STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3366–3381
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2258

Estimation in flexible two stage designs

Werner Brannath*,†, Franz König and Peter Bauer

Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Wien, Austria

SUMMARY

Adaptive test designs for clinical trials allow for a wide range of data driven design adaptations using all information gathered until an interim analysis. The basic principle is to use a test statistics which is invariant with respect to the design adaptations under the null hypothesis. This allows for a control of the type I error rate for the primary hypothesis even for adaptations not specified *a priori* in the study protocol. Estimation is usually another important part of a clinical trial, however, is more difficult in adaptive designs. In this research paper we give an overview of point and interval estimates for flexible designs and compare methods for typical sample size rules. We also make some proposals for confidence intervals which have nominal coverage probability also after an unforeseen design adaptation and which contain the maximum likelihood estimate and the usual unadjusted confidence interval. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: adaptive design; flexible confidence interval; invariance principle; maximum likelihood estimate; mean unbiased estimate; median unbiased estimate

1. INTRODUCTION

In the last two decades flexible or adaptive designs have been suggested which allow for mid-trial design modifications that are based on the unblinded interim data without compromising the overall type I error rate [1–10]. Examples for design modifications are the adaptation and reallocation of sample sizes, the adaptation of the study goal (non-inferiority and superiority) [11, 12], the test statistics [13–17], and the number of interim analyses [6–9], as well as the selection and addition of treatment arms, endpoints and subgroups [2, 18–20]. The crucial point is that the adaptations need not be specified in advance in order to keep the type I error rate at the pre-specified level α. This allows to deal with the situation where a mid-trial

*Correspondence to: Werner Brannath, Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Wien, Austria.
†E-mail: werner.brannath@meduniwien.ac.at

Contract/grant sponsor: FWF; contract/grant number: P15853-N04

Received November 2004
Accepted April 2005

Copyright © 2005 John Wiley & Sons, Ltd.

inspection of the data reveals violations of the *a priori* assumptions relevant for the choice of the study design.

The basic method of adaptive tests is to use a test statistic whose null distribution is invariant with respect to design modifications. This method allows controlling the type I error rate of the significance test. Estimation is usually another important part of a clinical trial which, however, is more difficult in a flexible design. The distribution of an estimator might be adaptation invariant for a specific hypothesis (e.g. the null hypothesis) but not for all parameter configurations simultaneously, otherwise, design adaptations are unlikely to have an impact on the performance of the trial. Nevertheless, point estimators and confidence intervals have been suggested in the recent years [5, 8, 12, 21–28]. We give an overview on these methods, make some new proposals, and compare the methods numerically for typical sample size adaptations. We also consider the maximum likelihood estimate which turns out to perform fairly good in terms of the mean square error (MSE) after changing sample sizes. Therefore, we introduce new flexible confidence intervals around the maximum likelihood estimate.

Since usual fixed size sample point estimates and confidence intervals are unbiased if sample sizes remain as prefixed we can focus on sample size adaptations. We further assume that recruitment is terminated at the interim analysis in a flexible and unscheduled way. For rejecting the null at the interim analysis, however, one must prefix an interim hypothesis test, otherwise, the null must be accepted when stopping the trial. Note that our results remain relevant for adaptive interim analyses with a selection of treatment groups as termination of a treatment group means to stop recruitment for this group. For didactical reasons most sections start assuming that there is no interim stopping before dealing with the general case. For simplicity we assume that the study goal is to test $H_0: \mu = 0$ *versus* the one-sided alternative $H_1: \mu > 0$ for the mean μ of a normal response with known variance σ^2. Extensions will be discussed in Section 4.4.

2. THE PRINCIPLE OF ADAPTIVE DESIGNS

Flexible designs are based on the following general invariance principle: for the final test decision one combines stagewise statistics whose common null distribution is invariant with respect to mid-trial design modifications. We will illustrate this general principle by the so-called weighted z-score method which combines stagewise z-scores [4, 5, 29].

2.1. Weighted z-score method without interim hypothesis test

We prefix weights $w_1, w_2 \geqslant 0$ with $w_1^2 + w_2^2 = 1$ for a combination of stagewise z-scores. We further prefix the first stage sample size n_1. After recruiting n_1 patients, we compute the first stage z-score $z_1 = \sqrt{n_1}\,\bar{x}_1/\sigma$ with $\bar{x}_1$ the mean of the first stage sample. Based on this and/or any other internal or external information we choose the second stage sample size $\tilde{n}_2$ and recruit additional $\tilde{n}_2$ patients. For simplicity let us first assume that H_0 is not tested at the interim analysis and hence always $\tilde{n}_2 \geqslant 1$. Often there will be a pre-planned sample size n_2 for the second stage which, however, might be adapted at the interim analysis, so that $\tilde{n}_2 \neq n_2$ is possible. At the end of the second stage we compute $z_2 = \sqrt{\tilde{n}_2}\,\bar{x}_2/\sigma$ from the second stage sample mean $\bar{x}_2$, and we build the weighted z-score $\tilde{z} = w_1 z_1 + w_2 z_2$. A natural choice of the

Copyright © 2005 John Wiley & Sons, Ltd.

weights is $w_i = \sqrt{n_i/(n_1+n_2)}$ where n_2 is the pre-specified second stage sample size. This weighted z-score has the striking property that if no adaptation is performed $(\tilde{n}_2 = n_2)$ it is equal to the usual fixed size sample z-score $z = \sqrt{n_1+\tilde{n}_2}\,\bar{x}/\sigma = (\sqrt{n_1}\,z_1 + \sqrt{\tilde{n}_2}\,z_2)/\sqrt{n_1+\tilde{n}_2}$ where $\bar{x}$ is the overall mean (combining the first and second stage sample). However, if $\tilde{n}_2 \neq n_2$ then $\tilde{z}$ is in general not equal to z.

The null distribution of the weighted z-score $\tilde{z}$ is standard normal independently from the adaptations. To see this, notice that the conditional distribution of z_2 given z_1 and $\tilde{n}_2$ is standard normal under the null hypothesis $\mu = 0$, independently from our adaptive choice of $\tilde{n}_2 \geqslant 1$. Therefore, z_1 and z_2 are independently standard normal under the null. Since the weights w_i are fixed, we get $E_{\mu=0}(\tilde{z}) = 0$ and $\mathrm{Var}_{\mu=0}(\tilde{z}) = w_1^2 + w_2^2 = 1$. Hence, rejecting H_0 if $\tilde{z} \geqslant z_\alpha$ for the $(1-\alpha)$%-percentile z_α of the standard normal distribution gives a test for H_0 with type I error probability equal to α independently from the adaptations.

2.2. *Weighted z-score method with interim hypothesis test*

If at the interim analysis the data indicate that the treatment is ineffective or that the chance to reject H_0 at the final analysis is small even after extending sample sizes then one should stop the trial for futility and accept H_0. Such an action does not inflate the type I error rate and hence can always be done even if unspecified in the protocol.

For the possibility of rejecting the null at the interim analysis without inflating the nominal level the weighted z-score method need to be modified. This can be done by pre-specifying rejection levels $\alpha_1, \alpha_2 < \alpha$ and rejecting H_0 at the interim analysis if $z_1 \geqslant z_{\alpha_1}$ and at the final analysis if $\tilde{z} \geqslant z_{\alpha_2}$ (cf. Reference [5]). Since the covariance between z_1 and $\tilde{z}$ equals $\mathrm{Cov}_{\mu=0}(z_1, \tilde{z}) = w_1^2$ independently from the adaptations, the nominal level α is met if choosing α_1 and α_2 from a classical two stage group sequential design with interim information fraction $t_1 = w_1^2$ (cf. Reference [30]).

2.3. *Sample size assessment rules*

We have investigated the performance of estimators and confidence intervals for several different sample size assessment rules given in terms of $\tilde{r} = \tilde{n}_2/n_1$. The second stage sample size $\tilde{r}$ is assumed to be restricted by some prefixed maximum and minimum, i.e. $\tilde{r} = 0$ if stopping at the interim analysis and $0 < r_{\text{cont}} \leqslant \tilde{r} \leqslant r_{\max} < \infty$ if continuing with the second stage. We let $r_{\min}$ denote the overall minimum of $\tilde{r}$ which equals r_{cont} in a trial excluding early stopping and 0 otherwise. The preplanned second stage sample size is denoted n_2.

We will, in particular, consider the *predictive power rule*, where one uses the weighted z-score for testing H_0 and assesses $\tilde{n}_2$ for a prefixed conditional power $P_{\hat{\mu}_1}$ (reject $H_0\,|\,z_1$) of $1-\beta_c$ at the estimated alternative $\hat{\mu}_1 = \max(0, \bar{x}_1)$ (cf. References [3, 5, 10, 31–34]). The trial is stopped for futility if $z_1 \leqslant z_{\alpha_0}$ for some $\alpha_0 > \alpha$ and/or stopped with a rejection of H_0 if $z_1 \geqslant z_{\alpha_1}$ for some $\alpha_1 < \alpha$. Accounting for r_{cont} and $r_{\max}$ this gives the sample size rule

$$\tilde{r} = \begin{cases} 0 & \text{if } z_1 \leqslant z_{\alpha_0} \text{ or } z_1 \geqslant z_{\alpha_1} \\ r_{\max} & \text{if } z_{\alpha_0} \leqslant z_1 \leqslant 0 \\ \max\left[r_{\text{cont}}, \min[r_{\max}, \left(\dfrac{z_\alpha/w_2 - z_1\, w_1/w_2 + z_{\beta_c}}{\max(0, z_1)} \right)^2 \right] & \text{if } \max(0, z_{\alpha_0}) < z_1 < z_{\alpha_1} \end{cases} \tag{1}$$

For illustrative purposes we also will consider designs with $\alpha_0 = 1$ and $\alpha_1 = 0$ in which case there is no early stopping ($r_{\min} = r_{\text{cont}} > 0$).

3. POINT ESTIMATION IN FLEXIBLE DESIGNS

3.1. Maximum likelihood estimate

Since $\tilde{n}_2$ is determined at the interim analysis, $\tilde{n}_2$ is formally a stopping rule. Hence the likelihood of the data is given by $\prod_{i=1}^{\tilde{n}_2} f(x_i)$ with $f(x)$ the density of a single observation, and the maximum likelihood estimate of μ is the overall mean $\bar{x} = (n_1 \bar{x}_1 + \tilde{n}_2 \bar{x}_2)/(n_1 + \tilde{n}_2) = (\bar{x}_1 + \tilde{r}\bar{x}_2)/(1+\tilde{r})$. If sample sizes are reassessed then $\bar{x}$ can become mean biased. Liu *et al.* [24] give a simple formula for the mean bias of $\bar{x}$ by noticing that $\bar{x} = \tilde{v}\bar{x}_1 + (1-\tilde{v})\bar{x}_2$ with $\tilde{v} = 1/(1+\tilde{r})$. Since given $\tilde{n}_2$ the conditional mean of $\bar{x}_2$ is μ, $E_\mu(\bar{x}) - \mu = \text{Cov}_\mu(\tilde{v}, \bar{x}_1)$; see Appendix A.1. If, for example, the sample size decreases with increasing $\bar{x}_1$ then the covariance $\text{Cov}_\mu(\tilde{v}, \bar{x}_1)$ is positive and hence also the bias. To know the bias, we would need to know $\tilde{n}_2$ at all interim outcomes. Flexible designs aim to deal with the case where $\tilde{n}_2$ does not follow a pre-fixed rule and, obviously, the mean bias is unknown in this case. However, as shown in Appendix A.1, the absolute mean bias is always bounded by $|E_\mu(\bar{x}) - \mu| \leqslant 0.4\,(\sigma/\sqrt{n_1})\,\{(1 + r_{\min})^{-1} - (1 + r_{\max})^{-1}\}$ which is at most 40 per cent of the standard deviation of the first stage mean. The variance is another important property of an estimator and is given in Appendix A.1. From the formula in the Appendix it can be seen that it also depends on the rule for $\tilde{n}_2$ and hence it is unknown as well.

3.2. Mean unbiased estimates for designs with $r_{\min} > 0$

If the trial is always continued beyond the interim analysis, there is a simple mean unbiased estimate for μ [24, 27]. Prefixing a number $0 \leqslant u \leqslant 1$, the estimate

$$\hat{x}_u = u\bar{x}_1 + (1-u)\bar{x}_2 \tag{2}$$

is mean unbiased for μ, since $\bar{x}_1$ and $\bar{x}_2$ are unbiased and hence $E_\mu(\hat{x}_u) = u\,\mu + (1-u)\,\mu = \mu$. If $u = n_1/(n_1 + n_2)$ for the pre-planned n_2, and $\tilde{n}_2 = n_2$ is as pre-planned, then $\hat{x}_u$ equals the maximum likelihood estimate $\bar{x}$. If $\tilde{n}_2 \neq n_2$, however, $\hat{x}_u$ and $\bar{x}$ are different in general. As shown in Appendix A.1, the variance of $\hat{x}_u$ is $\text{Var}_\mu(\hat{x}_u) = (\sigma^2/n_1)\,\{u^2 + (1-u)^2 E_\mu(1/\tilde{r})\}$ and depends on the rule for $\tilde{r}$. Note that the first stage mean $\bar{x}_1 = \hat{x}_1$ $(u = 1)$ has mean μ and variance σ^2/n_1 independently form the adaptations. The mean of the first $n_{\min} = n_1\,(1 + r_{\min})$ observations, i.e.

$$\bar{x}_{\min} = \sum_{i=1}^{n_{\min}} x_i/n_{\min} \tag{3}$$

is another mean unbiased estimate with invariant variance $\sigma^2/n_{\min}$. Clearly, this estimate is not of the form (2) and is more precise than $\bar{x}_1$.

3.3. Median unbiased estimates for designs with $r_{\min} > 0$

Another suggestion is to use an estimate which has median equal to μ independently from the adaptations [8, 26, 27]. Median unbiased point estimates can be constructed from the invariance

Copyright © 2005 John Wiley & Sons, Ltd.

principle of adaptive tests. If $r_{\min} > 0$, i.e. the trial is always continued with the second stage, then

$$\hat{x}_m = \tilde{u}\bar{x}_1 + (1-\tilde{u})\bar{x}_2 \quad \text{with } \tilde{u} = \frac{w_1\sqrt{n_1}}{w_1\sqrt{n_1} + w_2\sqrt{\tilde{n}_2}} = \frac{w_1}{w_1 + w_2\sqrt{\tilde{r}}} \tag{4}$$

is median unbiased. This follows from $(\hat{x}_m - \mu)(w_1\sqrt{n_1} + w_2\sqrt{\tilde{n}_2})/\sigma = w_1\sqrt{n_1}(\bar{x}_1 - \mu)/\sigma + w_2\sqrt{\tilde{n}_2}(\bar{x}_2 - \mu)/\sigma$, where the right side is standard normal independently from the adaptations by the same reason why $\tilde{z}$ is standard normal under the null hypothesis. Cheng [28] derived (4) by the method of moments. Note that the weight $\tilde{u}$ in (4) depends on the weight w_i and the choice of $\tilde{n}_2$. Like for the maximum likelihood estimate, mean bias and variance of $\hat{x}_m$ depend on the adaptation rule and hence are unknown in general; see Appendix A.1.

3.4. Point estimation in designs with $r_{\min} = 0$

We now consider the case where $r_{\min} = 0$, i.e. the trial can be stopped at the interim analysis (in an unscheduled way). Note that the discussion of the maximum likelihood estimate in Section 3.1 covers this case. With $r_{\min} = 0$ the absolute mean bias is bounded by $0.4(\sigma/\sqrt{n_1})\{1-(1+r_{\max})^{-1}\}$ whatever sample size rule is used. It is shown in Appendix A.1 that the bias is maximized when choosing $\tilde{r} = 0$ if $\bar{x}_1 \geqslant \mu$ and $\tilde{r} = r_{\max}$ otherwise. The same bias is achieved in a two stage group sequential design with minimum and maximum sample size n_1 and $n_1(1+r_{\max})$, respectively, and interim rejection rule $\bar{x}_1 \geqslant z_{\alpha_1}\sigma/\sqrt{n_1}$ if the true mean is $\mu = z_{\alpha_1}\sigma/\sqrt{n_1}$ (which e.g. in a one-sided Pocock design [30] at level $\alpha = 0.025$ equals 1.04 times the preplanned effect size for a power of 80 per cent). Hence, the maximum mean bias is in a flexible two stage design in general not larger than in a conventional group sequential design.

Also $\hat{x}_m$ is formally defined for $\tilde{r} = 0$ and equals $\bar{x}_1$ in this case, however, will in general become median biased if the trial can be stopped at the interim analysis. Median unbiased point estimates for general two- and multi-stage flexible designs with stopping rules are given in Reference [8], however, only for trials where the stopping rule is mandatory and prefixed.

If $u < 1$ the mean unbiased estimate $\hat{x}_u$ is not defined for $\tilde{r} = 0$. Only the first stage mean $\bar{x}_1$ $(u = 1)$ is defined and mean unbiased for $r_{\min} = 0$.

3.5. Comparison of point estimates

Obviously, the estimates $\bar{x}$ (maximum likelihood), $\bar{x}_1$ (first stage mean), $\bar{x}_{\min}$ (mean of first $n_{\min}$ patients), $\hat{x}_u$ (mean unbiased estimate) and $\hat{x}_m$ (median unbiased estimate) converge to μ in probability if $n_1 \to \infty$ and $\tilde{n}_2 \to \infty$ in probability. The maximum likelihood and median unbiased estimates are consistent in an even stronger sense: they converge to μ also if any one of the stagewise sample sizes becomes infinite whereas the other remains bounded. This stronger consistency property is not shared by the mean unbiased estimates $\bar{x}_1$, $\bar{x}_{\min}$, and $\hat{x}_u$.

Figure 1 gives plots of mean bias and square root of MSE in units of $\sigma/\sqrt{n_1}$ and in dependence of the non-centrally parameter $\delta_1 = \sqrt{n_1}\,\mu/\sigma$ of the first stage z-score z_1. Such plots are invariant with respect to σ and n_1 but depend on the rule for $\tilde{r}$ and on the choice of the weights u and w_1 for $\hat{x}_u$ and $\hat{x}_m$. In Figure 1(a) we have used the predicitive power rule of Section 2.3 with $\alpha_0 = 1$ and $\alpha_1 = 0$ (no early stopping), $r_{\text{cont}} = 0.1$ and $r_{\max} = 2$. In Figure 1(b) we used $w_1^2 = u = 0.5$ and $\alpha_0 = 0.5$, $\alpha_1 = 0.0026$, $r_{\text{cont}} = 0.1$ and $r_{\max} = 5$. The interim rejection level α_1 is from a one sided O'Brien and Fleming [30] design at overall level $\alpha = 0.025$ without

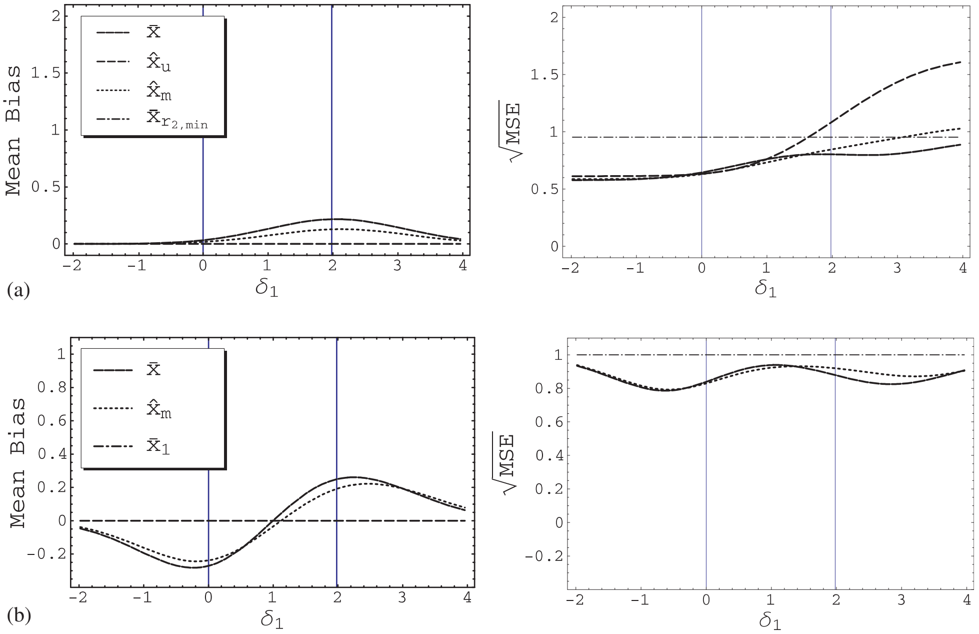

Figure 1. Mean bias and root of mean square error in units of $\sigma/\sqrt{n_1}$ and in dependence of the non-centrally parameter $\delta_1=\sqrt{n_1}\,\mu/\sigma$ of z_1 for the maximum likelihood estimate $\bar{x}$ (solid line), mean and median unbiased estimates $\hat{x}_u$ (dashed line) and $\hat{x}_m$ (dotted line) with $w_1^2=u=0.5$, when using the predictive power rule for $\tilde{r}$ with (a) $\alpha_0=1$, $\alpha_1=0$, $r_{\min}=0.1$, $r_{\max}=2$, and (b) $\alpha_0=0.5$, $\alpha_1=0.0026$, $r_{\mathrm{cont}}=0.1$, $r_{\max}=5$. The right vertical line is through $\delta_1=\sqrt{n_1}\,\mu/\sigma$ where μ/σ is the 80 per cent power alternative of the usual fixed size sample z-test with sample size $n_1/u=2n_1$. The horizontal dashed-dotted lines gives the standard deviation of the mean unbiased estimate $\bar{x}_{\min}$ using the first $n_1(1+r_{\min})$ observation.

accounting for the early acceptance boundary z_{α_0} (to keep the futility stopping optional). Since the estimate $\hat{x}_u$ is not defined for this rule it is not considered in Figure 1(b). The vertical line through $\delta_1=0$ marks the null hypothesis, the second vertical line is through $\delta_1=\sqrt{n_1}\,\mu/\sigma$ with μ/σ such that the one-sided fixed size sample z-test with sample size $n_1/u=2n_1$ has power 80 per cent. Using this δ_1 instead of the estimate $\max(0,z_1)$ in the conditional power rule (1) we have observed very similar mean biases and mean square errors.

The figures (and similar numerical investigations) indicate that the mean bias of median unbiased and maximum likelihood estimate is usually small compared to their MSE. Figure 1(a) also indicates that the mean unbiased estimate $\hat{x}_u$, although mean unbiased, may perform bad in terms of the MSE. In particular, the MSE of $\hat{x}_u$ can become much larger than the MSE of $\bar{x}_{\min}$ represented by the horizontal slashed-dotted line in Figure 1(a), or may even be larger than the MSE of $\bar{x}_1$ which is 1 in units of σ^2/n_1. As mentioned before, $\bar{x}_{\min}$ and $\bar{x}_1$ are mean unbiased and have adaptation invariant variances (equal to MSE), however, they use only part of the data. For this reason we do not recommend to use $\bar{x}_1$, $\bar{x}_{\min}$ or $\hat{x}_u$. The MSE of the median unbiased estimate can exhibit a similar adverse property which, however, is much less pronounced than for the mean unbiased estimate. It seems ignorable in our examples, in particular, in Figure 1(b) when using a stopping rule. We have also considered the weights

$w_1^2=u=0.25$ and $w_1^2=u=0.75$ (with the according conditional power rule) and have made very similar observations.

4. FLEXIBLE CONFIDENCE INTERVALS

4.1. Flexible confidence intervals centred at the median unbiased estimate

In flexible trials the classical confidence interval $\mathrm{CI}=(\bar{x}-\sigma z_\alpha/\sqrt{n_1+\tilde{n}_2},\ \bar{x}+\sigma z_\alpha/\sqrt{n_1+\tilde{n}_2})$ can have coverage probability less than $1-2\alpha$ for the same reasons why the naïve z-test for $H_0: \mu=0$ can have type I error probability larger than α. Hence CI is not a valid confidence interval in flexible designs. However, by the duality between confidence sets and hypothesis tests, the invariance principle can be used to construct flexible confidence intervals which have coverage probability of at least $1-2\alpha$ (cf. References [5, 8, 12, 22, 24–28]). In the following we illustrate this method for the weighted z-score test. Again we first focus on trials which are never stopped at the interim analysis ($r_{\min}>0$) and consider trials with $r_{\min}=0$ afterwards.

4.1.1. Trials with $r_{\min}>0$. We prefix weights $w_1, w_2 \geqslant 0$ with $w_1^2+w_2^2=1$ and exclude a parameter value μ if and only if $w_1 z_{1\mu}+w_2 z_{2\mu} \geqslant z_\alpha$ with $z_{1\mu}=\sqrt{n_1}(\bar{x}_1-\mu)/\sigma$ and $z_{2\mu}=\sqrt{\tilde{n}_2}(\bar{x}_2-\mu)/\sigma$. Since this is the rejection rule of an adaptive level α test for testing the parameter value μ, the set $\{\mu: w_1 z_{1\mu}+w_2 z_{2\mu}<z_\alpha\}$ has coverage probability $1-\alpha$ independently from the adaptations. This set can easily be inverted to the one-sided interval $(\hat{x}_m-\sigma z_\alpha(w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2})^{-1}, \infty)$. Similarly, a two-sided flexible confidence interval at level $1-2\alpha$ is defined by

$$\mathrm{FCI}_m=\left(\hat{x}_m-\frac{\sigma z_\alpha}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}},\ \hat{x}_m+\frac{\sigma z_\alpha}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}}\right) \tag{5}$$

Obviously, this confidence interval is symmetric around the median unbiased estimate $\hat{x}_m$. Furthermore, FCI_m excludes 0 if and only if the adaptive test for $H_0: \mu=0$ based on the weighted z-score $\tilde{z}$ with the same weights as in $\tilde{z}_\mu$ rejects H_0. If no sample size adaptation is performed ($\tilde{r}=r$) then FCI_m is equal to the classical confidence interval CI. One should also note that the coverage probabilities of FCI_m are exactly $1-2\alpha$ independently from the adaptations (as the type I error probability of the dual one-sided significance tests equal α).

4.1.2. Trials with $r_{\min}=0$. In Reference [5] the repeated confidence interval approach of Jennison and Turnbull [35, 36] is extended to flexible designs. Here one uses the weighted z-score method with some prefixed first and second stage rejection levels α_1 and α_2 satisfying $P_\mu(z_{1\mu} \geqslant z_{\alpha_1}$ or $\tilde{z}_\mu \geqslant z_{\alpha_2})=\alpha$ (cf. Section 2.2). One excludes μ at the interim analysis if $z_{1\mu} \geqslant z_{\alpha_1}$, and at the second stage if $\tilde{z}_\mu \geqslant z_{\alpha_2}$. Inverting these rejection rules for the first and second stage analysis, respectively, gives the sequential flexible confidence interval

$$\mathrm{SFCI}_m=\begin{cases}\left(\bar{x}_1-\dfrac{\sigma z_{\alpha_1}}{\sqrt{n_1}},\ \bar{x}_1+\dfrac{\sigma z_{\alpha_1}}{\sqrt{n_1}}\right) & \text{if } \tilde{r}=0\\[2ex] \left(\hat{x}_m-\dfrac{\sigma z_{\alpha_2}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}},\ \hat{x}_m+\dfrac{\sigma z_{\alpha_2}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}}\right) & \text{if } \tilde{r}>0\end{cases} \tag{6}$$

Copyright © 2005 John Wiley & Sons, Ltd.

which is centred at the median unbiased estimate $\hat{x}_m$ and extends (5) to trials with $r_{\min}=0$.

Since $P_\mu(\{\tilde{r}=0,\ z_{1\mu}\geqslant z_{\alpha_1}\}\cup\{\tilde{r}>0,\ \tilde{z}_\mu\geqslant z_{\alpha_2}\})\leqslant P_\mu(z_{1\mu}\geqslant z_{\alpha_1}$ or $\tilde{z}_\mu\geqslant z_{\alpha_2})=\alpha$, the interval SFCI_m has coverage probability of at least $1-2\alpha$ irrespective of the rule for $\tilde{r}$.

It has been demonstrated that (6) does not exhaust the level, i.e. has coverage probabilities larger than $1-2\alpha$ for most μ. In References [8, 12, 25] flexible confidence intervals are constructed which exhaust the level and/or uniformly improve (6) in trials with a prefixed (and mandatory) stopping rule. However, these intervals are not applicable if recruitment is stopped at the pre-scheduled interim analysis in a completely flexible way.

4.2. Flexible confidence intervals containing the classical confidence interval

We have seen in Section 3.5 that the maximum likelihood estimate $\bar{x}$ is a fairly good point estimator also in the adaptation case. Additionally, $\bar{x}$ is the posteriori mean of μ when using a flat prior. Hence, it is worthwhile to ask for confidence intervals which always have $\bar{x}$ in its interior. We next introduce such flexible confidence intervals by building intervals which contain the classical interval CI. Note that CI has the Bayesian interpretation of being the central 95 per cent range of the posterior distribution of μ (from the flat prior) also if $\tilde{n}_2$ is reassessed. Moreover, a rejection of parameter values not excluded by CI may be difficult to communicate. (Why do adaptations allow to reject parameter values not rejected in a fixed size sample test with similar sample size and mean?) The related property of an adaptive test to accept H_0 if the usual unadjusted test accepts is advocated in Denne [32] and Posch *et al.* [33]. Note that rejecting $H_0:\mu=0$ if 0 is excluded by the flexible confidence interval gives a flexible level α test for H_0. Hence, specifying a flexible confidence interval there is no need to specify a separate adaptive test for H_0.

4.2.1. Enlarging $SFCI_m$ and FCL_m to contain the classical confidence interval CI. Usually $\alpha_1<\alpha$ and SFCI_m contains the classical interval CI at the interim analysis, however, not necessarily also at the second stage. To cover CI at the second stage we could take

$$\text{SEFCI}_m=\left(\min\left[\hat{x}_m-\frac{\sigma z_{\alpha_2}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}},\ \bar{x}-\frac{\sigma z_{\alpha_2}}{\sqrt{n_1+\tilde{n}_2}}\right],\right.$$
$$\left.\max\left[\hat{x}_m+\frac{\sigma z_{\alpha_2}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}},\ \bar{x}+\frac{\sigma z_{\alpha_2}}{\sqrt{n_1+\tilde{n}_2}}\right]\right) \tag{7}$$

Clearly, the resulting interval has a flexible coverage probability of at least $1-2\alpha$ and always contains CI. For $r_{\min}>0$ the interval FCI_m could be enlarged in a similar way.

4.2.2. Flexible confidence intervals based on the sufficient test statistic. Adaptive tests based on the invariance principle have been criticized because they do not use the sufficient test statistics $(\tilde{n},\bar{x})$ where $\tilde{n}=n_1+\tilde{n}_2$ [37, 38]. Similarly, the confidence intervals (5) and (7) are not based on $(\tilde{n},\bar{x})$. Can one construct flexible confidence intervals using $(\tilde{n},\bar{x})$? Proschan and Hunsberger [3] show that if α_{ad} solves the equation $\alpha_{\text{ad}}+\exp(-z^2_{\alpha_{\text{ad}}}/2)/4=\alpha$, then $P_\mu(\sqrt{\tilde{n}}(\bar{x}-\mu)/\sigma<z_{\alpha_{\text{ad}}})\geqslant 1-\alpha$ independent from how we chose $\tilde{n}_2$ at the interim analysis. Inverting the

Copyright © 2005 John Wiley & Sons, Ltd.

inequality inside the probability gives the flexible one-sided level $1-\alpha$ confidence interval $(\bar{x}-\sigma z_{\alpha_{\mathrm{ad}}}/\sqrt{\tilde{n}},\infty)$. Similarly, a flexible two-sided $1-2\alpha$ confidence interval is

$$\mathrm{FCI}_{\mathrm{mle}}=(\bar{x}-\sigma z_{\alpha_{\mathrm{ad}}}/\sqrt{\tilde{n}},\ \bar{x}+\sigma z_{\alpha_{\mathrm{ad}}}/\sqrt{\tilde{n}}) \tag{8}$$

This interval, however, seems large compared to the classical confidence interval CI. If for example $\alpha=0.025$ then $z_{\alpha_{\mathrm{ad}}}=2.35$ compared to $z_\alpha=1.96$. By definition (8) controls the coverage probabilities for the two specific sample size rules which minimize the coverage probabilities of the lower and upper one-sided intervals, respectively. These rules are unique and hence the one-sided coverage probabilities are larger than $1-\alpha$ for all but a single sample size rule. Moreover, the rule which minimizes the upper coverage probability does not minimize the lower coverage probability and *vice versa*. Hence, the two-sided coverage probability (the probability that the two-sided interval covers μ) is always larger than $1-2\alpha$.

If $\tilde{r}$ is constraint, e.g. $r_{\mathrm{cont}}\leqslant\tilde{r}\leqslant r_{\max}$ at the second stage for some prefixed $0<r_{\mathrm{cont}}<r_{\max}$, then the maximum type I error inflation of the classical unadjusted test is smaller than without constraints. Hence, the adjusted level is smaller than for the unrestricted case. This allows for a smaller flexible confidence interval. Since in practice sample sizes are always constraint we further on refer to $\mathrm{FCI}_{\mathrm{mle}}$ as (8) with α_{ad} accounting for the constraints.

But how to determine α_{ad} with constraints on $\tilde{r}$? We show in Appendix A.2 that the maximum type I error rate of the rejection rule '$z_\mu:=\sqrt{\tilde{n}}(\bar{x}-\mu)/\sigma\geqslant c$' equals the probability of '$Z_{\max}(Z_1,Z_2)\geqslant c$' for independent standard normally distributed Z_1,Z_2 whereby

$$Z_{\max}(z_1,z_2):=\max_{j=1,\ldots,m}\frac{z_1+\sqrt{r_j}\,z_2}{\sqrt{1+r_j}} \tag{9}$$

with r_j $(j=1,\ldots,m)$ the possible values of $\tilde{r}$, and z_1, z_2 are assumed fixed and known when maximizing over r_j in (9). Note that by taking the maximum over the prefixed set of possible values of $\tilde{r}$, (9) does not depend on $\tilde{r}$ and is a monotone function of the stagewise z-scores.

Since $z_\mu=(z_{1\mu}+\sqrt{\tilde{r}}z_{2\mu})/\sqrt{1+\tilde{r}}\leqslant Z_{\max}(z_{1\mu},z_{2\mu})$ by definition (9), the probability of '$Z_{\max}(Z_1,Z_2)\geqslant c$' is an upper bound for the probability of '$z_\mu\geqslant c$'. In Theorem A.1 of Appendix A.2 we show that the probability of '$z_\mu\geqslant c$' can be as large as $P[Z_{\max}(Z_1,Z_2)\geqslant c]$ when choosing $\tilde{r}$ based on the interim data.

Knowing $Z_{\max}(z_1,z_2)$ one can determine α_{ad} such that $P[Z_{\max}(Z_1,Z_2)\geqslant z_{\alpha_{\mathrm{ad}}}]=\alpha$ by numeric integration and root finding. If using this α_{ad} then (8) has minimal coverage probability $1-2\alpha$. The determination of $Z_{\max}(z_1,z_2)$ is discussed in Appendix A.3.

4.2.3. Uniform improvement of FCI_{mle}. It is interesting to note that the flexible likelihood based confidence interval $(\bar{x}-\sigma z_{\alpha_{\mathrm{ad}}}/\sqrt{\tilde{n}},\ \infty)$ can be uniformly improved by inverting '$Z_{\max}(z_{1\mu},z_{2\mu})<z_{\alpha_{\mathrm{ad}}}$': since at every sample point $Z_{\max}(z_{1\mu},z_{2\mu})\geqslant\sqrt{\tilde{n}}(\bar{x}-\mu)/\sigma$ for all μ, inverting '$Z_{\max}(z_{1\mu},z_{2\mu})<z_{\alpha_{\mathrm{ad}}}$' always leads to a smaller interval than inverting '$\sqrt{\tilde{n}}(\bar{x}-\mu)/\sigma<z_{\alpha_{\mathrm{ad}}}$'. By the definition of $z_{\alpha_{\mathrm{ad}}}$ the critical region $\{Z_{\max}(z_{1\mu},z_{2\mu})\geqslant z_{\alpha_{\mathrm{ad}}}\}$ also has type I error α. The two sided interval $\mathrm{FCI}_{\mathrm{mle}}$ can be improved in a similar way. However, the improved intervals are not based on the maximum likelihood statistics $(\tilde{n},\bar{x})$.

4.3. Numerical comparison of flexible confidence intervals

The ratio of the length of the interval FCI_m centred at the median unbiased estimate and the interval $\mathrm{FCI}_{\mathrm{mle}}$ centred at the maximum likelihood estimate depends only on α, α_{ad} and $\tilde{r}$.

Figure 2(a) gives a plot of this ratio if $w_1^2 = 0.5$, for different preplanned $r_{\min} > 0$ (assuming that the trial is always continued with the second stage) and $r_{\max}$ with corresponding α_{ad} when using the (conservative) approximation for $Z_{\max}(z_1, z_2)$ given in Appendix A.3. Each curve is plotted for $\tilde{r}$ between $r_{\min}$ and $r_{\max}$ only. Note that the ratio is maximal for $\tilde{r} = w_1^2/w_2^2 = 1$ in which case median unbiased and maximum likelihood estimate are identical and $\text{FCI}_m = \text{CI}$ is smaller than FCI_{mle}. Note also that FCI_m remains shorter than FCI_{mle} for a wide range of $\tilde{r}$, i.e. the likelihood ratio test statistics leads to a larger confidence interval for most choices of $\tilde{r}$ and becomes smaller only for extreme $\tilde{r}$. Hence, the price paid for the potential of adaptations becomes higher with the likelihood ratio test statistic than with the weighted z-score test.

Figure 2(b) gives the ratio of the length of FCI_{mle} and SFCI_m if $w_1^2 = 0.5$, $\alpha_0 = 0.5$ and $\alpha_1 = \alpha_2 = 0.0147$ is according to a Pocock design at level $\alpha = 0.025$ (not accounting for the futility boundary), and α_{ad} is such that FCI_{mle} has coverage probability $1 - 2\alpha$ in a trial which can either be stopped at the interim analysis ($\tilde{r} = 0$) or continued with $r_{\text{cont}} \leqslant \tilde{r} \leqslant r_{\max}$. Again the interval SFCI_m based on the invariance principle is shorter than the likelihood ratio based confidence interval FCI_{mle} for $\tilde{r}$ close to 1 and is longer for small and large $\tilde{r}$. Which method to use depends on the strength and frequency of the sample size reassessments: if sample size adaptations are seldom and moderate then SFCI_m will be shorter. In designs with extensive sample size adaptations FCI_{mle} will be shorter more frequently. However, one should notice that SFCI_m is shorter whenever stopping the trial at the interim analysis. This is seen from the dots in Figure 1(b). Similar results are observed when using the weight $w_1^2 = 0.25$ or 0.75 for FCI_m and SFCI_m.

We have also compared the extended confidence interval SEFCI_m with FCI_{mle} for the predictive power rule (1) with $\alpha_0 = 0.5$ and $\alpha_1 = 0.0147$ as for SEFCI_m in a simulation study with 10^6 runs per scenario. Table I gives the average length of SFCI_m, SEFCI_m and FCI_{mle} as well as the probabilities that SEFCI_m is smaller than FCI_{mle}. For all $r_{\text{cont}}, r_{\max}$ considered in the table the enlarged interval SEFCI_m was in average smaller than FCI_{mle} under the null

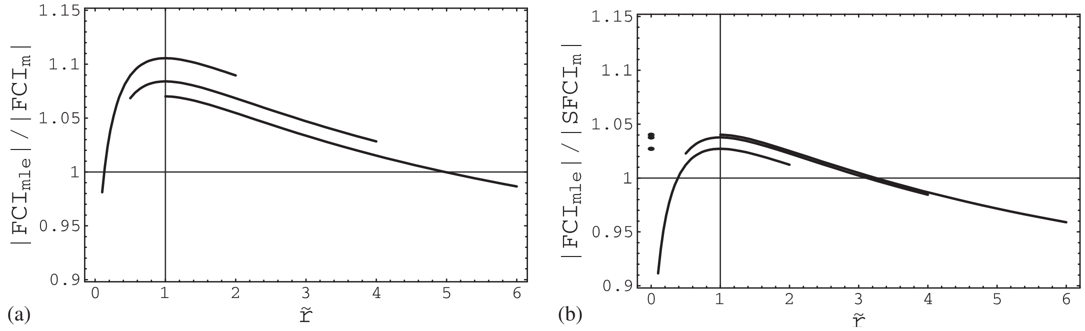

Figure 2. (a) The ratio $|\text{FCI}_{\text{mle}}|/|\text{FCI}_m|$ of the length of the interval FCI_m around the median unbiased estimate and FCI_{mle} around the maximum likelihood estimate in dependence of $\tilde{r}$ if $\alpha = 0.025$ and $w_1^2 = 0.5$. The three curves are for $r_{\min} = 0.1, 0.5, 1$ and $r_{\max} = 2, 4, 6$, respectively, and with the according α_{ad}. Each curve is plotted between $r_{\min}$ and $r_{\max}$. (b) The ratio $|\text{FCI}_{\text{mle}}|/|\text{SFCI}_m|$ with $w_1^2 = 0.5$, $\alpha_1 = \alpha_2$ and α_{ad} such that SFCI_m and FCI_{mle} have coverage probability 0.95. In the determination of α_{ad} for the three curves, $\tilde{r}$ is assumed to be either 0 or between the minimum r_{cont} and maximum $r_{\max}$ which were prefixed for the case of a continuation, with $r_{\text{cont}} = 0.1, 0.5, 1$ and $r_{\max} = 2, 4, 6$, respectively. For $\tilde{r} = 0$ the ratio $|\text{FCI}_{\text{mle}}|/|\text{SFCI}_m|$ is indicated by a black dot; it equals the respective maximum ratio.

Copyright © 2005 John Wiley & Sons, Ltd.

Table I. Comparison of the length of 95 per cent sequential flexible confidence intervals containing the classical confidence interval when using the predictive power rule with $w_1^2 = 0.5$, different r_{cont}, $r_{\max}$, and when stopping the trial ($\tilde{r} = 0$) if $z_1 \leqslant z_{\alpha_0}$ or $z_1 \geqslant z_{\alpha_1}$ with $\alpha_1 = 0.0147$. The table summarizes expected length and probabilities that SEFCI$_m$ with $\alpha_0 = 0.5$ and $\alpha_1 = \alpha_2 = 0.0147$ is shorter than FCI$_{\text{mle}}$ under the null hypothesis $\delta_1 = 0$ and the alternative $\delta_1 = \sqrt{n_1}\,\mu/\sigma$ under which a single stage trial with $2n_1$ observations would have 80 per cent power. Results are from a simulation study with 10^6 runs.

			Expected length of			Probability of
r_{cont}	$r_{\max}$	δ_1	SFCI$_m$	SEFCI$_m$	FCI$_{\text{mle}}$	length SEFCI$_m$ < length FCI$_{\text{mle}}$
0.1	2	0.00	3.51	3.51	3.58	0.99
0.1	2	1.98	3.51	3.51	3.59	0.97
0.5	4	0.00	3.30	3.30	3.38	0.59
0.5	4	1.98	3.39	3.40	3.49	0.77
1	6	0.00	3.20	3.20	3.27	0.59
1	6	1.98	3.30	3.31	3.39	0.78

hypothesis and the alternative $\delta_1 = \sqrt{n_1}\,\mu/\sigma$ for which a single stage trial with sample size $2\,n_1$ has power 80 per cent. Note that in average the extended interval SEFCI$_m$ is only slightly larger than SFCI$_m$. We observed similar results with $\alpha_0 = 1$ and $\alpha_1 = 0$.

4.4. Extensions

For simplicity we have focused on the mean value of a normal response with known variance. The flexible confidence intervals considered in Section 4.1, however, can be easily generalized to the mean difference of two normal responses (e.g. of a treatment and a placebo group) with common known variance and balanced group sizes at both sequential stages. In this case one need to replace in the formula for the estimate (4) and in the formulas of Section 4.1 the stagewise means by the stagewise mean differences, as well as n_1 and $\tilde{n}_2$ by $n_1/2$ and $\tilde{n}_2/2$, respectively. The modification of (4) is again median unbiased if the trial is never stopped at the interim analysis. The confidence intervals of Section 4.2 can be modified in a similar way. For instance, the likelihood based interval (8) with $\bar{x}$ replaced by the overall mean difference and $\tilde{n}$ replaced by $\tilde{n}/2$ is a flexible confidence interval for the mean difference. For an unbalanced design and/or an unknown variance there is no exact generalization of (8).

The flexible confidence intervals in Section 4.1 can also be generalized to parameters Δ for which independent stagewise p-values $p_{i\Delta}$ are available (either exactly or at least asymptotically). Such p-values exist, e.g. for the mean difference Δ of two normal responses with a common unknown variance and arbitrary group sizes, or for a rate or the difference of two rates. Using the 'inverse normal method' [5] we build stagewise z-scores $z_{i\Delta} = z_{p_{i\Delta}}$ by applying the inverse standard normal distribution function to $1 - p_{i\Delta}$, and use the weighted z-scores $\tilde{z}_\Delta = w_1\,z_{1\Delta} + w_2\,z_{2\Delta}$ for the dual flexible tests. The method can also be extended to flexible multi-stage designs [8]. Clearly, such intervals could always be enlarged to cover classical (unadjusted) point estimates and confidence intervals.

4.5. Numerical example

We illustrate flexible point estimates and confidence intervals with the flexible multi-centre randomized placebo-controlled trial reported in Zeymer *et al.* [39]. The primary efficacy end point was infarct size measured by the cumulative release of α-HDBH within 72 h after administration of the drug (area under the curve, α-HDBH AUC). Four dose and a placebo group were investigated at the interim analysis, where it was decided that two dose groups (and the placebo) we carried over to the second stage. The second stage sample size were chosen such that the conditional power is at least 90 per cent for a treatment difference of 0.25 times the standard deviation. The trial did not succeed in showing that the drug is superior to placebo.

We estimate the mean difference of α-HDBH AUC between the smaller selected dose and the placebo group. For simplicity we assume that the sample sizes of both treatment groups were equal to $n_1 = 88$ at the first stage and $n_2 = 322$ at the second stage. (The actual numbers in the treatment group were 91 at the first and 321 at second stage). Hence, the second stage sample size was $\tilde{r} = 3.66$ times the first stage size. The mean treatment difference (active treatment minus placebo) was -4.0 at the first stage and 1.8 at the second stage. This gives an overall mean difference of 0.6. Assuming that $w_1 = \sqrt{0.5}$ was prefixed before the study, the median unbiased estimate of the treatment difference equals -0.2. Let us assume for simplicity that the standard deviation is equal to the mean of the pooled estimate from the placebo and treatment group which is $\sigma = 26.7$ (assuming equal variances and ignoring the fact that σ is estimated). In this case the classical 95 per cent confidence interval becomes $\mathrm{CI} = (-3.1, 4.2)$. The trial could (and probably would) have been stopped if at the interim analysis the p-value of the linear trend test for no dose relationship were below a prefixed boundary. Hence, exact flexible confidence intervals must account for the possibility of $\tilde{r} = 0$. We give $\mathrm{FCI}_{\mathrm{mle}}$ assuming that either $\tilde{r} = 0$ or $r_{\mathrm{cont}} = 1 \leqslant \tilde{r} \leqslant r_{\max} = 6$ (leading to $\alpha_{\mathrm{ad}} = 0.0117$), and SFCI_m, SEFCI_m with $w_1^2 = 0.5$ and Pocock type boundaries $\alpha_1 = \alpha_2 = 0.0147$ for $\alpha = 0.025$. Clearly, in practice one must report only one interval and must fix *a priori* which interval will be used. Here we get $\mathrm{FCI}_{\mathrm{mle}} = (-3.6, 4.8)$, $\mathrm{SFCI}_m = (-4.4, 4.1)$, and $\mathrm{SEFCI}_m = (-4.4, 4.2)$. As indicated in Figure 2(b) for $\tilde{r} = 3.66$, the intervals $\mathrm{FCI}_{\mathrm{mle}}$ and SFCI_m have about the same length.

5. DISCUSSION AND EXTENSIONS

We have discussed the problem of parameter estimation in flexible designs by comparing different point estimates and confidence intervals. We have seen that for typical sample size rules the class of flexible mean unbiased point estimates can lead to unreasonable large mean square errors and hence should not be used in practice. The usual maximum likelihood estimate is mean biased, however, performs well in terms of MSE and hence remains useful for flexible trials. The median unbiased estimate performs well as long as sample size adaptations are not too extreme. Its performance is much improved by imposing stopping rules. It has the additional advantage to be the midpoint of a flexible confidence interval which is consistent with the test decision of a weighted z-score test.

Flexible confidence intervals can be constructed from the invariance principle, e.g. using the weighted z-score test. Another approach is to use the likelihood ratio test statistics and to adjust the critical boundary for potential adaptations. We have seen that the confidence interval from the weighted z-score test is usually smaller than the confidence interval from the adjusted likelihood ratio test, except for extreme sample size reassessments. If one must account for the possibility of terminating the trial at the interim analysis the advantage of

Copyright © 2005 John Wiley & Sons, Ltd.

the invariance principle based intervals can become less pronounced, and the likelihood ratio based intervals can be shorter already for moderate sample size adjustments. If stopping at the interim analysis, however, the likelihood ratio based intervals are larger.

Any flexible confidence interval can be forced to include the maximum likelihood estimate or even the classical (unadjusted) confidence interval by simply enlarging the interval. Simulation results indicate that enlarging the confidence interval from the weighted z-score test gives an interval which is in average smaller than the likelihood ratio based flexible confidence interval. The reason for the wider likelihood ratio based intervals is that a worst case adjustment of the critical boundary over all possible sample size reassessment rules is performed. Hence, the enlarged weighted z-score based confidence intervals seem to be the more useful option in practice, in particular, if sample size adjustments are expected to be rare and/or moderate.

We finally note that our conclusions are based on the investigation of a limited number of examples. In practice, the planning stage of an adaptive design should involve similar investigations of point and interval estimates for the most probable adaptations.

APPENDIX A

A.1. Mean bias, MSE and variance of estimators

The estimates considered in Section 3 are all of the form $\hat{x} = \tilde{u}\bar{x}_1 + (1-\tilde{u})\bar{x}_2$ where $0 \leqslant \tilde{u} \leqslant 1$ is either a constant (for the mean unbiased estimate) or depends on $\tilde{n}_2$. A formula for the mean bias can be obtained from the fact that $E_\mu[\bar{x}_2 - \mu|\tilde{n}_2] = 0$ for all $\tilde{n}_2$, which gives $E_\mu[\hat{x} - \mu] = E_\mu[\tilde{u}(\bar{x}_1 - \mu)] + E_\mu[(1-\tilde{u})E_\mu(\bar{x}_2 - \mu \mid \tilde{n}_2)] = E_\mu[\tilde{u}(\bar{x}_1 - \mu)] = \mathrm{Cov}_\mu(\tilde{u}, \bar{x}_1)$.

To obtain the bound on the absolute mean bias we maximize and minimize the conditional bias $E_\mu(\bar{x} - \mu|\bar{x}_1, \tilde{r}) = (\bar{x}_1 - \mu)/(1 + \tilde{r})$ by choosing $\tilde{r}$ based on the interim data. Clearly, maximizing (minimizing) the conditional bias maximizes (minimizes) the overall bias. Obviously, the conditional mean bias is maximized if $\tilde{r} = r_{\min}$ for $\bar{x}_1 \geqslant \mu$ and $\tilde{r} = r_{\max}$ for $\bar{x}_1 < \mu$ which gives $E_\mu(\bar{x} - \mu|\bar{x}_1, \tilde{r}) = \max(0, \bar{x}_1 - \mu)/(1 + r_{\min}) - \max(0, \mu - \bar{x}_1)/(1 + r_{\max})$. Since $E_\mu[\max(0, \bar{x}_1 - \mu)] = E_\mu[\max(0, \mu - \bar{x}_1)] = 0.4\,\sigma/\sqrt{n_1}$ the overall bias becomes $E_\mu[\bar{x}_1 - \mu] = 0.4\,(\sigma/\sqrt{n_1})\{(1 + r_{\min})^{-1} - (1 + r_{\max})^{-1}\}$. It can be shown in a similar manner that the minimum bias is $-0.4\,(\sigma/\sqrt{n_1})\{(1 + r_{\min})^{-1} - (1 + r_{\max})^{-1}\}$ and is achieved by $\tilde{r} = r_{\max}$ if $\bar{x}_1 \geqslant \mu$ and $\tilde{r} = r_{\min}$ otherwise.

The MSE equals $E_\mu[(\hat{x} - \mu)^2] = E_\mu[\tilde{u}^2(\bar{x}_1 - \mu)^2] + \sigma^2 E_\mu[(1 - \tilde{u})^2/\tilde{n}_2]$, because $E_\mu[\bar{x}_2 - \mu|\bar{x}_1, \tilde{n}_2] = 0$ and $E_\mu[(\bar{x}_2 - \mu)^2|\bar{x}_1, \tilde{n}_2] = \sigma^2/\tilde{n}_2$ for all $\tilde{n}_2$. Consequently, the MSE of $\hat{x}$ relative to the MSE of the first stage mean is

$$E_\mu[(\hat{x} - \mu)^2]/(\sigma^2/n_1) = E_\mu[\tilde{u}^2(z_1 - \sqrt{n_1}\,\mu/\sigma)^2] + E_\mu[(1 - \tilde{u})^2/\tilde{r}] \tag{A1}$$

For the maximum likelihood estimate (A1) becomes $E_\mu[(z_1 - \sqrt{n_1}\,\mu/\sigma)^2/(1+\tilde{r})^2] + E_\mu[\tilde{r}/(1+\tilde{r})^2]$, for the median unbiased estimate $w_1^2 E_\mu[(z_1 - \sqrt{n_1}\,\mu/\sigma)^2/(w_1 + w_2\sqrt{\tilde{r}})^2] + w_2^2 E_\mu[\tilde{r}/(w_1 + w_2\sqrt{\tilde{r}})^2]$. For the mean unbiased estimates MSE and variance are the same and equal $(\sigma^2/n_1)\{u^2 + (1-u)^2 E_\mu(1/\tilde{r})\}$.

In general, the variance of $\hat{x}$ is $\mathrm{Var}_\mu(\hat{x}) = E_\mu[(\hat{x} - \mu)^2] - E_\mu[\hat{x} - \mu]^2 = E_\mu[\tilde{u}^2(\bar{x}_1 - \mu)^2] + \sigma^2 E_\mu[(1 - \tilde{u})^2/\tilde{n}_2] - E_\mu[\tilde{u}(\bar{x}_1 - \mu)]^2 = \mathrm{Var}_\mu[\tilde{u}(\bar{x}_1 - \mu)] + (\sigma^2/n_1)E_\mu[(1 - \tilde{u})^2/\tilde{r}]$.

Copyright © 2005 John Wiley & Sons, Ltd.

A.2. Maximum type I error probability with constraints on the sample size

Theorem A.1

Let $Z_{\max}(\cdot,\cdot)$ be as defined in (9). Choosing at the interim analysis $\tilde{r}$ from the values r_j considered in (9) based on the unblinded interim data, the maximum type I error probability of the z-test $z \geqslant z_{\alpha_{\text{ad}}}$ for $H_0: \mu = 0$ equals $P[Z_{\max}(Z_1, Z_2) \geqslant z_{\alpha_{\text{ad}}}]$ for independent and standard normally distributed Z_1, Z_2.

Proof

By definition, $Z_{\max}(z_1, z_2) \geqslant z = (z_1 + \sqrt{\tilde{r}}\, z_2)/\sqrt{1+\tilde{r}}$ whatever $\tilde{r}$ is chosen at the interim analysis. Hence, $P[Z_{\max}(Z_1, Z_2) \geqslant z_{\alpha_{\text{ad}}}]$ is an upper bound for the maximum type I error probability. Next we verify that with a sample size recalculation at the interim analysis the test '$z \geqslant z_{\alpha_{\text{ad}}}$' can have type I error probability equal to $P[Z_{\max}(Z_1, Z_2) \geqslant z_{\alpha_{\text{ad}}}]$. To this end we show in the next paragraph how one can choose $\tilde{r} = \tilde{r}^{(z_1)}$ from the information on z_1 such that '$Z_{\max}(z_1, z_2) \geqslant z_{\alpha_{\text{ad}}}$' implies '$(z_1 + \sqrt{\tilde{r}^{(z_1)}}\, z_2)/\sqrt{1+\tilde{r}^{(z_1)}} \geqslant z_{\alpha_{\text{ad}}}$' for any z_2.

Let $z_2^{(z_1)} = \inf\{z_2 : Z_{\max}(z_1, z_2) \geqslant z_{\alpha_{\text{ad}}}\}$ which is a function of z_1, and notice that $Z_{\max}(z_1, z_2^{(z_1)}) = z_{\alpha_{\text{ad}}}$ by continuity of $Z_{\max}(z_1, z_2)$ in z_2. Continuity of $Z_{\max}(z_1, z_2)$ follows from the continuity of $(z_1 + \sqrt{r_j}\, z_2)/\sqrt{1+r_j}$ and from taking the maximum over finitely many r_j. Now choose $\tilde{r}^{(z_1)}$ such that $(z_1 + \sqrt{\tilde{r}^{(z_1)}}\, z_2^{(z_1)})/\sqrt{1+\tilde{r}^{(z_1)}} = Z_{\max}(z_1, z_2^{(z_1)})$. Obviously, $\tilde{r}^{(z_1)}$ is a function of z_1. Clearly, '$Z_{\max}(z_1, z_2) \geqslant z_{\alpha_{\text{ad}}}$' implies '$z_2 \geqslant z_2^{(z_1)}$' which in turn implies

$$(z_1 + \sqrt{\tilde{r}^{(z_1)}}\, z_2)/\sqrt{1+\tilde{r}^{(z_1)}} \geqslant (z_1 + \sqrt{\tilde{r}^{(z_1)}}\, z_2^{(z_1)})/\sqrt{1+\tilde{r}^{(z_1)}} = Z_{\max}(z_1, z_2^{(z_1)}) = z_{\alpha_{\text{ad}}} \qquad \square$$

A.3. On the determination of $Z_{\max}(z_1, z_2)$

If there are only few possible values for $\tilde{r}$ then $Z_{\max}(z_1, z_2)$ can be obtained by determining the maximum (9) directly. It is helpful to notice that for $\alpha < 0.5$ one gets $z_{\alpha_{\text{ad}}} > 0$ and hence one can replace $Z_{\max}(z_1, z_2)$ by $Z^{+}_{\max}(z_1, z_2) := \max[0, Z_{\max}(z_1, z_2)]$ in the determination of α_{ad}. The latter combination function is easier to compute, for instance, $Z^{+}_{\max}(z_1, z_2) = 0$ whenever $z_1 < 0$ and $z_2 < 0$.

For a large number of r_j between $r_{\min}$ and $r_{\max}$ we can approximate $Z^{+}_{\max}(z_1, z_2)$ conservatively by taking the maximum with respect to all real values between $r_{\min}$ and $r_{\max}$. We show next that this approximation gives

$$Z^{+}_{\max}(z_1, z_2) \approx \begin{cases} \sqrt{z_1^2 + z_2^2} & z_1, z_2 > 0 \text{ and } \sqrt{r_{\min}}\, z_1 \leqslant z_2 \leqslant \sqrt{r_{\max}}\, z_1 \\ \max(0, \tilde{z}^{(2,\min)}, \tilde{z}^{(2,\max)}) & \text{otherwise} \end{cases} \tag{A2}$$

with $\tilde{z}^{(mxx)} = (z_1 + \sqrt{r_{mxx}}\, z_2)/\sqrt{1+r_{mxx}}$ for $mxx \in \{\min, \max\}$. If $\tilde{r}$ either 0 or $r_{\text{cont}} \leqslant \tilde{r} \leqslant r_{\max}$ then $Z^{+}_{\max}(z_1, z_2)$ is the maximum between $\max(0, z_1)$ and (A2) with $r_{\min}$ replaced by r_{cont} and $\tilde{z}^{(\min)}$ replaced by $\tilde{z}^{(\text{cont})} = (z_1 + \sqrt{r_{\text{cont}}}\, z_2)/\sqrt{1+r_{\text{cont}}}$.

To verify (A2) we can restrict attention to z_1, z_2 where either $z_1 \geqslant 0$ or $z_2 \geqslant 0$. We consider the derivative $(\mathrm{d}/\mathrm{d}r)(z_1 + \sqrt{r}\, z_2)/\sqrt{1+r}$ which has the same sign as $s_r := z_2 - \sqrt{r}\, z_1$ for $r > 0$. If $z_2 < 0$ and $z_1 \geqslant 0$ then s_r is negative for all $r > 0$ and hence $Z_{\max}(z_1, z_2) = (z_1 + \sqrt{r_{\min}}\, z_2)/\sqrt{1+r_{\min}}$. If $z_1 \leqslant 0$ and $z_2 > 0$ then s_r is positive for all $r > 0$ and so $Z_{\max}(z_1, z_2) = (z_1 + \sqrt{r_{\max}}\, z_2)/\sqrt{1+r_{\max}}$. If $z_2 = 0$ then $Z_{\max}(z_1, 0)$ equals either $\tilde{z}^{(\min)}$ or $\tilde{z}^{(\max)}$ depending on the sign of z_1. If $z_1 > 0$ and $z_2 > 0$ then $(z_1 + \sqrt{r}\, z_2)/\sqrt{1+r}$ has the unique maximizer $r = z_2^2/z_1^2$, the solution of $z_2 - \sqrt{r}\,z_1 = 0$. If $z_2^2/z_1^2 < r_{\min}$ then $(z_1 + \sqrt{r}\,z_2)/\sqrt{1+r}$ is decreasing for all $r \geqslant r_{\min}$ and hence is maximal for $r = r_{\min}$. If $z_2^2/z_1^2 > r_{\max}$ then $(z_1 + \sqrt{r}\,z_2)/\sqrt{1+r}$ is increasing for all $r \leqslant r_{\max}$ and hence is maximal for $r = r_{\max}$. If $r_{\min} \leqslant z_2^2/z_1^2 \leqslant r_{\max}$ then $r = z_2^2/z_1^2$ gives $Z_{\max} = \sqrt{z_1^2 + z_2^2}$.

Note that setting $r_{\min} = 0$ and letting $r_{\max} \to \infty$ we get $\tilde{z}_{\min} = z_1$ and $\tilde{z}_{\max} = z_2$ and a limiting combination function (9) which is closely related to the circular conditional error function of Proschan and Hunsberger [3].

ACKNOWLEDGEMENTS

The authors would like to thank Martin Posch and two anonymous referees for their helpful comments.

REFERENCES

1. Bauer P. Multistage testing with adaptive designs (with Discussion). Biometrie und Informatik in Medizin und Biologie 1989; **20**:130–148.
2. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
3. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
4. Cui L, Hung HMJ, Wang S. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:321–324.
5. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.
6. Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 1999; **55**:190–197.
7. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
8. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
9. Hartung J, Knapp G. A new class of completely self-designing clinical trials. *Biometrical Journal* 2003; **45**:3–19.
10. Bauer P, König F. The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine*, 2005, in press.
11. Wang SJ, Hung HMJ, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2001; **20**:1903–1912.
12. Brannath W, Maurer W, Posch M, Bauer P. Sequential tests for non-inferiority and superiority. *Biometrics* 2003; **59**:106–114.
13. Lang T, Auterith A, Bauer P. Trend tests with adaptive scoring. *Biometrical Journal* 2000; **42**:1007–1020.
14. Neuhäuser M. An adaptive location-scale test. *Biometrical Journal* 2001; **43**:809–819.
15. Lawrence J. Strategies for changing the test statistics during a clinical trial. *Journal of Biopharmaceutical Statistics* 2002; **12**:193–205.
16. Kieser M, Schneider B, Friede T. A bootstrap procedure for adaptive selection of the test statistic in flexible two-stage designs. *Biometrical Journal* 2002; **44**:641–652.
17. Friede T, Kieser M, Neuhäuser M, Büning H. A comparison of procedures for adaptive choice of location tests in flexible two-stage designs. *Biometrical Journal* 2003; **45**:292–310.
18. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
19. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**:261–277.
20. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**: 581–589.
21. Coburger S, Wassmer G. Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal* 2001; **43**:821–833.
22. Liu Q, Chi GYH. On sample size and inference for two-stage adaptive designs. *Biometrics* 2001; **57**:172–177.
23. Coburger S, Wassmer G. Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biometrical Journal* 2003; **45**:812–825.
24. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 2002; **97**:1034–1041.

Copyright © 2005 John Wiley & Sons, Ltd.

25. Brannath W, König F, Bauer P. Improved repeated confidence bounds in trials with a maximal goal. *Biometrical Journal* 2003; **45**:311–324.
26. Lawrence J, Hung HMJ. Estimation and confidence intervals after adjusting the maximum information. *Biometrical Journal* 2003; **45**:143–152.
27. Proschan MA, Liu QL, Hunsberger S. Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials* 2003; **24**:4–15.
28. Cheng Y, Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* 2004; **60**:910–918.
29. Banik N, Köhne K, Bauer P. On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical Journal* 1996; **38**:25–37.
30. Jennison C, Turnbull BW. *Group Sequential Tests with Applications to Clinical Trials*. Chapman & Hall/CRC: London/Boca Raton, FL, 2000.
31. Posch M, Bauer P. Interim analysis and sample size reassessment. *Biometrics* 2000; **56**:1170–1176.
32. Denne JD. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; **20**:2645–2660.
33. Posch M, Bauer P, Brannath W. Issues on flexible designs. *Statistics in Medicine* 2003; **22**:953–969.
34. Brannath W, Bauer P. Optimal conditional error functions for the control of conditional power. *Biometrics* 2004; **60**:715–723.
35. Jennison C, Turnbull BW. Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* 1984; **5**:33–45.
36. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1989; **51**:305–361.
37. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials. *Statistics in Medicine* 2003; **22**:971–993.
38. Tsiatis AA, Metha C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
39. Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis G, Linssen G, U T, Schröder R, Tiemannn R, Machnig T, Neuhaus KL. The Na^+/H^+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. *Journal of the American College of Cardiology* 2001; **38**:1644–1651.

Copyright © 2005 John Wiley & Sons, Ltd.