A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials

J. Bowden * and A. Mander

In this paper, we review the adaptive design methodology of Li et al. (Biostatistics 3:277–287) for two-stage trials with mid-trial sample size adjustment. We argue that it is closer in principle to a group sequential design, in spite of its obvious adaptive element. Several extensions are proposed that aim to make it even more attractive and transparent alternative to a standard (fixed sample size) trial for funding bodies to consider. These enable a cap to be put on the maximum sample size and for the trial data to be analysed using standard methods at its conclusion. The regulatory view of trials incorporating unblinded sample size re-estimation is also discussed. © 2014 The Authors. Pharmaceutical Statistics published by John Wiley & Sons Ltd.

Keywords: sample size re-estimation; two-stage trial; conditional power; median unbiased estimation

1. INTRODUCTION

When designing a randomised controlled trial (RCT) to test the efficacy of a treatment in a chosen patient population, assumptions need to be made about the mean and spread of patient responses to treatment in order to derive an appropriate sample size. However, these assumptions may be subject to considerable uncertainty and, if their validity is not subsequently checked, could lead to a hopelessly underpowered or overpowered study. Adaptive designs incorporating sample size re-estimation offer a potential solution to this problem, by enabling interim patient data to be used to decide whether the initial assumptions were sensible and, if necessary, to alter the size and scope of the trial. Methods to update a trial’s required sample size using the current estimate of the pooled response’s standard deviation are well used and accepted by the regulatory authorities [1], because this does not require unblinding of the treatment and control groups. See for example Gould and Shih [2] and Kieser and Friede [3]. Conversely, there has been a poor uptake of methods that allow unblinding to explicitly estimate the difference in response levels across groups, that is, the treatment effect. This is due in part to fundamental concerns over the trials’ perceived validity or scientific rigour after unblinding has occurred. However, objections of a more theoretical nature have also been raised. For example, common methods proposed in this context such as p-value combination or variance spending approaches [4,5] can assign unequal weight to patients before and after the sample size re-estimation (SSR). This violates the sufficiency principle and is criticised for being inefficient, compared with more established group-sequential methods [6,7]. Moreover, the statistical complexity of many methods, and their use of abstract conditional error functions [8] with non-standard critical thresholds, may also serve to discourage their application in real clinical settings.

In this paper, we review a two-stage adaptive design incorporating SSR proposed by Li et al. [9]. Following the convention of Wang et al. [10], we refer to this as the ‘LSW’ (for Li, Shih and Wang) approach. We feel that the LSW approach strikes a nice balance between the flexibility of an adaptive design and the rigour of group-sequential design, as well as being comparatively simple to implement. In Section 2, we introduce our notation and describe the motivation for an adaptive SSR design over a fixed sample size design. In Section 3, we introduce the LSW method, show how it can be modified to accommodate capping of the maximum sample size and evaluate the operating characteristics of these two approaches compared with a fixed sample size design. In Section 4, an alternative method for choosing the design parameters of the LSW method is introduced. We conclude with a wide-ranging discussion of the adaptive approach and point to further research in Section 5.

Before proceeding any further, we firstly describe our motivation for this review.

1.1. Motivation

Current treatment options for knee osteoarthritis (OA) are not suitable or ineffective for large numbers of patients [11], and surgery is often the only remaining option. A grant application sought funding to conduct an RCT into the effectiveness of a standard rheumatoid arthritis oral therapy (with acceptable toxicity profile) to relieve pain in OA patients. An initial open-label pilot study in patients with knee OA had shown promising results, so the case for an RCT appeared to be strong. However, the funding council decided that the trial’s design should be substantially...
revised. They were concerned in particular about the lack of evidence on the effect size likely to be seen in the RCT context. The funder offered the opportunity to re-submit the application under the proviso that the new trial incorporated an interim analysis, with clear criteria to stop the study or to proceed with the full-scale recruitment (and new sample-size calculation). The authors were contacted through an advisory service to aid in the trial’s re-design in light of the funder’s response. We were looking for a simple, transparent method that

1. can be fully specified before any recruitment begins;
2. can be understood and easily implemented by an independent data monitoring committee (DMC);
3. is not motivated via a complex conditional error function;
4. is implemented through a clear decision framework that links the interim effect size estimate with future sample size via a simple, familiar formula;
5. allows the trial data to be analysed at the end using standard methods;
6. is a practical and understandable alternative to a specific fixed design for funding bodies and trialists to consider.

Despite feeling that the LSW approach satisfied points 1–4 in our checklist, the aim of this paper is to suggest further changes that address points 5 and 6.

2. NOTATION AND THE STANDARD SAMPLE SIZE CALCULATION

Assume that observations in the (experimental) treatment group $X$ and (standard therapy) control group $Y$ are normally distributed with means $\mu_X$ and $\mu_Y$, respectively, and have a common known variance of $\sigma^2$. The standardised mean difference, $\delta$, is defined as $\frac{\mu_X - \mu_Y}{\sigma}$ and is the measure of treatment effect we are interested in estimating. An estimate for $\delta$, $\hat{\delta}$, could be obtained from a trial with $n$ patients per arm by plugging in estimates for $\mu_X$ and $\mu_Y$, as below:

$$\hat{\delta} = \frac{\bar{x} - \bar{y}}{\sigma}$$

where $\bar{x} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right)$, $\bar{y} \sim N\left(\mu_Y, \frac{\sigma^2}{n}\right)$.

The test statistic, $z = \frac{\hat{\delta}}{\sqrt{n/2}}$, which follows a $N(\delta, \sqrt{n/2}, 1)$ distribution, can be used to test the null hypothesis $H_0: \delta \leq 0$. Under $H_0$, $z$ follows a standard normal distribution, so that $H_0$ is rejected if $z > Z_{\alpha}$, where $Z_{\alpha} = \Phi^{-1}(1 - \alpha)$ and $\alpha$ is the type I error rate. In a fixed sample size two-arm trial, the number of participants per arm, $n$, can be determined from the formula:

$$n = \frac{2}{\delta^2} \left(Z_{\alpha} + Z_{\beta}\right)^2,$$

where $1 - \beta$ is the power reject $H_0$ at $\delta$. For example, to obtain 80% power to detect a difference of $\delta = \delta_H = 0.35$ with a one-sided type I error rate of 2.5% ($\alpha = 0.025$ and $\beta = 0.2$), $n = 129$ patients per arm are needed.

There may, however, be a considerable uncertainty and/or lack of information on the parameter $\delta$ with which to base this calculation. If $\delta$ is truly less than 0.35, then substantially more than 129 people would be needed. Equally, if $\delta$ is truly much larger than 0.35, then the trial may be needlessly large.

2.1. A two-stage alternative

Suppose instead that $n_1 (< n)$ subjects are initially recruited per arm into the trial, and an interim analysis is conducted after their responses are observed. This would enable a $\hat{\delta}_1$ and $z_1$ to be obtained as

$$\bar{x} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right), \quad \bar{y} \sim N\left(\mu_Y, \frac{\sigma^2}{n}\right),$$

$$\hat{\delta}_1 = \frac{\bar{x} - \bar{y}}{\sigma}, \quad z_1 = \frac{\hat{\delta}_1}{\sqrt{2/n_1}} \sim N\left(\frac{\delta}{\sqrt{2/n_1}}, 1\right).$$

A decision could then be made on the number of additional subjects per arm needed, $n_2$, say, in the remainder of the trial, given the magnitude of $\hat{\delta}_1$ (or equivalently $z_1$). It is fairly intuitive to see that if one wanted to maintain the same a priori power of $1 - \beta$ conditional on the value of $z_1$ (assuming that $z_1$ is positive), then $n_2$ would be a decreasing function of $z_1$ — this will be shown explicitly in the next section. However, if $z_1$ were sufficiently small, then it may be decided to stop the trial altogether; on the grounds that the treatment effect was well short of the clinically relevant difference and, furthermore, the trial would need to be so large to detect this effect that it would be extremely unlikely to attract the necessary funding. Alternatively, the interim data could suggest stopping the trial for positive reasons if $\hat{\delta}_1$ and $z_1$ were so large as to provide overwhelming evidence against $H_0$.

However, care must be taken when interim data is used to make decisions about the trial’s eventual size and is then subsequently used in the final analysis. Such practices, if unaccounted for, can inflate the type I error rate above the nominal level [8,12]. In the next section, we describe the LSW method [9,13] that was proposed for this two-stage design framework. Given an initial sample of patient data, it provides a rationale for deciding whether to stop the trial or continue, and if continuing how many additional patients to recruit. Crucially, it does this whilst controlling the overall type I error rate and also setting the minimum power to reject the null hypothesis conditional on reaching the second stage.

3. A REVIEW OF THE LSW METHOD

Suppose that $n_1 + n_2(z_1)$ people are recruited to each arm of the trial according to some as yet unspecified ‘rule’, $n_2(z_1)$, save that $n_2(z_1)$ is 0 when $z_1$ is less than $h$ or greater than $k$ (the futility and efficacy boundaries, respectively). Suppose further that if $z_1 \in (h, k)$, then we desire a constant conditional power of $1 - \beta_1$ to reject $H_0: \delta \leq 0$ at the trial’s conclusion, on the basis of the sufficient statistic and likelihood ratio test:

$$z = \frac{\sqrt{n_1 z_1} + \sqrt{n_2(z_1) z_2}}{\sqrt{n_1 + n_2(z_1)}} \text{ reject } H_0 \text{ if } z \geq C,$$

where $z_2$ is the test statistic based on $n_2(z_1)$ people, derived in an identical fashion to equation (1). Li et al. [9] provide a methodology for choosing the critical value $C$ given the design parameters $h, k, \beta_1$ and overall type I error rate $\alpha$, that is independent of the interim test statistic $z_1$. It can be understood as a simple but clever modification of the general approach of Proschan and Hunsberger [8], which does not share this independence property. The method is now explained in detail using the original
notation of Li et al. First, define the conditional power function $CP_\delta(n_2, C|z_1, n_1)$ to be

$$P(z > C|z_1, n_1, \delta) = 1 - \Phi\left(\frac{C\sqrt{n_1 + n_2(z_1)} - z_1\sqrt{n_1} - n_2(z_1)\delta}{\sqrt{n_2(z_1)}}\right),$$

(3)

which is the probability of rejecting the null hypothesis at the trial’s conclusion given a sample of $n_1 + n_2(z_1)$ per arm. Let $CP_0(n_2, C|z_1, n_1)$ be this conditional power when $\delta$ is set to 0, so that the null hypothesis is true. The overall type I error rate for the design is equal to

$$P(z_1 > k|\delta = 0) + P(z > C|\delta = 0; h < z_1 < k)$$

$$= \alpha + \int_h^k CP_0(n_2, C|z_1, n_1)\phi(z_1)dz_1$$

(4)

is the conditional type I error rate in the region $(h, k)$. From equation (3), if we desire a conditional power of $1 - \beta_1$ to reject $H_0$ at stage two, then the following must hold:

$$\left(\frac{C\sqrt{n_1 + n_2(z_1)} - z_1\sqrt{n_1} - n_2(z_1)\delta}{\sqrt{n_2(z_1)}}\right) = -\Phi^{-1}(1 - \beta_1) = -Z_{\beta_1}.\quad (5)$$

The two unknowns defining the two stage design are $n_2(z_1)$ and $C$. They could be obtained by solving equation (5) for $n_2(z_1)$ (as a function of $C$ and $z_1$), and then plugging in its value to (4) to solve for $C$. Li et al. provide a simpler solution; make equation (5) an inequality by replacing the left hand side’s denominator, $\sqrt{n_2(z_1)}$, with $\sqrt{n_1 + n_2(z_1)}$, thereby guaranteeing a conditional power of at least $1 - \beta_1$. Further substituting $\delta$ with its stage one maximum likelihood estimate, $\hat{\delta}_1 = z_1\sqrt{2/n_1}$, leads to the solution:

$$n_2(z_1) = \left(\frac{(C + Z_{\beta_1})^2}{z_1^2} - 1\right)n_1, \quad \text{for } z_1 \in (h, k). \quad (6)$$

Note that the equivalent formula to (6) in Li et al. [9] is incorrect. Plugging $n_2(z_1)$ into equation (4) and re-arranging yields the following formula for $C$:

$$1 - \Phi(h) - \alpha = \int_h^k \phi\left(\frac{C(C + Z_{\beta_1}) - u^2}{\sqrt{(C + Z_{\beta_1})^2 - u^2}}\right)du. \quad (7)$$

Equation (7) can thus be solved to yield the critical value $C$ needed for the stage two test. Because (7) does not depend on $n_1, n_2$ or $z_1$, $C$ can be found before the trial starts and any data is observed. Furthermore, as long as $n_2$ is chosen via equation (6), then any $n_1$ can be used, meaning the timing of the interim analysis need not be specified in advance. The constant $C$ and constant minimal conditional power (set by $\beta_1$) of the LSW method are in marked contrast to the original approach proposed by Proschan and Hunsberger. In their approach, a conditional power function $Z_{\beta_1}(z_1)$ must be specified, and $C$ can only be calculated once $z_1$ has been observed. Note a technical detail; the upper limit on the integral is constrained by the fact that the square-rooted term in the denominator of (7) must be positive, so that $k < C + Z_{\beta_1}$. Li et al. set the integral’s upper limit to $k_1 = \min(k, C + Z_{\beta_1})$ to address this.

The LSW approach’s sequential nature, use of pre-specified stopping rules based on sufficient test statistics and its strict control of type I error rate, means that it bears a strong resemblance to a traditional group-sequential trial (GSD). This, we believe, is one of its strengths.

### 3.1. Example: a standard implementation of the LSW method

Following the funder’s response to the original arthritis trial application, the LSW method was investigated as a possible alternative. It was decided that it would be feasible to recruit an initial sample of $n_1 = 50$ patients per arm across the seven study centres within 6 months of the trial commencing. The outcome (change in knee pain from baseline at 24 weeks) would therefore be available for all patients 1 year after study initiation. Trial recruitment would be frozen in this period. If $\hat{\delta}_1 \leq 0.2$ (equivalent to $h = 1$ on the $z_1$ scale), then the trial would stop for futility and not recruit any further patients. This would equate to a p-value for the null hypothesis $\delta \leq 0$ of $\approx 0.16$. If on the other hand $\hat{\delta}_1 \geq 0.55$ (equivalent to $k = 2.76$ on the $z_1$ scale), then the trial would stop for efficacy. This would equate to a p-value of 0.003. If however the estimate was between 0.2 and 0.55, then additional participants would be recruited to each arm according to equation (6),

| Approach: Design | $h$ | $k$ | $1 - \beta_1$ | $C$ | $\alpha$ | $n_1$ | $n_{\text{max}}$ |
|-----------------|-----|-----|----------------|-----|---------|------|--------------|
| Standard
| Implementation | 1   | 2.76| 0.8            | 1.92| 0.025   | 50   | 333         |
| Modified LSW: 2| 1   | 2.76| 0.8            | 1.93| 0.025   | 50   | 90          |
| Reverse
| Implementation | 1.14| 2.24| 0.8            | 1.96| 0.025   | 70   | 353         |
| Modified LSW: 4| 1.08| 2.32| 0.8            | 1.96| 0.025   | 71   | 121         |

$n_{\text{max}}$ is the maximum stage 2 sample size.
guaranteeing at least 80% conditional power to reject $H_0$, with a type I error rate of 2.5%. By substituting the implied values of $k$, $\alpha$, $\beta_1$ and $h$ into equation (7), $C$ is found (via numerical integration) to be 1.923. The parameters defining this design are listed in Table I, and it is referred to as ‘design 1’. The value of $k$ was chosen to be as large as possible to give the smallest chance of stopping for efficacy at stage one. It is equal to 2.76 = $C + Z_{\beta_1} = 1.923 + 0.84$.

Thus, for the chosen values of $(h, \alpha, \beta_1)$, one is forced by the design of Li et al. to spend a minimum of 0.003 of the total type I error rate of 0.025 at stage one. However, the futility threshold $h$ effectively buys back this type I error rate (and more) because the final threshold $C$ is less than $Z_{\alpha} = 1.96$.

Figure 1 plots the total number of patients needed as a function of $\delta_1$ under design 1. We see that at the interim, if $\delta_1 = 0.35$, then only 125 patients per arm are required for the trial in total. The dotted line in Figure 1 shows the distribution of the estimate $\hat{\delta}_1$ when $\delta = 0.35$ to indicate the proportion of times the study would stop early for efficacy or futility at stage one, or continue to stage two with the specified sample size.

### 3.2. Incorporating a maximum sample size constraint

The LSW approach seemed to provide an alternative two-stage trial design that addressed the funders concerns. Although they were not prepared to pay for the fixed trial of 129 patients per arm based on existing evidence, they were potentially prepared to provide full funding if results from an initial phase were sufficiently promising. However, in design 1, the total sample size per arm could be anything from 50 to 383 depending on the value of $z_1$. Although the research team were keen to define the promising region for $z_1$ as $(1.2, 7.6)$, it was felt that there would be a maximum sample size (well below 380) beyond which the trial would probably not be funded. A total sample size of 140 per arm, or of $n_2 = n_{\text{max}} = 90$, was thought a plausible upper limit. To incorporate this constraint into the LSW method and preserve the property that the final stage test statistic threshold is independent of $z_1$, equations (6) and (7) must be modified. Equation (6) becomes

$$n_2(z_1) = \min\left[n_{\text{max}}, \left(\frac{\left(C_{\text{max}} + Z_{\beta_1}\right)^2}{z_1^2} - 1\right) n_1\right], \text{ for } z_1 \in (h, k)$$

$$= \left(\frac{\left(C_{\text{max}} + Z_{\beta_1}(z_1)\right)^2}{z_1^2} - 1\right) n_1, \text{ for } z_1 \in (h, k).$$

Although it is convenient to use formula (8) when calculating $n_2$, we must make use of the equivalent formula (9) in subsequent calculations. It tells us that the constant minimal conditional power term $Z_{\beta_1}$ is, in effect, replaced by a simple step function, $Z_{\beta_1}(z_1)$, where

$$Z_{\beta_1}(z_1) = \begin{cases} Z_{\beta_1}, & \text{if } \left(C_{\text{max}} + Z_{\beta_1}\right) \sqrt{\frac{n_1}{n_1 + n_{\text{max}}} < z_1 < k} \\ z_1 \sqrt{\frac{n_{\text{max}} + n_1}{n_1}} - C_{\text{max}}, & \text{if } h < z_1 < \left(C_{\text{max}} + Z_{\beta_1}\right) \sqrt{\frac{n_1}{n_1 + n_{\text{max}}}} \\ \end{cases}$$

and the relevant stage two threshold, $C_{\text{max}}$ (which is different to $C$), is found by solving the integral

$$1 - \Phi(h) - \alpha = \int_h^{k_1} \Phi\left[\frac{C_{\text{max}} \left(C_{\text{max}} + Z_{\beta_1}(u)\right) - u^2}{\sqrt{(C_{\text{max}} + Z_{\beta_1}(u))^2 - u^2}}\right] \phi(u)du. \quad (11)$$

The upper limit on the integral is defined as before. The denominator of the $\Phi(\cdot)$ function in (11) remains well defined for $Z_{\beta_1}(z_1) < Z_{\beta_1}$, because, from (10), when this occurs

$$C_{\text{max}} + Z_{\beta_1}(z_1) = C_{\text{max}} + z_1 \sqrt{\frac{n_{\text{max}} + n_1}{n_1}} - C_{\text{max}}$$

$$= z_1 \sqrt{\frac{n_{\text{max}} + n_1}{n_1}} \geq z_1.$$

### 3.3. Example revisited

Adding in the extra constraint $n_{\text{max}} = 90$ to the remaining parameters of design 1 in Section 3.1, we calculate $C_{\text{max}}$ to be 1.936. This is listed as ‘design 2’ in Table I, and we refer to capping the maximum sample size in this way as the modified LSW approach. Figure 2 (left) shows the total sample size of design 2 as a function of $\delta_1$. Figure 2 (right) shows the minimum conditional power guaranteed by this design, as a function of $\delta_1$. It starts at close to 40% when $\delta_1 = 0.2$ and increases up to a maximum of 80% by $\delta_1 = \delta_1^* = \left(C_{\text{max}} + Z_{\beta_1}\right) \sqrt{\frac{2}{n_1 + n_{\text{max}}}} = 0.332$. So, artificially constraining the sample size to not exceed a maximum value leads to some loss of power when $\delta_1$ is small. It is therefore important to assess this constraint’s effect on both on the overall power and expected sample size of the adaptive trial.

### 3.4. An assessment of designs 1–2

Figure 3 highlights the operating characteristics of the original fixed design proposal ($n = 129$, $\alpha = 0.025$, $\beta = 0.2$) and adaptive designs 1 and 2 as a function of $\delta$. They are calculated from the list of expressions given in Table S1 in the appendix (available online as Supporting Information). Figure 3 (top left) shows,
for adaptive designs 1 and 2, how the probability of stopping for efficacy or futility changes as $\delta$ increases from 0 to 1. The two probabilities are equal when $\delta$ equals the mid-point of $\sqrt{\frac{2}{m}} h$ and $\sqrt{\frac{2}{m}} k$. Under design 1, the probability that the total sample size is greater than the maximum of 140 per arm is maximised at around 26% when $\delta$ equals the mid-point of $\sqrt{\frac{2}{m}} h$ and $\delta_1$. The same value of $\delta$ maximises the probability that $n_2 = n_{\text{max}}$ under design 2. Figure 3 (top right) shows that the expected sample size of designs 1 and 2 is always less than that of the fixed design. The maximum expected sample size of design 2 is over 20 patients less than that of design 1. Figure 3 (bottom left) shows the overall unconditional power, $P(\text{Reject } H_0)$, of all three designs. Formula (12) in Table S1 gives this quantity, as well as a more standard formula for the power of the fixed design. The fixed design’s overall power is greater than the adaptive designs for all reasonable values of $\delta$. At the originally hypothesised value $\delta = \delta_{H_1} = 0.35$, the overall power is 80% by definition, whereas adaptive designs 1 and 2 only achieve an overall power of $\approx 71\%$ and $69\%$, respectively. This shortcoming of the adaptive designs is returned to in Section 4. Figure 3 (bottom right) shows the ratio of the design’s overall power with their expected sample size (which is of course constant for the fixed design). Comparisons of power between designs with different expected sample sizes can be misleading, so the power per unit of expected sample size provides a new and potentially useful standardised measure. Indeed, it has been recently employed by the second author to compare the relative merits of competing development strategies for phase II trials [14]. Despite design 2 being the least powerful of the three, it is the superior of the three for all values of $\delta$ according to
this measure. It also highlights how unnecessarily large the fixed
design is when $\delta$ is over 0.5.

4. AN ALTERNATIVE IMPLEMENTATION OF
THE LSW METHOD

The standard implementation of the LSW approach gives the user
freedom to choose the $h$ and $k$ (albeit subject to some weak con-
straints) determining whether or not the trial continues to stage
two. It also allows the user to specify the minimum conditional
power level desired at stage two and leaves the choice of $n_1$ com-
pletely open. As we have seen, this can help identify adaptive
designs with a far smaller expected sample size compared with
the fixed design. However, because the method is constructed to
control the minimum conditional power at stage two, we have
also demonstrated that it generally has a lower overall power,
as given by equation (11) (available online as Supporting Informa-
tion), compared with the fixed design. This power loss is especially
evident, and indeed relevant, at the originally hypothesised value
of $\delta_{H_1}$. Furthermore, the use of a non-standard critical thresh-
old ($C$ or $C_{\text{max}}$) at stage two that is different (and especially
lower) than the nominal $\alpha$ level will doubtlessly sit uncomfort-
ably with some trialists. For example, one could envisage the
following scenario: a clinical trial using a specific LSW design
proceeded to stage two and suggested a rejection of the null
hypothesis, because the final test statistic, $z$, was $> C$. However,
a standard analysis of the data based around the maximum likeli-
hood estimate (MLE) at the $\alpha$ level of significance—which would
invariably be preferred by the trial committee and general medi-
cal community—did not. With these two things in mind, we now
propose a different rationale for choosing the parameters in an
LSW or modified LSW design.

4.1. Reverse implementation of the LSW design

For the standard LSW method, rather than choosing $h, k, \alpha, Z_{\beta}$,
and determining $C$, we instead propose to identify a family of
possible designs by implementing the following algorithm:

(1) Identify a fixed sample size design with type I error $\alpha$ and
power $1-\beta$ at $\delta = \delta_{H_1}$.
(2) Find all joint values of $(h, k, Z_{\beta})$ consistent with $\alpha$ and
$C = Z_{\alpha}$ from equation (7).
(3) For each specific value of $(h, k, Z_{\beta})$, find the minimum value of
$n_1$ that sets the unconditional power in equation S1 at $\delta = \delta_{H_1}$
equal to $1 - \beta$.

Fixing $C$ to $Z_{\alpha}$ means that rejection of $H_0$ at stage two via
the adaptive design must coincide with a rejection based on a
standard analysis using $z$. The algorithm can be split into the
aforementioned steps two and three because equation (7) is inde-
dependent of $n_1$, and this also makes the numerical optimisation
an easier task. The solid line in Figure 4 shows the values of $(h, k,
Z_{\beta}, n_1)$ consistent with this strategy. Scales for $h$ and $k$ are shown
alongside the $p$-values for early stopping due to efficacy and futility
($P_H$ and $P_E$) they imply. For scales $Z_{\beta}$ and $1-\beta_1$, the expected
sample size per arm at $\delta_{H_1}$ and $n_1$ are also shown. The red point
highlights an interesting and appealing design, where the mini-
mum conditional power equals the unconditional overall power,
or $\beta = \beta_1$. This occurs at (approximately) $h = 1.14, k = 2.24$ and
$n_1 = 70$. This is listed as ‘design 3’ in Table I. The expected sample

4.2. Reverse implementation of the modified LSW design

Our motivation for proposing the modified LSW design was to
allow the user to limit the overall trial size through the second
stage sample size, $n_2$, given that the trial had already accrued $n_1$
patients in stage one. However, under the reverse implementa-
tion, $n_1$ is considered as an additional parameter in the design
space. This suggests that, rather than simply controlling the stage
two sample size via $n_{\text{max}}$, it would be more sensible to control
the maximum total sample size per arm, $(n_1 + n_2(z_1))$. An algo-

Figure 4. Possible parameter choices under the reverse implementation LSW
design.

To find possible modified LSW designs under this strategy is
now described. The algorithm is more computationally demand-
 ing than before, because equation (11) depends implicitly on $n_1$
through $Z_{\beta_1}(z_1)$.

(1) Identify a fixed sample size design with type I error $\alpha$ and
power $1-\beta$ at $\delta = \delta_{H_1}$. Additionally fix the maximum value of
$(n_1 + n_2(z_1)), n_{\text{max}}$ say, and set $C = Z_{\alpha}$
(2) Given $n_{\text{max}} = n_{\text{max}} - n_1$, find the joint values of
$(h, k, Z_{\beta}, n_1, n_{\text{max}})$ such that:

(a) $(h, k, Z_{\beta}, n_1, n_{\text{max}})$ are consistent with $\alpha$ and
$C = Z_{\alpha}$ from equation (11).
(b) $n_1$ is minimised given the joint values of $(h, k, Z_{\beta}, n_1, n_{\text{max}})$.
(c) The unconditional power in equation S1 at $\delta = \delta_{H_1}$
equals $1-\beta$.

Table II shows the joint values of $(h, k, Z_{\beta}, n_1, n_{\text{max}})$ consistent
with this strategy when $n_{\text{max}}$ is fixed at 192:50% larger than the
original fixed design’s sample size per arm of 129. A simple plot
is not possible because the design parameters do not all increase
or decrease together. We again highlight the design for which
the minimum conditional power equals the unconditional over-
all power, or $\beta = \beta_1$. This occurs at $h = 1.08, k = 2.32, n_1 = 71$ and
$n_{\text{max}} = 121$. It is listed as ‘design 4’ in Table I and high-
lighted in bold in Table II. The expected sample size of the design
at $\delta = \delta_{H_1} = 0.35$ is approximately 111.

Table I shows the joint values of $(h, k, Z_{\beta}, n_1, n_{\text{max}})$ consistent
with this strategy when $n_{\text{max}}$ is fixed at 192:50% larger than the
original fixed design’s sample size per arm of 129. A simple plot
is not possible because the design parameters do not all increase
or decrease together. We again highlight the design for which
the minimum conditional power equals the unconditional over-
all power, or $\beta = \beta_1$. This occurs at $h = 1.08, k = 2.32, n_1 = 71$ and
$n_{\text{max}} = 121$. It is listed as ‘design 4’ in Table I and high-
lighted in bold in Table II. The expected sample size of the design
at $\delta = \delta_{H_1} = 0.35$ is approximately 111.

Figure 5 compares the operating characteristics four designs
featured. Designs 1 and 2 have a far smaller expected sample size

© 2014 The Authors. Pharmaceutical Statistics published by John Wiley & Sons Ltd. Pharmaceutical Statist. 2014, 13 163–172
exhibit a low conditional power when in Figure 2 (right), the price paid for this property is that it can sample size that is always lower than the fixed design. As is shown additionally caps the maximum sample size—has an expected is provided by the adaptive approach. However, design 4—which to stop altogether if the interim data suggests so. This flexibility the size of (or bigger than) the fixed design, but want the chance funding bodies should generally be prepared to support a trial with SSR and suggested two extensions to make the trialist's are encouraged to continue recruiting up to the original sample size of \( n \). However, if \( z_1 \) is in the promising region, the user is encouraged to increase the sample size (over and above \( n \) up to a maximum level) and to fix the conditional power at \( 1 - \beta \) using the data derived treatment effect estimate. Following this decision framework is guaranteed to not inflate the type I error rate of a standard \( \alpha \)-level analysis at the trials end. Mehta and Pocock therefore argue that the use of standard analysis methods, as opposed to non-standard adaptive design methods (e.g [4,8]), makes the method attractive.

5. DISCUSSION

In this paper, we have reviewed the LSW method for a two-stage trial with SSR and suggested two extensions to make the approach more amenable for use in practice. We firstly modified their standard approach to allow the second stage sample size to be capped. This restriction may be useful in practice, because it would enable funders to maintain a reasonable grip on the maximum cost of the trial when presented with preliminary findings from stage one. The maximum size of the trial may not be limited by the funding body, but instead by the trialists, because of practical constraints on its duration, the number of study centres available and likely recruitment rate. Our modification would also be useful in this case. We secondly suggested an alternative ‘reverse’ implementation of the LSW approach. Our aim was to make the resulting LSW designs a more obvious and transparent alternative to a standard fixed sample size design. The reverse implementation can identify adaptive designs whose unconditional power is near identical with a specific fixed design (across a large range of treatment effects) whilst ensuring that its expected sample size is generally well below that of the fixed design. Furthermore, there is no danger that the resulting decision to reject the null hypothesis at stage two will be odds with inference based around the final test statistic.

### Table II. Possible parameter choices under the reverse implementation modified LSW design.

| \( h \) | \( k \) | \( 1-\beta_1 \) | \( n_1 \) | \( n_1 + E[n_2] \) | \( n_{max} \) |
|---|---|---|---|---|---|
| 0.700 | 2.76 | 0.847 | 54 | 117 | 138 |
| 0.751 | 2.66 | 0.835 | 56 | 115 | 136 |
| 0.802 | 2.59 | 0.826 | 58 | 114 | 134 |
| 0.853 | 2.52 | 0.818 | 61 | 114 | 131 |
| 0.904 | 2.47 | 0.813 | 63 | 113 | 129 |
| 0.955 | 2.41 | 0.808 | 65 | 112 | 127 |
| 1.010 | 2.37 | 0.804 | 67 | 111 | 125 |
| 1.060 | 2.34 | 0.801 | 70 | 111 | 122 |
| **1.080** | **2.32** | **0.800** | **71** | **111** | **121** |
| 1.110 | 2.30 | 0.799 | 72 | 111 | 120 |
| 1.160 | 2.27 | 0.796 | 75 | 111 | 117 |
| 1.200 | 2.25 | 0.795 | 77 | 111 | 115 |

than 3 and 4 but because of this, do not control the overall power at \( \delta = \delta_H \), at the original desired level of 80%. As well as being identical to the fixed design at \( \delta = \delta_H \), the unconditional power curves of designs 3 and 4 are very close to that of the fixed design across all values of \( \delta \).

Design 3 has a higher expected sample size than the fixed design at \( \delta = 0.33 \). This may raise concerns in some quarters, but we do not view it as an inherent weakness. Our argument is that funding bodies should generally be prepared to support a trial the size of (or bigger than) the fixed design, but want the chance to stop altogether if the interim data suggests so. This flexibility is provided by the adaptive approach. However, design 4—which additionally caps the maximum sample size—has an expected sample size that is always lower than the fixed design. As is shown in Figure 2 (right), the price paid for this property is that it can exhibit a low conditional power when \( z_1 \) is just above the futility threshold \( h \).

5.1. The approach of Mehta and Pocock

Although different in its exposition, our approach is similar in spirit to the recent work of Mehta and Pocock [15], that has generated a lot of discussion in the literature (see [16,17]). In the same context of two-stage trials with mid-trial sample size adjustment, Mehta and Pocock were motivated to find a simple composite design strategy that would be attractive to trialists who were contemplating moving away from a traditional fixed sample size design. Using the notation of this paper, they encourage the trialist to specify a sample size based on a fixed design (with sample size \( n \), type I error \( \alpha \) and power \( 1 - \beta \)), but to allow for an interim after \( n_1 \) patients. At this interim, they identify values of the test statistic, \( z_1 \), that correspond to so-called unfavourable, promising and favourable regions of conditional power—defined analogously to equation (3). If \( z_1 \) is in the favourable or unfavourable regions then the trialist’s are encouraged to continue recruiting up to the original sample size of \( n \). However, if \( z_1 \) is in the promising region, the user is encouraged to increase the sample size (over and above \( n \) up to a maximum level) and to fix the conditional power at \( 1 - \beta \) using the data derived treatment effect estimate. Following this decision framework is guaranteed to not inflate the type I error rate of a standard \( \alpha \)-level analysis at the trials end. Mehta and Pocock therefore argue that the use of standard analysis methods, as opposed to non-standard adaptive design methods (e.g [4,8]), makes the method attractive.

![Figure 5. Expected sample size (left) and overall power (right) of the LSW methods under designs 1–4, as a function of \( \delta \).](image-url)
Our reverse implementation of the LSW approach in designs 3 and 4 allow for early stopping at stage one but provide the termination thresholds and a simple SSR rule such that a standard $\alpha$-level analysis is still possible at the final stage. They can deliver substantial reductions in the expected sample size compared with a fixed design when $\delta$ is much smaller or larger than expected (especially modified LSW design 4), but with no appreciable loss in overall power. Furthermore, the $\alpha$-level analysis conducted at the end of stage two is correct in the sense that it preserves the theoretical type I error rate at $\alpha$, whereas Mehta and Pocock’s $\alpha$-level threshold is actually over-conservative and therefore inefficient.

5.2. Regulatory support for the use of unblinded SSR

In recent US Food and Drug Administration (FDA) guidance to industry on adaptive trials [1], the FDA is positive about the use of blinded SSR, citing its ability to improve study efficiency and ability to achieve the study goal, without affecting the type I error rate of the trial (Section B, lines 668–674). In contrast, no explicit endorsement can be found in this document on the use of unblinded SSR and so it is necessary to look for principles of guidance more widely.

Revising a study’s design in light of an unblinded interim analysis has the potential to induce bias and type I error inflation. Therefore, revisions should be prospectively defined and carefully implemented to avoid bringing the interpretation of study results into doubt. This legitimises the use of GSDs in the FDA’s eyes [1] (Section D, lines 817–829), where formal testing of a null hypothesis is carried out at one or more interim analyses to make decisions regarding the continuation of the trial. This is under the proviso that (a) methods for controlling resulting type I error inflation are incorporated as standard; (b) an independent DMC is tasked with reviewing the data at the interim analyses, and furthermore, a statistician independent of the study prepares the report for the DMC.

Throughout this paper, we have pointed the LSW approach’s similarity with GSDs, both in their sequential design and in their analysis. So, can current FDA guidance on GSDs be invoked to cover its use? Our opinion is a qualified ‘yes’. Although (a) and (b) would be sufficient for a GSD, the LSW approach (and any SSR procedure for that matter) is arguably more vulnerable to the de-masking of interim results for the following reason: A keen and suitably qualified individual in the study team could potentially transform the recommended sample two sample size into the stage one effect estimate and use this to influence the trial going forward. We therefore believe that an additional condition is necessary; information about the precise value of the stage two sample size must not be fed back to the study team unnecessarily. Rather, trial recruitment should be allowed to continue or stop, until the independent statistician can reveal that the planned size is about to be reached.

5.3. Estimation following an adaptive design

The MLE of $\delta$ at the end of the adaptive design will generally be biased, because it ignores the trial’s sequential nature. In the appendix, we provide a detailed investigation of the MLEs properties in this context and contrast it with that of the median unbiased estimate (MUE) suggested by Wang et al. [10]. The MUE is shown to provide estimates with a reduced bias and mean squared error compared with the MLE, when $\delta$ is small and positive.

5.4. Implementation of the adaptive approach when $\sigma$ unknown

We assume that $\sigma$ is known in the calculations used to both find our designs and report their operating characteristics. The simple mathematical formulae would not work if $\sigma$ were treated as a random variable. However, in practice one will need to estimate it from the data to implement any of the design proposals. It is important therefore to verify that this estimation does not cause a design’s true operating characteristics to differ substantially from its theoretical counterpart. Figure 6 (left) shows the expected sample size of Design’s 1 and 2 as a function of $\delta$ using (a) theoretical calculation (i.e. using formulae from Table S1) and (b) via simulation (incorporating estimation of $\sigma$ separately at stage 1 and 2). To clarify, treatment and control group data were simulated from equation (1) for specific values of $\mu_x$, $\mu_y$ and $n_1$, but with a common value of $\sigma = 20$. This defined the theoretical value of $\delta$. A pooled estimate for $\bar{\delta}$, $\bar{\delta}_2$ was then obtained from these two populations and $\hat{\delta}_2$ was estimated as $(\bar{k} - \bar{y})/\bar{\delta}$. If the trial proceeded to stage 2, $\sigma$ was re-estimated from the

![Figure 6](image-url)

**Figure 6.** Left: expected sample size of the LSW (design 1) and modified LSW (design 2) using theoretical calculation (black) and using empirical simulation (red). Right: type I error rate inflation (above the nominal 0.025 level) when the data is used to estimate $\sigma$ under the modified LSW design. Design 4 highlighted in red.
n2(z1) additionally simulated patients in each arm in the same manner, and used to calculate 1, 2 and 2 for equation (2). The difference between the theoretical expected sample size and those obtained in practice (with estimation of ) is tiny, which is reassuring. The theoretical and practical power curves for these designs are also near identical (results not shown). However, it is of crucial importance to check that the type I error rate is not drastically inflated (i.e. the power when 0).

Figure 6 (right) the theoretical and practical type I error rate of the reverse implementation modified LSW design (with a nominal -level of 0.025) as a function of the stage 1 sample size, n1. Each point corresponds to a row (possible design) in Table II, with the red dot corresponding to design 4. Some inflation is clearly present. As n1 increases this inflation reduces and when n1 is equal to 71 (design 4) the inflation is 0.0018 over the stated 2.5% level. For a comparison we also plot the type I error rate of the fixed design (n = 129, 0 = 0.025) when 2 is estimated from the data. Design 4 is 0.001 above this level.

In summary, there is a minimal difference in the operating characteristics when 2 is estimated from the data and n1 is of a reasonable size. It therefore seems a sensible strategy to use the known variance assumption to identify sensible designs and, if one so desires, to then add small perturbations to the parameters in conjunction with empirical simulations until the observed operating characteristics are acceptable. This might be efficiently achieved by substituting threshold parameters (e.g. 1, 2, 3) that are close to the equivalent quantiles mapped from the -distribution. For example, if the maximum sample size of 192 is used under design 4, then a value for 4 close to 0.025,191 = 1.972 (instead of 1.96) may be sensible first guess.

5.5. Further work

One may wish to extend the modified LSW method to allow not only a maximum cap to be put on the future sample size (given continuation) but also a minimum cap. This extra design facet may be needed in practice if the study team do not wish to halt recruitment whilst waiting for the stage outcome data to be observed. In our motivating example, this was up to 6 months, which is a lengthy delay. Of course, if the interim decision is to stop the trial but further patients end up being recruited, then methods for dealing with trial overrun must be employed. Koyama and Chen [18] have investigated this issue for two-stage trials with a binary response, and it would be interesting to see if this could be generalise to the setting we have discussed.

Using the interim effect estimate to evaluate, the conditional power has been criticised, because for a fixed 3, CP3(1,2|z1, n1) is a random variable containing a substantial amount of variability [19]. Thus, it is very important to understand the operating characteristics of any design procedure that utilises conditional power in this way. We have tried to do this here, for a large range of possible values for the parameter . Unwanted variability in CP3(1,2|z1, n1) can be mitigated to a certain extent by restricting n1 to be greater than a minimal value. As we have seen, this will also limit the need to correct for any type I error inflation caused by estimating . However, as further work, we plan to extend the LSW approach to explicitly account for the uncertainty in the estimation of and for using Bayesian and semi-Bayesian approaches, as in Wang [20].

Multi-arm multi-stage trials—in which several active treatments are sequentially tested against a standard therapy—are becoming increasingly popular in the era of stratified medicine. The STAMPEDE trial is a prime example [21]; it has the additional interesting feature whereby early outcome data (on progression free survival) is used to decide whether specific treatment arms should remain active in the trial, whereas the final analysis of a treatment’s effect will be based on overall survival. So far SSR has not been considered for such designs but may offer some utility. To apply the LSW approach, one would need to generalise it to account for interim estimation of the primary endpoint based on a correlated secondary endpoint. The definition of power used to guide the calculation would also need to be carefully chosen, as different definitions are possible when multiple hypotheses are being tested.

Software is made available at http://www.mrc-bsu.cam.ac.uk/Software/download.html to reproduce the set of reverse implementation LSW and modified LSW designs shown in Figure 4 and Table II respectively. This work was funded by the Medical Research Council (grant number G0800860). The authors would like to acknowledge the reviewers’ for their helpful comments which greatly improved this manuscript.

REFERENCES

[1] US Food and Drug Administration. Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics. 2010. Available from: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf [Accessed on June 2012].
[2] Gould AL, Shih W. Modifying the design of ongoing trials without unblinding. Statistics in Medicine 1998; 17:89–100.
[3] Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not effect the type I error rate. Statistics in Medicine 2003; 22:3571–3581.
[4] Bauer P, Kohne K. Evaluation of experiments with adaptive interim analysis. Biometrics 1994; 50:1029–1041.
[5] Fisher L. Self-designing clinical trials. Statistics in Medicine 1998; 17:1551–1562.
[6] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine 2003; 22:971–993.
[7] Jennison C, Turnbull B. Adaptive and nonadaptive group sequential tests. Biometrika 2006; 93:1–21.
[8] Proshans M, Hunsberger S. Designed extension of studies based on conditional power. Biometrics 1995; 51:1315–1324.
[9] Li G, Shih W, Xie T. A sample size adjustment procedure for clinical trials based on conditional power. Biostatistics 2002; 3:277–287.
[10] Wang Y, Li G, Shih W. Estimation and confidence intervals for two-stage sample size flexible design with lsw likelihood approach. Statistics in Biosciences 2010; 2:180–190.
[11] Zhang W, Nuki G, Moskowitz R, Abramsom R, Arden N, Bierma-Zeinstra S, Brandt K, Croft P, Doherty M, Dougados M, Hochberg M, Hunter D, Kwoh K, Lohmander L, Tugwell P. Oarsi recommendatons for the management of hip and knee osteoarthristis; part iii: Changes in evidence following systematic cumulative update of research published through january 2009. Osteoarthritis Cartilage 2010; 18:476–499.
[12] Graf A, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. Statistics in Medicine 2011; 30:1637–1647.
[13] Li G, Shih W, Wang Y. Two-stage adaptive design for clinical trials with survival data. Journal of Biopharmaceutical Statistics 2005; 15:707–718.
[14] Grayling M, Mander A. Do single-arm trials have a role in Phase II development plans? 2014, Submitted.
[15] Mehta C, Pocock S. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine 2011; 30:3267–3284.
[16] Emerson SS, Levin C, Emerson SC. Comments on ‘Adaptive increase in sample size when interim results are promising: A practical guide with examples’. Statistics in Medicine 2011; 30:3285–3301.
[17] Glimm E. Comments on ‘Adaptive increase in sample size when interim results are promising: A practical guide with examples’ by C.R. Mehta and S.J. Pocock. *Statistics in Medicine* 2012; 31:98–99.

[18] Koyama T, Chen H. Proper inference from Simon’s two-stage designs. *Statistics in Medicine* 2008; 27:3145–3154.

[19] Bauer P, Koenig F. The reassessment of trial perspectives from interim data: a critical view. *Statistics in Medicine* 2006; 25:23–36.

[20] Wang M. Sample size re-estimation by Bayesian prediction. *Biometrical Journal* 2007; 49:365–377.

[21] Sydes MR, Parmar MKB, James ND, Clarke NW, Deenaley DP, Mason MD, Morgan RC, Sanders K, Royston P. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* 2009; 10:39.

[22] Fairbanks K, Madsen R. P-values for tests using a repeated significance test design. *Biometrika* 1982; 69:69–74.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article at the publisher’s web site.