a combination rule for $P_1$ and $P_2$ is implicit in the original plan and the "meta-analysis approach" can be followed using this rule. Specifically, design modifications can be made after observing the first $m_1$ observations per treatment; for example, the total sample size can be increased beyond the originally planned $n_f = 100$. The P-value $P_2$ is calculated from the data collected after this re-design point and this is combined with $P_1$ using the rule (11), keeping these original weights.

The equivalent representations of the combination rule discussed in Section 4 are also applicable. The decision rule can be represented as a critical region $\mathcal{C}$ for the pair $(P_1, P_2)$, that is, $H_0$ is rejected if and only if the pair $(P_1, P_2)$ lies in $\mathcal{C}$. Alternatively, if the function $\tilde{A}(P_1)$ defines the boundary of the critical region, $H_0$ is rejected if $P_2 < \tilde{A}(P_1)$. As we saw in Section 4, $\tilde{A}(P_1)$ is the conditional Type I error probability given the observations at the re-design point and this gives rise to a simply stated principle for re-design: first calculate the current conditional Type I error probability if the study were to continue following the originally specified design, $\tilde{A}(P_1)$, then conduct the remainder of the study with whatever modifications are deemed appropriate and reject $H_0$ at the end if the P-value for the remainder of the study, taken by itself, is less than $\tilde{A}(P_1)$.

Using (10), we can just as easily work with Z-statistics and derive a critical value for the standardized statistic $Z_2$ based on the data from the remainder of the study. In our example, $H_0$ is rejected if

$$w_1 Z_1 + w_2 Z_2 \;>\; z(\alpha)$$

where $w_1 = \surd\}m_1/100\}$ and $w_2 = \surd\{(100 - m_1)/100\}$, and the criterion for rejecting $H_0$ is

$$Z_2 \;>\; \{z(\alpha) - w_1 Z_1\}/w_2.$$

This approach to mid-course design changes is by no means limited to such simple examples but generalizes readily to other response types and test statistics. It also encompasses modifications during a group sequential trial. Suppose a trial is being conducted according to the group sequential design outlined in Section 6.1, the trial has continued up to analysis $\tilde{k}$ where the statistic $Z_{\tilde{k}}$ lies in the continuation region $(a_{\tilde{k}}, b_{\tilde{k}})$, and re-design is considered at this point. It is possible to work in terms of P-values, $P_1$ based on data up to analysis $\tilde{k}$ and $P_2$ on data from analysis $\tilde{k}$ onwards, and identify the critical region $\mathcal{C}$ of pairs $(P_1, P_2)$ for which $H_0$ is rejected. However, the account given by Müller and Schäfer (2001) shows this approach is complex and the construction is not particularly intuitive. It is simpler to move straight on to the equivalent definition in terms of conditional Type I error probability, $\tilde{A}(P_1)$. This is the probability under $H_0$ that the $\{Z_k\}$ process, starting at the current value of $Z_{\tilde{k}}$, will cross the upper boundary by analysis $K$, and it is easily found using standard group sequential calculations. The remainder of the study can be run as a new group sequential test with Type I error probability $\tilde{A}(P_1)$ incorporating whatever design changes are desired. A real advantage of this construction is that it allows investigators to make an initial choice of an efficient group sequential design knowing that under normal conditions this will be used throughout the trial but there is the reassurance that, if necessary, it will be possible to make modifications in a way that preserves the Type I error rate.

Other methods for adapting group sequential designs fit into this framework. In the proposal by Cui et al. (1999), group sizes after the re-design point are multiplied by a factor chosen to increase power and the original boundary values $(a_k, b_k)$ are applied to modified versions of the usual Z-statistics. The authors give a direct proof that the overall Type I error probability is preserved but it can also be seen that the method preserves the conditional Type I error probability. In fact, this conditional property must hold by the argument of Jennison and Turnbull (2003, Sec. 3.2) which implies that any unplanned mid-course re-design that maintains the overall Type I error rate must preserve the conditional Type I error probability whenever re-design occurs.

The "variance spending" scheme proposed by Shen and Fisher (1999) is described in terms of Z-statistics. This method starts with a fixed sample test and at a re-design point the criterion for this test to reject $H_0$ is written as $w_1 Z_1 + w_2 Z_2 > z(\alpha)$ where $w_1^2 + w_2^2 = 1$, $Z_1$ is the Z-statistic summarising current data and $Z_2$ is the Z-statistic still to come from future data. The basic design modification is to change the sample size on which $Z_2$ will be based and to use the Z-statistic based on this new sample size in the original rule. Applying such a change recursively gives a form of group sequential test.

## 6.3 Merits of various uses of re-design

Many different reasons have been given for why investigators may wish to make mid-course design modifications. Some of these appear more reasonable than others and in investigating some proposals we have found a danger of producing seriously inefficient experimental designs. We conclude this paper with a description of five commonly cited reasons for design modification and offer our views on the merits of each.

1. *Changes in response to external factors.* An example is the withdrawal of a competing product which leaves a gap in the market. If this had been known earlier, a smaller treatment effect could have been deemed clinically beneficial and there would also have been good financial motivation for a larger study, able to detect a smaller effect size. When such information arrives during the course of a study, it is reasonable to enlarge the current study in order to increase power at smaller effect sizes. If the people making this decision are not completely blind to the data accumulated thus far, it is advisable to make any sample size change in a manner that maintains the Type I error rate in order to preserve the study's credibility. The methods described in Section 6.2 allow such changes to be made. Clearly, it is unfortunate that the relevant information should not have been known sooner as it would have been simpler to plan a larger, more powerful study at the outset. But, when information arrives during a study it is very valuable to have methods available that allow adaptation to the changed objectives.

2. *Responding to internal factors: nuisance parameters.* For some studies it is difficult to calculate the sample size needed to satisfy a power requirement as this depends on unknown factors such as the variance of normal responses or the baseline hazard rate and censoring rates for survival data. Information on such factors will accrue during the course of a study and, if initial estimates have been off target, it is good to be able to adjust the sample size later. Adaptive methods provide a framework for doing this. Since this problem is likely to be recognized in advance, these modifications may be pre-planned. In the case of normal responses with unknown variance, $t$-statistics calculated from the data in each stage can

be converted to exact P-values so the Type I error rate is attained precisely. It should be noted that there are other ways of coping with the problem of nuisance parameters: there is a large literature on internal pilot designs (see, for example, Wittes and Brittain (1990) and further references in Jennison and Turnbull (2000, Chap. 14)) or, in group sequential applications, error spending tests with maximum information designs (Jennison and Turnbull (2000, Chap. 7) provide a good solution.

3. *Responding to internal factors: safety variables.* Several authors mention the impact that information on the new treatment's safety can have on the suitability of a study's power objective. One possibility is that evidence of good safety properties for the experimental treatment could indicate that high power should be sought at a lower effect size than that originally used in setting the study's power. This would require an increased sample size and adaptive methods provide a means to do this. Another conclusion might be that, with good safety outcomes, it should suffice to demonstrate non-inferiority, rather than superiority, of the new treatment. This corresponds to shifting the power curve along the effect size scale: this may not need a change in sample size but could prompt a change to the group sequential stopping rule being applied to the primary endpoint. A problem here is that it is not a requirement that the original Type I error rate should be preserved at the null hypothesis of zero effect size; so the adaptive methods we have described do not solve the relevant problem. Of course, interim results on safety are subject to sampling variability and one should be wary of basing substantial design changes on noisy information. An alternative is available in Jennison and Turnbull's (1993) group sequential tests for bivariate data which monitor both safety and efficacy outcomes together.

4. *Rescuing an under-powered study.* The example Cui et al. (1999) present to motivate their work concerns a study of a treatment designed to reduce the risk of myocardial infarction in patients undergoing coronary artery bypass graft surgery. Initially, the investigators hoped for a reduction in incidence of 50% and the study was designed to achieve 95% power in this case. Interim data showed a smaller decrease in incidence of around 25% which was still a worthwhile improvement both clinically and commercially. At this late stage, it became clear that an increase in power at smaller effect sizes was desirable, but at the time there were no methods to implement such a change. Currently available methods would have allowed this

modification to be made while safeguarding the Type I error rate.

With hindsight, the simple solution would have been to design the study with a larger sample size and higher power at the outset, avoiding the need for a mid-course design change. We would certainly recommend that investigators think through scenarios with disappointing results in advance to gain a fuller understanding of what they really wish to see in a study's power curve. Then, they can choose a design achieving the correct objective from the start. This approach should yield benefits of lower expected sample size as modifying a study's objectives in mid-course leads to inefficient procedures: see Jennison and Turnbull (2004b, Sec. 2.3). However, the number of accounts of under-powered studies suggests this problem will continue to occur and, while it does, the availability of adaptive methods will be valuable.

5. *Responding to internal information on the primary endpoint.* Whereas we have presented the increase of sample size in an under-powered study as a "rescue" process with implicit criticism of the poor initial design, some authors view the flexibility of adaptive procedures as a positive advantage and advocate their use to modify power as new estimates $\widehat{\theta}$ of the effect size are obtained. It may well be that modifying an initial fixed sample size procedure yields an overall adaptive scheme with higher power but a smaller increase in average sample size than the fixed sample test needed to achieve this power: the examples Proschan and Hunsberger (1995) give of their "designed extension procedure" illustrate this point. However, if the adaptive procedure has two stages, the appropriate comparison is with a group sequential test with $K = 2$ analyses and in this comparison the Proschan and Hunsberger (1995) designs prove inferior.

   We remarked in Section 4 that there is a well developed methodology for constructing group sequential tests optimized to specific criteria. These designs use the sequence of estimates $\widehat{\theta}$ in their stopping rules and, by definition searching for new types of group sequential procedure to beat the optimum tests is futile. The freedom to let group sizes depend adaptively on current data does represent a possible source of improved efficiency but calculations reported by Jennison and Turnbull (2004a, 2004b) show the benefits of this to be insignificant. On the other hand, it is quite possible that adaptive schemes defined with what appear to be sensible sample size rules and stopping rules may be substantially less efficient than optimal group sequential tests with the same number of analyses. We have assessed a number of adaptive schemes described in the literature in terms of their overall power curves and average sample size functions, and found a high proportion to be quite inefficient. Given the availability of simply defined, efficient group sequential tests (Jennison and Turnbull (2004a)), our simple recommendation is to look no further. If, for some reason, an adaptive scheme really is deemed desirable, we suggest that its power curve and average sample size function be computed and compared against a standard group sequential design to make sure it has adequate efficiency.

Overall, our conclusion is that the key advantages of adaptive methods are (a) the ability to make mid-study modifications in response to external factors or, possibly, internal information on safety outcomes and (b) the ability to rescue an under-powered study when this problem becomes apparent during the study. These adaptations can be implemented in fixed sample and group sequential tests and, by ensuring the conditional Type I error probability remains unchanged, the overall Type I error rate will be maintained. On the other hand, we see no point in pursuing varieties of adaptive design as a novel alternative to standard group sequential tests: simple and highly efficient designs are already known and well understood and, unless the adaptive designs have other advantageous features, the very best they will do is gain equality with existing methods.

## REFERENCES

Barber, S. and Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika* **89**, 49–60.

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* **20**, 130–148.

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

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.

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.

Eales, J.D. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24.

Egger, M., Smith, G.D. and Altman, D.G. (Eds.) (2001). *Systematic Reviews in Health Care: meta-analysis in Context, 2nd Ed.* BMJ Publishing Co., London.

FDA. (1995). Statement regarding the demonstration of effectiveness of human drug products and devices. *Federal Register* **60** (No. 147), 39180–39181 (1 August 1995).

FDA. (1998). E9: Statistical Principles for Clinical Trials. *Federal Register* **63** (No. 179) 49583—49598 (16 September 1998).

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. and Richardson, B.A. (2004). Some design issues in trials of microbicides for the prevention of HIV infection. *Journal of Infectious Diseases* **190**, 666–674.

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

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, New York.

Hedges, L.V., Cooper, H.M. and Bushman, B.J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin* **111**, 188–194.

Jennison, C. and Turnbull, B.W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741–752.

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. (2004a). Efficient group sequential designs when there are several effect sizes under consideration. *University of Bath, Mathematics preprint*, 04/11.
(Available at `http://www.bath.ac.uk/~mascj/` )

Jennison, C. and Turnbull, B.W. (2004b). Adaptive and non-adaptive group sequential tests. *University of Bath, Mathematics preprint*, 04/112.
(Available at `http://www.bath.ac.uk/~mascj/` )

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

Li, Q.H. and Huque, M.F. (2003). A decision rule for evaluating several independent trials collectively. *J. Biopharmaceutical Statistics* **13**, 621–628.

Liu, Q., Proschan, M.A. and Pledger, G.W. (2002). A unified theory of two-stage designs. *J. Amer. Statist. Assoc.* **97**, 1034–1041.

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

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.

Padian, N.S. (2004). Evidence-based prevention: Increasing the efficiency of HIV intervention trials. *Journal of Infectious Diseases* **190**, 663–665.

Posch, M. and Bauer, P. (1999). Adaptive two-stage designs and the conditional error function. *Biometrical Journal* **41**, 689–696.

Posch, M., Bauer, P. and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–969.

Proschan, M.A. (2003). The geometry of two-stage tests. *Statistica Sinica* **13**(1), 163–177.

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

Robins, J.M., van der Vaart, A. and Ventura, V. (2000). Asymptotic distribution of $P$ values in composite null models. *J. Am. Statist. Assoc.* **95**, 1143–1156.

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

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190–197.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. and Williams, R.M. (1949). *The American Soldier: Adjustment during Army Life* (Vol.1). Princeton University Press.

Sutton, A.J., Abrams, K., Jines, D., Sheldon, T. and Song, F. (2000). *Methods for Meta-analysis in Medical Research*, Wiley, Chichester, England.

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

Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers* **41**, 253–279.

Whitehead, A. (2002). *Meta-analysis of Controlled Clinical Trials*, Wiley, Chichester, UK.

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