WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or, Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/130512

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 4.0 International license (CC BY 4.0) and may be reused according to the conditions of the license. For more details see: http://creativecommons.org/licenses/by/4.0/.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

An adaptive two-arm clinical trial using early endpoints to inform decision making: design for a study of sub-acromial spacers for repair of rotator cuff tendon tears

Nick Parsons[1*], Nigel Stallard[1], Helen Parsons[2], Philip Wells[2], Martin Underwood[2,3], James Mason[2] and Andrew Metcalfe[2,3]

Email: Nick Parsons, nick.parsons@warwick.co.uk; Nigel Stallard, n.stallard@warwick.ac.uk; Helen Parsons, h.parsons@warwick.ac.uk; Philip Wells, philip.wells@warwick.ac.uk; Martin Underwood, m.underwood@warwick.ac.uk; James Mason, j.mason@warwick.ac.uk; Andrew Metcalfe, a.metcalfe@warwick.ac.uk

Abstract

Background: There is widespread concern across the clinical and research communities that clinical trials, powered for patient reported outcomes, testing new surgical procedures are often expensive and time-consuming, particular when the new intervention is shown to be no better than the standard. Conventional (non-adaptive) randomized controlled trials (RCTs) are perceived as being particularly inefficient in this setting. Therefore, we have developed an adaptive group sequential design that allows early endpoints to inform decision making and show, through simulations and a worked example, that these designs are feasible and often preferable to conventional non-adaptive designs. The methodology is motivated by an on-going clinical trial investigating a saline-filled balloon, inserted above the main joint of the shoulder at the end of arthroscopic debridement, for treatment of tears of rotor cuff tendons. This research question and setting is typical of many studies undertaken to assess new surgical procedures.

Methods: Test statistics are presented, based on the setting of two early outcomes, and methods for estimation of sequential stopping boundaries are described. A framework for the implementation of simulations to evaluate design characteristics is also described.

Results: Simulations show that designs with one, two and three early looks are feasible and, with appropriately chosen futility stopping boundaries, have appealing design characteristics. A number of possible design options are described that have good power, and have high probability of stopping for futility if there is no evidence of a treatment effect at early looks. A worked example, with code in R, provides a practical demonstration of how the design might work in a real study.

Conclusions: In summary, we show that adaptive designs are feasible and could work in practice. We describe the operating characteristics of the designs and provide guidelines for appropriate values for the stopping boundaries for the START:REACTS (Sub-acromial spacer for Tears Affecting Rotator cuff Tendons: a Randomised, Efficient, Adaptive Clinical Trial in Surgery) study.

Trial registration: ISRCTN Registry: ISRCTN17825590. Registered on 5th March 2018.

Keywords: Adaptive design; Stopping for futility; Early endpoints

*Correspondence: nick.parsons@warwick.co.uk
[1] Statistics and Epidemiology Unit, Warwick Medical School, University of Warwick, CV4 7AL Coventry, UK
Full list of author information is available at the end of the article

Background

New surgical procedures are usually introduced based on what a surgeon believes might benefit patients and nothing more. While pharmaceuticals undergo rigorous clinical trials before introduction, this is not the case for surgical procedures, which are often introduced based purely on basic science (such as cadaveric testing) or small case series data only. There is a need to develop new processes and methodol-

ogy to introduce surgical procedures safely [1–3], with early randomised controlled trials (RCTs) in specialist centres used to determine whether a treatment is likely to be safe, clinically effective and cost effective prior to widespread uptake. Large clinical trials powered for patient reported outcomes are typically expensive and often take more than five years from award to completion. Ineffective, unsafe and costly treatments may be used for many years before they are removed from practice. This is clearly unacceptable and unethical. Conversely, very effective treatments may be withheld from widespread practice until trials are complete, leading to long delays in the delivery of worthwhile treatments for patients. Trial designs are required which can efficiently and rapidly determine that a procedure is ineffective or harmful, but will also adapt to demonstrate superiority if the technique is a genuine improvement on standard care. There is a growing awareness amongst both funders and researchers that conventional clinical trial designs are not the best option in many settings, and that novel adaptive design methods offer the potential to undertake clinical trials in a much more flexible manner, whilst retaining trial integrity.

An adaptive clinical trial allows for prospectively planned changes to be made to some aspects of the design, as it proceeds, using data collected from participants recruited into the study. These types of designs have grown in popularity in recent years [4] providing flexibility for trialists to, for instance, refine sample sizes, drop interventions (or doses of a drug), identify and focus recruitment on responsive subgroups (enrichment) or stop studies early [5]. For trials of new surgical interventions the option to potentially stop the study early, has particular appeal.The advantage of stopping a trial early are twofold. First in many widely encountered settings it is likely to make the trial design more efficient [5, 6]. For instance, if a test treatment is in truth much less effective than initially anticipated (or is totally ineffective), then the expected sample size and duration of a design that allows early stopping will be less than a comparable conventional fixed sample size (non-adaptive) design. Second, stopping a study early because an intervention is shown to be ineffective (under the null hypothesis) or conversely is shown to be effective (under the alternative) is clearly ethically beneficial, as it allows people to receive better treatments faster. Adaptive designs offer the potential of considerable advantages when compared to more conventional fixed designs, however there are often barriers to their implementation [7], and disadvantages such as the requirement to use or develop more complex statistical tools, the additional pressures on data monitoring and collection and the maintenace of trial integrity [8].

In surgical trials, participants are often routinely followed-up at a number of occasions (e.g. 3, 6 and 12 months) and the main study outcome(s) collected at each occasion. Therefore, at an interim analyses there will be some participants with 3m data, some with 3m and 6m data and some with 3m, 6m and 12m data. If interim analyses are limited to only those participants with 12m data (primary outcome), then the opportunities for early stopping if there is evidence to support either treatment *futility* or *efficacy* may well be severely limited due to time constraints; i.e. recruitment may well have completed before enough 12m outcome data are available for reliable decision making. If early endpoints are correlated with the definitive (final) study endpoint, then clearly an analysis that ignores the early endpoints for interim decision making is likely to be inefficient. Stallard [9] showed

that using *short-term* (or what others often call *early endpoint*) data, in the setting of a seamless phase II/III clinical trial with treatment selection with a single early endpoint, leads to increases in statistical power when these data are correlated with the primary endpoint.

As a consequence of the perceived lack of efficiency and inflexibility of traditional RCTs, the UK National Institute for Health Research [10] is funding a surgical RCT that will use a novel adaptive study design approach, developed specifically for the evaluation of new surgical procedures (Efficacy and Mechanism Evaluation Programme: 16/61 Evaluation of new surgical procedures through the use of novel study designs). This RCT provides the motivation for the work outlined here. In this paper we adapt the approach previously described by Stallard [9], which used a single early endpoint in a treatment selection design. Here we generalise to the setting with more than one early endpoint for comparing two treatment groups [11], and outline how the methodology can be used for interim decision making using an ongoing study of sub-acromial spacers for rotator cuff tendon tears as an exemplar. We start by providing the clinical context and then develop a model for the distribution of the outcomes, and give an expression for an appropriate test statistic and describe how inferences and decisions about stopping are made in the chosen setting. Simulations are undertaken and operating characteristics are illustrated for a wide range of design options. The aim of the work described here is to outline the process undertaken to develop a design for the specific trial that motivated this work, the final selection of the design options for that study will be made by and remain confidential within the study team. A practical worked example, using synthetic data, is used to explain how the selected design would work in practice. Although the focus here is on a particular surgical intervention and a specific trial, we believe that the methodology described will have wider application for many other clinical procedures in areas outside of the chosen setting.

Clinical context

The rotator cuff is a group of muscles around the shoulder that help to stabilise the joint and initiate movement. Tears of the tendons of the rotator cuff, typically where they attach onto the humerus, are very common. Patients may present with persisting pain, loss of movement, and substantial limitations in their activities of daily living. Treatment often consists of physiotherapy but if this is not successful then surgery to repair the tear may be required. Sometimes the tears cannot be repaired and there are very few effective treatments in this situation. Arthroscopic debridement has traditionally been used in this setting, it is an operation to clear space around the tendons and shoulder to allow it to move more freely and with less pain. There are concerns that this operation has little benefit over non-operative care [12], leading to calls for innovative solutions to treat this painful and disabling condition [13]. A newly available treatment option is a saline-filled balloon inserted above the main joint of the shoulder at the end of an arthroscopic debridement; the *InSpace* balloon device [14]. It is simple to deploy and adds less than 10 minutes to the operation. However, it is costly and evidence for efficacy is scant [15]. It provides a cushion inside the shoulder joint that should improve biomechanics

and hence reduce pain and improve shoulder function. We are running an adaptive, patient-assessor blinded, RCT across multiple centres in the UK, comparing standard arthroscopic debridement to standard arthroscopic debridement *plus* insertion of the *InSpace* balloon.

Methods

START:REACTS study

The START:REACTS study [16] (Sub-acromial spacer for Tears Affecting Rotator cuff Tendons: a Randomised, Efficient, Adaptive Clinical Trial in Surgery) commenced recruitment in autumn 2018; ISRCTN registration ISRCTN17825590 [17]. Recruitment is expected to take 24 months. In the following subsections we discuss important issues that motivated and determined the final study design, and provide a mathematical description of the methods that will be used to allow the possibility of early stopping.

Study outcomes

The primary outcome for the START:REACTS study is the Constant-Murley (C-M) shoulder score at 12 months (12m) [18, 19], which is widely used in trials, accepted by surgeons and has good reliability and responsiveness [20–23]; early outcomes will also be collected at 3m and 6m post-operation. Based on a recent meta-analysis, it is expected that the C-M score reaches a plateau by 12m after intervention for a rotator cuff tear [24]. The scoring system consists of four subscales (pain, activities of daily living, strength and range of motion) that are combined to give a score out of 100 (perfect function).

Sample size

A minimum clinically important difference (MCID) for the Constant-Murley (C-M) score of 10 units has been widely used for other trials [12, 25, 26]. For the purposes of analysis, the C-M score is considered to be approximately normally distributed with a standard deviation of 20 giving a moderate standardised mean difference of 0.5 [12, 27]. A recent meta-analysis [24] reported that standard deviations did not differ much between 3m, 6m and 12m, which is consistent with our own more detailed analysis of data available from another study reporting C-M scores [26]. For a costly invasive procedure of this nature an effect size smaller than 10 units is unlikely to be considered worthwhile. For a power of 90% to detect an effect of this size and two-sided type I error rate of 5% a study without early stopping would require 170 participants (85 in each intervention group). The START:REACTS study was initially powered on this basis, with a 20% allowance for some loss to follow-up, giving a maximum sample size of 212.

Recruitment is planned to take 24 months at 15 centres; recruitment will begin with a single centre at month 1, increasing to 2 centres at 2 months, 3 centres at 3 months, 6 centres at 4 months, 9 centres at 5 months, 12 centres at 6 months and 15 centres at 7 to 24 months. A total of 303 months of recruitment, which, assuming a constant recruitment rate at each centre, for a target of 170 participants means a

rate of (approximately) 0.56 participants per centre per month.
Pilot work from a survey of shoulder surgeons, undertaken immediately prior to the start of the study, indicated that a treatment difference in the range 7.5 to 10 points on the C-M scale provided moderate to strong evidence in favour of the balloon intervention. Therefore when considering options for stopping boundaries for the adaptive design, we would want to set these such that we had a low probability of stopping for futility for effects sizes of this magnitude, whilst at the same time stopping with high probability (for futility) for treatment differences in the range 0 to 2.5 points on the C-M scale.

Correlations between early and long-term outcomes
The best available evidence for correlations between early endpoints and the variance of the C-M shoulder score at 3m, 6m and 12m comes from a study undertaken in an analogous setting but in a different population to that planned for the START:REACTS study [26]. These data give estimates for the correlation between C-M shoulder scores at 3m and 6m as $\rho_{3m,6m} = 0.51$, between 6m and 12m scores as $\rho_{6m,12m} = 0.59$ and between 3m and 12m scores as $\rho_{3m,12m} = 0.46$. Therefore for the purposes of the simulations exploring the characteristics of the adaptive designs we will assume a uniform correlation model (i.e. correlations between 3m, 6m and 12m data are equal) with a value of 0.5.

Stopping window
The likely pattern of recruitment suggests that the window of opportunity for early stopping for the START:REACTS study will be relatively short. Presuming collection of primary 12m outcome data commences promptly and proceeds to plan, and as we will not want to take an interim look before some 12m data are available, then it is likely that only after 18m of recruitment could early looks at the data begin. Early looks at the data will need to complete by the end of recruitment at 24m. Therefore, in practice, there will likely be a period of approximately 6 months when early looks at the data are possible. If this is the case, then the feasible number of early looks at the data will be small. Therefore for the simulations exploring the characteristics of the adaptive designs we will assume that there are either one, two or three early looks at the data.

Statistical model
In the START:REACTS study the early endpoints at 3m and 6m are monitored in addition to the primary 12m endpoint. At the time of an interim analysis, before recruitment is complete, many more participants will have early endpoint data than 12m (primary) endpoint data. Although the 3m and 6m early endpoint data are useful for monitoring purposes, participant retention and safety issues, from a clinical perspective a treatment effect observed at 3m or 6m will not necessarily translate to a treatment effect at the definitive 12m endpoint; i.e. early benefit for the active intervention may not be sustained to the primary (clinically relevant)

12m endpoint. Therefore, at the early looks we wish to gain information on the final 12m endpoint from the early endpoints based on their expected within-participant correlations, irrespective of any early treatment effects. Stallard [9] shows that using early endpoint data, in a treatment selection (phase II/III) setting, leads to increases in power when these data are correlated with the primary endpoint, even if treatment effects on endpoints are unrelated. In the following sections we briefly outline the methods developed by Stallard [9] to control the familywise error rate in this setting and provide explicit expressions to estimate tests statistics when there are two early endpoints.

Distribution of outcomes

Suppose participants in a study are followed-up and data are collected on the same endpoint at a number of occasions, then let X_{ijK} be the final long-term outcome and $X_{ij1} \dots X_{ij(K-1)}$ be $K-1$ early (short-term) outcomes, for participant i in intervention arm j. We assume outcomes are independent for different participants and that the distribution of outcomes $(X_{ij1}, \cdots, X_{ijK})$ is multivariate normal, with mean $(\mu_{1j}, \cdots, \mu_{Kj})$ and variance

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_K\rho_{1K} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \cdots & \sigma_2\sigma_K\rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_K\sigma_1\rho_{K1} & \sigma_K\sigma_2\rho_{K2} & \cdots & \sigma_K^2 \end{pmatrix},$$

where σ_k^2 is the variance of the outcome X_k and $\rho_{kk'}$ is the correlation between endpoints X_k and $X_{k'}$.

Test statistic

For a two-arm study, participants are randomized to either the control ($j = 0$) or active intervention ($j = 1$) arms, and at an interim analysis long-term (final) outcomes are available from N_K subjects and early (short-term) outcomes from $N_1 \dots N_{K-1}$ subjects in each arm of the study. For the settings we are interested in, we assume that at any time during follow-up $N_1 \geq N_2 \geq \cdots \geq N_K$; i.e. there are always more or equal numbers of subjects providing data for the earlier outcome X_{k-1} than the later outcome X_k. The parameter of primary interest is the effect of the test intervention on the long-term (primary) outcome X_K. Following Galbraith and Marschner [11], the treatment effect B, which uses all the available early endpoint data for two short-term outcomes (X_1 and X_2), for instance at 3m and 6m such as in our chosen setting, and a single long-term outcome X_3 (at 12m) is given by

$$B = \frac{1}{N_3}\Bigg[\sum_{i=1}^{N_3}(X_{i13} - X_{i03}) + \rho_{13}\frac{\sigma_3}{\sigma_1}\sum_{i=N_3+1}^{N_1}\Big(X_{i11} - X_{i01} - \frac{1}{N_1}\sum_{m=1}^{N_1}(X_{m11} - X_{m01})\Big) + \rho_{23}\frac{\sigma_3}{\sigma_2}\sum_{i=N_3+1}^{N_2}\Big(X_{i12} - X_{i02} - \frac{1}{N_2}\sum_{m=1}^{N_2}(X_{m12} - X_{m02})\Big)\Bigg], \quad (1)$$

with variance

$$\mathrm{var}(B) = \frac{2\sigma_3^2}{N_3}\Bigg[1 - \rho_{13}^2\frac{N_1 - N_3}{N_1} - \rho_{23}^2\frac{N_2 - N_3}{N_2} + 2\rho_{13}\rho_{23}\rho_{12}\Big(1 - \frac{N_3}{N_2}\Big)\Bigg]. \quad (2)$$

Estimates $\hat{B}$ and $\mathrm{var}(\hat{B})$ follow from estimates of the correlations ρ_{13}, ρ_{23} and ρ_{12} and standard deviations, σ_1, σ_2 and σ_3, obtained from the appropriate regression models, using all available data. Expressions (1) and (2) are presented for the special case of equal numbers of subjects in each arm of the study. However, they can be modified easily for the case of unequal numbers in the study arms. These and more general expressions for B and $\mathrm{var}(B)$, for $K-1$ early outcome are provided in the supplementary material. From expressions (1) and (2) it is clear that if long-term outcome X_3 is uncorrelated with short-term outcomes X_1 and X_2 (i.e. if $\rho_{13} = \rho_{23} = 0$), then B and $\mathrm{var}(B)$ simplify to conventional expressions we would use to estimate the mean treatment effect (and variance) for X_3 alone, without reference to the early endpoints. As correlations between X_3 and X_1 and X_2 increase in magnitude then $\mathrm{var}(B)$ decreases, provided that the two early putcomes X_1 and X_2 are not themselves strongly correlated. In general, $\mathrm{var}(B)$ is minimized as both $\rho_{13} \to 1$ and $\rho_{23} \to 1$, and $\rho_{12} \to 0$; i.e. X_1 and X_2 are strongly correlated with X_3, but are themselves uncorrelated.

Implementation for a two-arm trial

For a two-arm study, with two short-term outcomes, study participants are randomized to either the control or active intervention arms, and data collection proceeds until the first interim analysis when N_{31} long-term data and N_{11} and N_{21} short-term data are available per arm; N_{3w}, N_{2w} and N_{1w} are the number of study participants with long and short-term data available at early look w. Expressions (1) and (2) are used to obtain the test statistic $S_1 = \hat{B}_1/\mathrm{sd}(\hat{B}_1)$, and observed information $\hat{I}_1 = 1/\mathrm{var}(\hat{B}_1)$, using estimates $\hat{\sigma}_3^2$, $\hat{\rho}_{12}$, $\hat{\rho}_{13}$ and $\hat{\rho}_{23}$ obtained from the observed data. The observed test statistic is then compared to pre-defined lower and upper stopping boundaries l_1 and u_1, which are determined by the expected information

I_1 at the first look, and either the trial is stopped, for futility or efficacy, or it continues to the next interim analysis. At each subsequent interim analysis, the test statistic $S_w = \hat{B}_w/\text{sd}(\hat{B}_w)$ is calculated in the same way as at the first analysis, using all available data on short-term and long-term outcomes, and compared to stopping boundaries u_w and l_w that determine whether the study is stopped early. If the trial is stopped early at an interim analysis, then long-term data will continue to be collected on all those recruited up to that point and these data will be used for final (definitive) inferences in an overrunning analysis [28].

The timing of the first and subsequent looks is typically specified at the commencement of the study via the selected values for N_{3w}, N_{2w} and N_{1w} at each early look w. These values are used, together with expected values of σ_3^2, ρ_{12}, ρ_{13} and ρ_{23}, to give the expected information I_w, at each planned early look w, using expression (2). The observed information $\hat{I} = 1/\text{var}(\hat{B})$ is monitored during data accrual, and interim analysis w occurs when the observed information equals the expected information at look w (see later worked example).

Sequential stopping boundaries

We are interested in a sequential trial with two short-term endpoints where a series of W interim analyses (looks) are undertaken to compare the two groups. The number of study participants increases in the two groups, and thus the long-term and short-term data available for analysis also increases through the course of the trial. Tests are performed at each of a series of interim analyses in order to make inferences about the superiority of the active intervention group (over the control) in terms of the long-term endpoint. The tests are undertaken at interim analysis w, using test statistic S_w, and must control the type I error rate across the W interim analyses. For a one-sided alternative at overall level α, with possible stopping for futility, the type I error rate spent is such that $\alpha_U^*(1) < \cdots < \alpha_U^*(W) = \alpha$ and $\alpha_L^*(1) < \cdots < \alpha_L^*(W) = 1 - \alpha$, where $\alpha_U^*(w)$ is the probability of stopping and rejecting H_0 in favour of $B > 0$ at look w (efficacy) and $\alpha_L^*(w)$ is the probability of stopping without rejecting H_0 at look w (futility). The type I error rates spent is determined by $\alpha_U^*(w)$ and $\alpha_L^*(w)$, which are specified in advance of the study begining. Stallard [9] proposes a method for construction of stopping boundaries in this scenario for the more general setting of T intervention arms, and a single control arm. For a two-arm study, standard group sequential methods and widely available software allow one to calculate the lower and upper stopping boundaries (l_w and u_w) at each look w [29].

Simulations

The statistical methodology described here provides a framework for how decisions about early stopping will be made. In order to understand how our assumptions about the likely size of the treatment effect, settings for nuisance parameters and the number of planned interim analysis affects design characteristics (e.g. how often we stop early for futility), we simulate data from the full multivariate distribution of outcomes $(X_{ij1}, \cdots, X_{ijK})$ for each of the i study participants and undertake

interim and final analyses many times. A Poisson model [30] is used to simulate the likely pattern of participant recruitment into the study. A constant monthly recruitment rate at each centre is assumed, with a smooth increase up to the target number of centres during the first 6 months of the planned 24 months of recruitment. The pattern of follow-up data collection at 3m, 6m and 12m is assumed to mirror that for recruitment. The timing of the interim looks are set at the start of a study using selected (feasible) values for N_3 and, based on the expected patterns of early data accrual, N_2 and N_1. These together with the expected values of ρ_{12}, ρ_{13}, ρ_{23} and σ_3 determine the expected information content of the data at each look $I_w = 1/\text{var}(B_w)$, using expression (2). The pre-specified stopping boundaries follow directly from I_w, α_L^* and α_U^*. The temporal pattern of participant recruitment, data collection and ultimately information are simulated for a single realisation of the study. For each simulation, a series of estimates for ρ_{12}, ρ_{13}, ρ_{23} and σ_3 are calculated using progressively increasing amounts of data as each new participant is recruited into the study. The pattern of (simulated) information accrual follows from these estimates, and the temporal pattern of data collection, using expression (2). Interim looks at the data occur when the simulated information is equal to the information content at the pre-specified stopping boundaries. The estimated test statistics are compared to stopping boundaries, with decisions on stopping following directly from these. Thus, the simulations emulate how the study would have evolved, and how decisions about stopping would have been made in a manner as close to a real life setting as we can feasibly create. Undertaking these simulated analyses many times allows us to estimate expected stopping probabilities and overall power (to reject the null hypothesis) that inform our decisions about the overall study design.

Results

Recruitment and data accrual

Simulating data from the recruitment model suggested that within the window of opportunity for early stopping (between 18 and 24 months from commencement of recruitment), 12m data will be available from between 15 and 40 participants per intervention arm (N_3). Figure 1 shows the expected patterns of recruitment, data and information acrrual during follow-up for our chosen correlation model $\rho_{12} = \rho_{13} = \rho_{23} = 0.5$, obtained from the simulations. The figure also shows information accrual (i.e. $1/\text{var}(B)$) for two extreme scenarios where (i) $\rho_{12} = \rho_{13} = \rho_{23} = 0$ and (ii) where $\rho_{12} = \rho_{13} = 0$ and $\rho_{23} = 1$, that represent the patterns of accrual when the early outcomes (3m and 6m) provide no information on the final 12m outcome and when the 6m outcome is exactly the same as the 12m outcome. In these two scenarios the pattern of information accrual are for scenario (i) exactly as would be observed if 12m outcome only provided all the relevant information, and in scenario (ii) exactly as would be observed if all the information was provided by the 6m data alone. For the purposes of motivating the simulations, it is useful to divide the likely recruitment numbers available in the window of opportunity for early stopping interval (a period of 6 months) equally. Figure 1 indicates that the likely pattern of data accrual at six potential interim looks for 12m, 6m and

3m data to be approximately as follows; at the first possible look $N_3 = 15$, $N_2 = 35$ and $N_1 = 50$, at the second look $N_3 = 20$, $N_2 = 40$ and $N_1 = 55$, at the third look $N_3 = 25$, $N_2 = 45$ and $N_1 = 60$, at the fourth look $N_3 = 30$, $N_2 = 50$ and $N_1 = 65$, at the fifth look $N_3 = 35$, $N_2 = 55$ and $N_1 = 70$ and at the sixth look $N_3 = 40$, $N_2 = 60$ and $N_1 = 75$. Under the expected correlation model $\rho_{12} = \rho_{13} = \rho_{23} = 0.5$ and expected standard deviation of the 12m outcome ($\sigma_1 = 20$), the information at each of these possible looks at the data is 21.4%, 28.0%, 34.4%, 40.8%, 47.1% and 53.3% , expressed as a percentage of the expected information at the study endpoint given by $N/2\sigma_3^2 = 85/800 = 0.106$. If $\rho_{12} = \rho_{13} = \rho_{23} = 0$, then this reduces to 17.6%, 23.5%, 29.4%, 35.3%, 41.2% and 47.1%; a correlation of 0 implies there is no information, on 12m outcomes, from the early 3m and 6m outcomes.

Type I error rate

As a prelude to simulations exploring overall study power and as a check of the software implementation, a number of simulations were undertaken to explore study characteristics under the null hypothesis (no treatment effect). The results of these simulations, for a selection of three likely data accrual patterns, are shown in Table 1. It is apparent from Table 1 that the estimated type I error rates for the three selected settings (i) one early look $N_1 = 60, N_2 = 45, N_3 = 25$, (ii) two early looks $N_1 = (55, 70), N_2 = (40, 55), N_3 = (20, 35)$ and (iii) three early looks $N_1 = (50, 65, 75), N_2 = (35, 50, 60), N_3 = (15, 30, 40)$ are well controlled at the 2.5% level. Also, the estimated cumulative probabilities of stopping for futility at early looks $p_{w,F}$ are equal (within simulation error) to the pre-specified lower error spending values, α_L^*.

Power

Overall study power and stopping probabilities were estimated for a range of plausible 12m treatment differences for the C-M score scale (0, 2.5, 5, 7.5 and 10); these corresponded to standardized effect sizes, for the selected value of $\sigma_Y = 20$, of 0, 0.125, 0.25, 0.375 and 0.5. A range of values for the lower bounds α_L^* were tested for one, two and three early looks at the data, using the same values for N, N_3, N_1 and N_2 as described above for type I error rate estimation, using the uniform correlation model ($\rho = \rho_{13} = \rho_{23} = \rho_{12}$) with a value of $\rho = 0.5$. Efficacy stopping boundaries were set to $\alpha_U^* = (0.001, 0.025)$, $\alpha_U^* = (0, 0.001, 0.025)$ and $\alpha_U^* = (0, 0, 0.001, 0.025)$, at one, two and three early looks respectively. The main initial clinical focus of our design is to determine whether the balloon procedure is ineffective or harmful. Therefore, the emphasis in the simulations, and the planned designs, will be on early stopping for futility, which is determined by α_L^*. The chosen settings for the upper (efficacy) boundaries α_U^* favour collecting as much information as possible if there is emerging evidence of efficacy. Early stopping for efficacy will only be considered at the last interim look, with boundaries set such that only if there is very strong evidence that the balloon procedure is superior to standard care will early stopping be considered. Figure 2 shows results for one early look at the data, Figure 3 for two early looks at the data and Figure 4 for three early looks at the

data.

There are strong trends for increasing power as the treatment difference increases from 0 to 10 points on the C-M score scale and corresponding decreases in the futility stopping probabilities. Estimates for early stopping for efficacy from the simulations, which were planned for the last of the interim looks only, increased from approximately 10% for one early look to 20% for two early looks and 25% for three early looks, for a treatment difference of 10 points. This was due to more data being available at the look when stopping for efficacy can occur ($n = 15$ for one look, $n = 35$ for two looks and $n = 40$ for three looks).

Four options for futility stopping were investigated for α_L^* that represented a sequence of increasingly agressive options, from a low probability of stopping, labelled as (a), to a high probability, labelled as (d), with (b) and (c) intermediate to these. For one early look at the data α_L^* was set to either (a) $(0.24, 0.975)$, (b) $(0.48, 0.975)$, (c) $(0.72, 0.975)$ or (d) $(0.96, 0.975)$, for two early looks to either (a) $(0.08, 0.24, 0.975)$, (b) $(0.16, 0.48, 0.975)$, (c) $(0.24, 0.72, 0.975)$ or (d) $(0.32, 0.96, 0.975)$ and for three early looks to either (a) $(0.08, 0.16, 0.24, 0.975)$, (b) $(0.16, 0.32, 0.48, 0.975)$, (c) $(0.24, 0.48, 0.72, 0.975)$ or (d) $(0.32, 0.64, 0.96, 0.975)$.

Under the null hypothesis (C-M treatment difference equal to 0), α_L^* represented the expected stopping probabilities (for futility) at each look. For the largest treatment differences (10 on C-M score scale) and the most agressive stopping options, the futility stopping rates were 44.4% for one early look (Figure 2(d)), 31.9% for two early looks (Figure 3(d)) and 27.1% for three early looks (Figure 4(d))). For this most agressive futility stopping setting, study power was lowered significantly due to (incorrect) early stopping. Power was reduced to only 55.5%, 68.0% and 72.7%, in these three settings, rather than the 90% we would expect for a non-adaptive design. The least agressive futility stopping option (Figures 2(a), 3(a) and 4(a)) showed good power (89.5%, 89.7% and 89.7%), but poor early stopping under the null hypothesis (24.3%, 25.1% and 26.7%). The two extreme futility stopping options (Figures 2(a,d) 3(a,d) and 4(a,d)), therefore, do not have the characteristics we are seeking in the design.

The intermediate options (Figures 2(b,c) 3(b,c) and 4(b,c)), however, have more desirable characteristics as they have reasonable power for a strong treatment effect (C-M treatment difference of 10) whilst retaining the ability to stop early for futility, with high probability, under the null hypothesis. For example, for two early looks when $\alpha_L^* = (0.24, 0.72, 0.975)$ (Figure 3(c)), overall power was 87.6% for a treatment difference of 10, with a stopping rate of 24.5% at the first look and 72.9% at the first or second look combined.

The expected sample size (ESS), calculated from the expected stopping probabilities and expected pattern of patient and data accrual, provides a useful summary of the design characteristics that complements study power. The right hand y-axes of Figures 2, 3 and 4 are annotated to provide a useful informal comparator to the fixed study design with a sample size of 170; this provides 90% power to detect a C-M score treatment difference of 10 points between intervention arms, at the 5% level. The ESS decreases, for all numbers of early looks, from the least (a) to the most aggressive (d) futility stopping options; increasing the probability of stopping early, for either futility and efficacy, lowers the overall study sample size from that

we would need for the non-adaptive (fixed) study design (sample size; $2N = 170$). The pattern of variation for ESS, across treatment differences, reflects the dominance of either futility stopping (for zero and small differences) or efficacy (for large differences). In selecting a good design, we aim to find settings of the stopping boundaries that maintain overall power at or as close as possible to the nominal (non-adaptive) 90% level, whilst at the same time lowering the expected sample size across the range of treatment effects we might expect to see in the study.

The number of study participants required to reach the required information levels at the early looks was also assessed in the simulations. The expected (mean) numbers were very close to the sample sizes used to motivate the simulations, as we would expect; i.e. $N_3 = 25$ for one early look, $N_3 = (20, 35)$ for two early looks and $N_3 = (15, 30, 40)$ for three early looks. The simulations were set-up such that early looks at the data took place, even if recruitment had been completed; whereas in reality, the early looks would have been abandoned. Recruitment had been completed at the final early look at the data for (approximately), 0%, 3% and 12% of the simulations for one, two and three early looks. The high value for three early looks reflects the fact that the final early look at the data occurs when approximately 40 participants in each arm of the study have 12m outcome data, which is quite close to 50, the point when the recruitment model expects that recruitment will have completed.

Worked example

In order to illustrate how the design will work in practice we briefly work through the necessary calculations, using purely synthetic data, for a much smaller and simpler example than those used in the simulations. The data and R code [31] for implementation are provided in the supplementary material.

A study is planned with $\alpha_L^* = (0.200, 0.600, 0.975)$ and $\alpha_U^* = (0.000, 0.001, 0.025)$ for two early looks, with group sample sizes of $N_3 = (10, 15)$, $N_2 = (15, 20)$, $N_1 = (20, 25)$ and $N = 30$; we assume equal group sizes, and two early outcomes and a final outcome as previously, for ease of exposition. Let us suppose that data available from a pilot study suggest correlations between outcomes of $\rho_{13} = \rho_{23} = 0.5$ and $\rho_{12} = 0$, with $\sigma_3 = 18$. Using these values in expression (2), indicates that the expected information at the early looks will be $I_1 = 0.019$ and $I_1 = 0.028$, and at the final analysis $I_{\text{Final}} = N/2\sigma_Y^2 = 30/648 = 0.046$ (for $\sigma_Y = 18$); expressed as a percentage of the information available at the final analysis, this corresponds to 42% and 60%, for the two early looks. The boundaries can be calculated using widely available software, for instance the `gsDesign` [32] package in R. For our selected values for α_L^* and α_U^*, and the expected information at our planned looks, the function `gsBound` provides the following boundaries for decision making; at look 1, $l_1 = -0.842$ (lower boundary) and $u_1 = \infty$ (upper boundary), at look 2 $l_2 = 0.247$ and $u_2 = 3.09$, and at the final analysis $l_{\text{Final}} = u_{\text{Final}} = 1.96$.

Data collection proceeded as planned, with information monitored during follow-up. After the twentieth participant had provided final outcome data the estimated information (0.02) reached the pre-set value for the first look (0.019). Figure 5 shows the distributions of outcome data at the first look. The estimate of the mean

treatment difference (in favour of the test group) for the final outcome (X_3) was -10.2; i.e. the outcome score for the test intervention was considerably lower than the control intervention. Estimates of the correlations between outcomes and the standard deviation of the final outcome were as follows; $\hat{\rho_{13}} = 0.45$, $\hat{\rho_{23}} = 0.20$, $\hat{\rho_{12}} = 0.04$ and $\hat{\sigma}_3 = 16.8$. Calculating B and var(B) (using equations 1 and 2), provides estimates of the mean treatment difference for the outcome of -9.77, with variance 50.18 (see Supplementary material). Therefore, the test statistic at look 1, $S_1 = -1.38$, is less than the lower boundary (-0.842) indicating that the study should be *stopped* for futility.

Continuing to follow-up all those in the study, after the desicison to stop at look 1, in an *overrunning* analysis [28] provides estimates of $B = -3.70$ and var(B) = 20.5 ($p = 0.419$). Confirming that the decision made to stop at look 1 appears to have been correct, and leads us to conclude that there is no evidence that the test group performs better than the control group.

If different settings for α_L^* had been selected then the study may have proceeded in a different manner. For instance, if a less agressive lower stopping criterion had been used at the first look (e.g. $\alpha_L^* = (0.080, 0.600, 0.975)$), then the lower boundary at the first look would be $l_1 = -1.41$, and the study would not have stopped for futility.

Discussion

This manuscript describes work to develop an adaptive clinical trial design motivated by a trial for testing a novel surgical approach for repair of rotator cuff tendon tears. The design, that builds and expands on previously published methodology [9, 11], uses early observations of the primary outcome at 3m and 6m to augment 12m outcome data to inform decision making on early stopping. The main focus in the development of the design is on *futility* stopping, rather than *efficacy* stopping; i.e. stopping for *efficacy* in the simulations is limited to the last interim look at the data and is such that very strong evidence is required to stop. This reflects the clinical perspective, that if a new intervention shows promise then it is prudent, within reason, to continue to collect data to the planned study sample size, rather than stop early, in order to provide more precise effect estimates and increase the chances of detecting any adverse events.

The simulations showed that with more looks at the data the chance of recruitment completing before the final look increased; recruitment completed before the final look in 3% and 12% of simulations for two and three early looks. More looks offer more possibilities for early decision making, but at a greater risk of not completing the planned early looks before the end of recruitment. The estimated rates of recruitment completing before the last early look, are clearly in part at least dependent on the veracity of the recruitment model. If recruitment was much higher or faster than expected at times during recruitment, then this could be problematic for the design. For instance, a rapid unexpected rise in the recruitment rate could cause recruitment to be completed before the early looks at the data had happened. We do not think this will happen in our setting, as there are structural (study-based) limitations in the number of centres, clinicians and timings of clinics which make this highly unlikely. However, recruitment will be monitored closely. In

the START:REACTS study it is likely that early looks will be dropped if recruitment completes much more rapidly than expected. However, it may be desirable in other settings to close centres or temporarily suspend recruitment if this were feasible.

As with conventional sample size calculations, the results of the simulations are dependent on assumptions made about the variance of the primary outcome (12m C-M score) and the correlations between the early 3m and 6m and 12m scores. We have good evidence on these *nuisance* parameters from a recently published systematic review [24] and relevant data [26]. A larger than expected value has been deliberately selected for the 12m C-M score standard deviation ($\sigma_3 = 20$); close inspection of the data from [24] suggest that the standard deviation is likely to be nearer to 15, than 20. Conservatively, a value of 20 was chosen for the simlations. If σ_3 is lower than 20, then we will reach the planned study information points, that determine the timings of the early looks at the data, sooner than the simulations indicate.

The simulations assume a relatively moderate correlation model for the study outcomes; $\rho_{13} = \rho_{23} = \rho_{12} = 0.5$. If the correlation model was stronger than expected (e.g. $\rho_{13} = \rho_{23} = \rho_{12} = 0.9$), and all other things were unchanged, we would reach the information thresholds for the early looks sooner than planned (i.e. with fewer participants) and potentially gain more from the adaptive design than we estimate from the simulations. Conversely, if the correlations are such that the early outcomes tell us nothing about the definitive outcome (i.e. $\rho_{13} = \rho_{23} = \rho_{12} = 0$), then we would acculmulate information more slowly than the simulations suggest and recruitment is likely to have completed before the information required for the first look at the data is reached. In such a setting the design would proceed to the fixed recruitment target, in the conventional manner. The *loss* in such a setting would be the increase in sample size, relative to the fixed design, that we would need for the adaptive design, For example, for the START:REACTS study described previously, the sample size would need to increase from 170 participants to between 180 and 188, dependent on the choice of boundaries and early looks. A relatively modest increase in sample size, given the potential gains from early stopping, for this study, but in other application areas this may be an unacceptable increase in sample size if there is little evidence for even moderate associations between the early and final study outcomes.

The simulations show that the error rate is controlled at the specified rate, provided that the stopping rules are *binding* [33]. Where by *binding*, we mean that stopping for futility at the early look is *essential* whenever the futility boundaries are crossed; irrespective, for instance, of reasons external to the study, such as new or emerging evidence on the interventions. The simulations show study power based on a sample size of 170 (85 in each group). This provided 90% power for the non-adaptive design. For the adaptive designs with appealing operating characteristics discussed here the power is somewhat lower than 90%. For the definitive adaptive study design the overall sample size will be increased to provide 90% power. The final selection of overall sample size, stopping boundaries and number of looks will be made by the START:REACTS data monitoring and safety committee (DSMC) and confirmed by the trial steering committee (TSC). The boundaries, timings of

the interim looks and agreement on binding will be incorporated into the DSMC charter and will be kept confidential within the study team.

The work described here is focussed primarily on the design of the START:REACTS study, and this is reflected in the set-up of the simulations and data generating model. For instance, we have assumed that the correlations between the outcomes are the same within the intervention arms. This need not be the case in other applications, and it would be relatively straightforward to modify the set-up of the simulations to allow different correlations in the intervention arms or different variances for each of the early outcomes. We believe that the designs discussed will have much wider application in many analogous settings particular where trials are undertaken to assess new surgical and other interventions where outcomes are assessed over a long period of time. Typically in studies of this type designs are non-adaptive, and early outcomes, usually available as part of routine monitoring of patients, are simply reported as secondary outcomes. This is both inefficient and wasteful. With increased methodological understanding and availability and ease of use of software tools for implementing adaptive designs, we believe that this situation will change in the future.

Conclusion

In this manuscript we present a methodology for the design of an adaptive clinical trial motivated by testing a novel surgical approach for repair of rotator cuff tendon tears. The design uses early observations of the 12m primary outcome at 3m and 6m to augment 12m outcome data to inform decision making on early stopping. We derive estimators for the treatment effect and test statistics based on the setting of two early outcomes, and present methods for estimation of sequential stopping boundaries. Simulations are undertaken for one, two and three early looks with a range of options for stopping boundaries. We show that a design with two early looks is feasible and, with appropriately chosen futility stopping boundaries, has appealing design characteristics. A number of possible design options are described that have good power, and have high probability of stopping for futility if there is no evidence of a treatment effect at early looks. A worked example provides a practical demonstration of how the design might work in a real study. In summary, the work shows that an adaptive design is feasible and could work in practice, and provides some guidelines for appropriate values for the stopping boundaries for the START:REACTS study.

Abbreviations

C-M: Constant-Murley shoulder score; DSMC: Data monitoring and safety committee; EME: (UK) Efficacy and Mechanism Evaluation programme; ESS: Expected sample size; MCID: Minimum clinically important difference; MRC; (UK) Medical Research Council; NHMRC: (Australian) National Health and Medical Research Council; NICE: (UK) National Institute for Health and Care Excellence; NIHR: (UK) National Institute of Health Research; RCT: Randomized controlled trial; START:REACTS: Sub-acromial spacer for Tears Affecting Rotator cuff Tendons: a Randomised, Efficient, Adaptive Clinical Trial in Surgery; TSC: Trial steering committee

Declarations

Ethics approval and consent to participate

Ethics approval is not applicable for this manuscript as it describes methodological development work, and uses simulated data only.

Consent for publication

Not applicable.

Availability of data and material
The datasets used and analysed and code written as part of this study are available from the corresponding author on reasonable request.

Competing interests
All authors have previously received or are currently in receipt of funding from the NIHR. MU was Chair of the NICE accreditation advisory committee until March 2017 for which he received a fee. MU is also chief investigator or co-investigator on multiple previous and current research grants from the NIHR, Arthritis Research UK and is a co-investigator on grants funded by the Australian NHMRC, and an NIHR Senior Investigator. MU has received travel expenses for speaking at conferences from the professional organisations hosting the conferences, and is a director and shareholder of Clinvivo Ltd that provides electronic data collection for health services research. MU is part of an academic partnership with Serco Ltd related to return to work initiatives, an editor of the NIHR journal series, and a member of the NIHR Journal Editors Group, for which he receives a fee. The makers of the InSpace balloon, Orthospace Ltd, have had no involvement in the conception, design, conduct or analysis of this work, and have had no involvement in preparing the manuscript except to check its content for intellectual property transgressions. They are providing 50 free balloons for the trial and have provided training for surgeons in using the device, but have no other involvement in the trial, the full independence of the trial team is clearly laid out contractually in line with standard NIHR terms

Funding
The work reported here is funded by the Efficacy and Mechanism Evaluation (EME) Programme, an MRC and NIHR partnership. The funding body had no other role in the work reported here and played no part in writing the manuscript. The views expressed in this publication are those of the authors and not necessarily those of the funding body or the MRC, NIHR or the UK Department of Health and Social Care.

Author's contributions
NS and NP developed the methods, NP conducted the simulations and exemplary applications and was the major contributor in writing the manuscript. AM, HP, PW, JM and MU critically reviewed, discussed and adapted the methodology. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Author details
[1] Statistics and Epidemiology Unit, Warwick Medical School, University of Warwick, CV4 7AL Coventry, UK. [2] Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, CV4 7AL Coventry, UK. [3] University Hospital Coventry and Warwickshire, CV2 2DX Coventry, UK.

References
1. McCulloch, P., Cook, J.A., Altman, D.G., Heneghan, C., Diener, M.K., Group, I.: Ideal framework for surgical innovation 1: the idea and development stages. BMJ **346**, 3012 (2013)
2. Ergina, P.L., Barkun, J.S., McCulloch, P., Cook, J.A., Altman, D.G., Group, I.: Ideal framework for surgical innovation 2: observational studies in the exploration and assessment stages. BMJ **346**, 3011 (2013)
3. Cook, J.A., McCulloch, P., Blazeby, J.M., Beard, D.J., Marinac-Dabic, D., Sedrakyan, A., Group, I.: Ideal framework for surgical innovation 3: randomised controlled trials in the assessment stage and evaluations in the long term study stage. BMJ **346**, 2820 (2013)
4. Bauer, P., Bretz, F., Dragalin, V., Konig, F., Wassmer, G.: Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. Stat Med **35**(3), 325–47 (2016)
5. Pallmann, P., Bedding, A.W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L.V., Holmes, J., Mander, A.P., Odondi, L., Sydes, M.R., Villar, S.S., Wason, J.M.S., Weir, C.J., Wheeler, G.M., Yap, C., Jaki, T.: Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med **16**(1), 29 (2018)
6. Chow, S.C., Corey, R.: Benefits, challenges and obstacles of adaptive clinical trial designs. Orphanet J Rare Dis **6**, 79 (2011)
7. Kairalla, J.A., Coffey, C.S., Thomann, M.A., Muller, K.E.: Adaptive trial designs: a review of barriers and opportunities. Trials **13**, 145 (2012)
8. Bauer, P., Brannath, W.: The advantages and disadvantages of adaptive designs for clinical trials. Drug Discov Today **9**(8), 351–7 (2004)
9. Stallard, N.: A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Stat Med **29**(9), 959–71 (2010)
10. National Institute for Health Research. https://www.nihr.ac.uk/
11. Galbraith, S., Marschner, I.C.: Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. Stat Med **22**(11), 1787–805 (2003)
12. Kukkonen, J., Joukainen, A., Lehtinen, J., Mattila, K., Tuominen, E., T, T.K., Aarimaa, V.: Treatment of non-traumatic rotator cuff tears: a randomised controlled trial with one-year clinical results. Bone and Joint Journal **96**(1), 75–81 (2014)
13. Rangan, A., Upadhaya, S., Regan, S., Toye, F., Rees, J.L.: Research priorities for shoulder surgery: results of the 2015 james lind alliance patient and clinician priority setting partnership. BMJ Open **6**(4), 010412 (2016)
14. OrthoSpace Massive Rotator Cuff Repair. http://orthospace.co.il/
15. Burks, R.T., Crim, J., Brown, N., Fink, B., Greis, P.E.: A prospective randomized clinical trial comparing arthroscopic single- and double-row rotator cuff repair: magnetic resonance imaging and early clinical evaluation. Am J Sports Med **37**(4), 674–82 (2009)
16. START:REACTS. https://warwick.ac.uk/fac/sci/med/research/ctu/trials/startreacts/

17. ISRCTN Registry. http://www.isrctn.com/ISRCTN17825590
18. Constant, C., Murley, A.: A clinical method of functional assessment of the shoulder. Clinical Orthopaedics and Related Research **214**, 260–164 (1987)
19. Constant, C., Gerber, C., Emery, R., Sojbjerg, J., Gohlke, F., Boileau, P.: A review of the constant score: modifications and guidelines for its use. Journal of Shoulder and Elbow Surgery **17**(2), 355–361 (2008)
20. Roy, J., MacDermid, J., Woodhouse, L.: A systematic review of the psychometric properties of the constant-murley score. Journal of Shoulder and Elbow Surgery **19**(1), 157–164 (2010)
21. Blonna, D., Scelsi, M., Bellato, E.M.E., Tellini, A., Rossi, R., Bonasia, D., Castoldi, F.: Can we improve the reliability of the constant-murley score? Journal of Shoulder and Elbow Surgery **21**(1), 4–12 (2012)
22. Ban, I., Troelsen, A., Christiansen, D., Svendsen, S., Kristensen, M.: Standardised test protocol (constant score) for evaluation of functionality in patients with shoulder disorders. Danish Medical Journal **60**(4), 4608 (2013)
23. Christiansen, D., Frost, P., Falla, D., Haahr, J., Frich, L., Svendsen, S.: Responsiveness and minimal clinically important change: comparison between 2 shoulder outcome measures. Journal of Orthopaedic and Sports Physical Therapy **45**(8), 620–625 (2015)
24. Khatri, C., Ahmed, I., Parsons, H., Smith, N., Lawrence, T., Modi, C., Drew, S., Bhabra, G., Parsons, N., Underwood, M., Metcalfe, A.: The natural history of full-thickness rotator cuff tears in randomized controlled trials: a systematic review and meta-analysis. The American Journal of Sports Medicine **1**(1), 1–1 (2018)
25. Haahra, J., Ostergaard, S., Dalsgaard, J., Norup, K., Frost, P., Lausen, S., Holm, E., Andersen, J.: Exercises versus arthroscopic decompression in patients with subacromial impingement: a randomised, controlled study in 90 cases with a one year follow up. Annals of the Rheumatic Diseases **64**(5), 760–764 (2005)
26. Karthikeyan, S., Kwong, H., Upadhyay, P., Parsons, N., Drew, S., Griffin, D.: A double blind randomised controlled study comparing subacromial nsaid (tenoxicam) injection with steroid (methylprednisolone) injection in patients with subacromial impingement syndrome. Journal of Bone and Joint Surgery (British) **92**(1), 77–82 (2010)
27. Senekovic, V., Poberaj, B., Kovacic, L., Mikek, M., Adar, E., Dekel, A.: Prospective clinical study of a novel biodegradable sub-acromial spacer in treatment of massive irreparable rotator cuff tears. European Journal of Orthopaedic Surgery and Traumatology **23**(3), 311–316 (2013)
28. Whitehead, J.: Overrunning and underrunning in sequential clinical trials. Controlled Clinical Trials **13**, 106–121 (1992)
29. Jennison, C., Turnbull, B.W.: Group Sequential Methods with Applications to Clinical Trials, p. 390. Chapman & Hall/CRC, Boca Raton (2000)
30. Barnard, K., Dent, L., Cook, A.: A systematic review of models to predict recruitment to multicentre clinical trials. BMC Medical Research Methodology **10**(63) (2010)
31. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/
32. Anderson, K.: gsDesign: Group Sequential Design. (2016). R package version 3.0-1. https://CRAN.R-project.org/package=gsDesign
33. Bretz, F., Koenig, F., Brannath, W., Glimm, E., Posch, M.: Adaptive designs for confirmatory clinical trials. Stat Med **28**(8), 1181–217 (2009)

Tables

Table 1 Estimated type I error rates, where $p_{w,F}$ is the cumulative probability of stopping for futility at look w or earlier, p_E is the probability of stopping early for efficacy and p_{12m} is the probability of stopping for efficacy at the end of the study; $N = 85$, for (a) one look $N_1 = 60, N_2 = 45, N_3 = 25$, (b) two looks $N_1 = (55, 70), N_2 = (40, 55), N_3 = (20, 35)$ and (c) three looks , $N_1 = (50, 65, 75), N_2 = (35, 50, 60), N_3 = (15, 30, 40)$, $\rho = \rho_{13} = \rho_{23} = \rho_{12}$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 20$ (10,000 simulations).

Futility bound (α_L^*)	ρ	p_E	$p_{1,F}$	$p_{2,F}$	$p_{3,F}$	p_{12m}
(a) One look; $\alpha_U^* = (0.001, 0.025)$						
$(0.0, 0.975)$	0.0	0.002	0.000	-	-	0.025
$(0.5, 0.975)$	0.0	0.002	0.504	-	-	0.023
$(0.0, 0.975)$	0.5	0.002	0.000	-	-	0.026
$(0.5, 0.975)$	0.5	0.002	0.504	-	-	0.026
(b) Two looks; $\alpha_U^* = (0, 0.001, 0.025)$						
$(0.0, 0.0, 0.975)$	0.0	0.001	0.000	0.000	-	0.025
$(0.2, 0.5, 0.975)$	0.0	0.001	0.202	0.499	-	0.025
$(0.0, 0.0, 0.975)$	0.5	0.001	0.000	0.000	-	0.024
$(0.2, 0.5, 0.975)$	0.5	0.002	0.199	0.505	-	0.025
(c) Three looks; $\alpha_U^* = (0, 0, 0.001, 0.025)$						
$(0.0, 0.0, 0.0, 0.975)$	0.0	0.001	0.000	0.000	0.000	0.024
$(0.1, 0.3, 0.5, 0.975)$	0.0	0.002	0.110	0.306	0.503	0.025
$(0.0, 0.0, 0.0, 0.975)$	0.5	0.001	0.000	0.000	0.000	0.025
$(0.1, 0.3, 0.5, 0.975)$	0.5	0.001	0.108	0.307	0.506	0.025

Figures

Figure 1 Recruitment, data and information accrual during follow-up. Expected recruitment, data and information accrual during 24 months, estimated from simulations. Information accrual is plotted for three possible correlation models; $\rho_{12} = \rho_{13} = \rho_{23} = 0.5$, $\rho_{12} = \rho_{13} = \rho_{23} = 0$ and, $\rho_{12} = \rho_{13} = 0$ and $\rho_{23} = 1$.

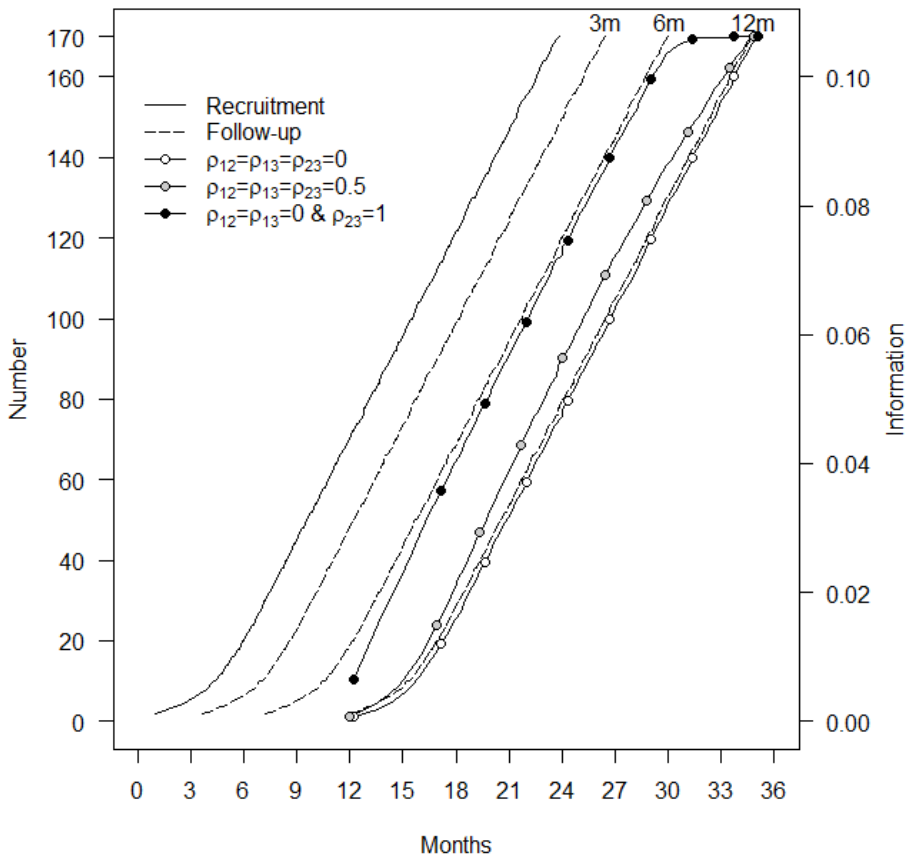


Figure 2 Design characteristics for one early look. Estimated probabilities of stopping for *futility* and *efficacy* at the first look, expected sample size (ESS) and overall study power, for effect sizes in range 0 to 10 for (a) $\alpha_L^* = (0.24, 0.975)$, (b) $\alpha_L^* = (0.48, 0.975)$, (c) $\alpha_L^* = (0.72, 0.975)$ and (d) $\alpha_L^* = (0.96, 0.975)$. Where $\alpha_U^*(1) = 0.001$, $\rho = 0.5$ and other settings are as Table 1.

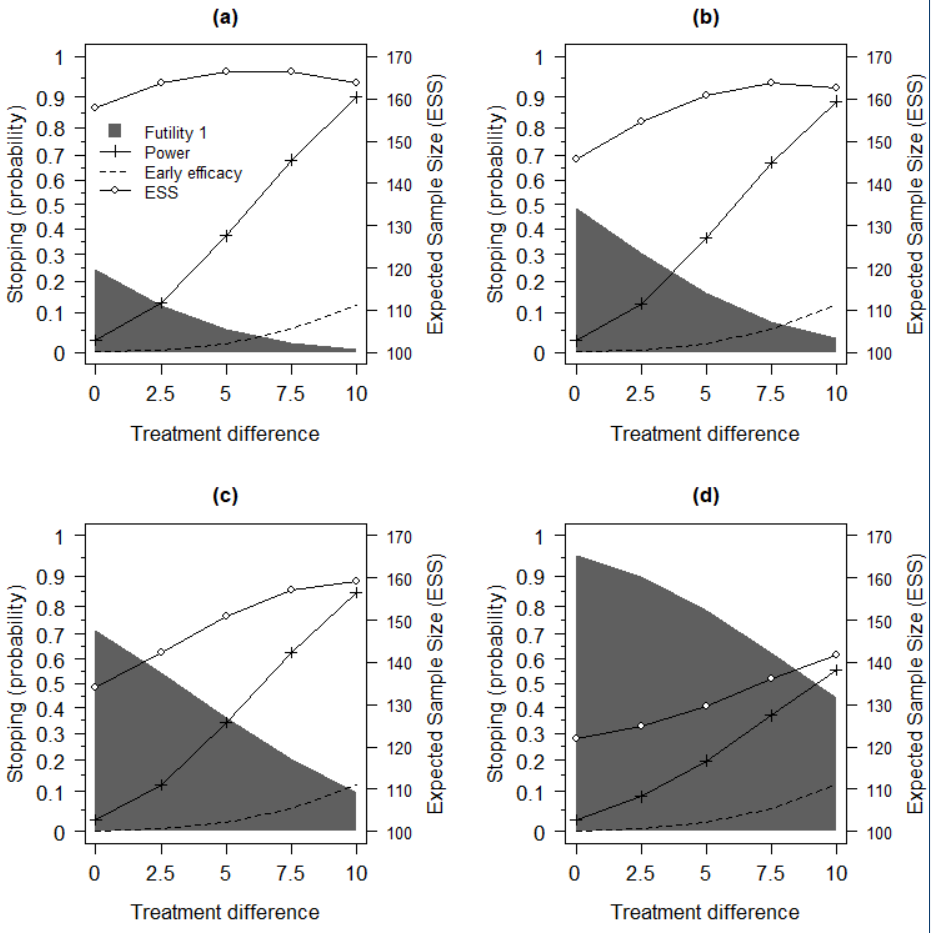


Figure 3 Design characteristics for two early looks. Estimated probabilities of stopping for *futility* and *efficacy* at the first and second looks, expected sample size (ESS) and overall study power, for effect sizes in range 0 to 10 for (a) $\alpha_L^* = (0.08, 0.24, 0.975)$, (b) $\alpha_L^* = (0.16, 0.48, 0.975)$, (c) $\alpha_L^* = (0.24, 0.72, 0.975)$ and (d) $\alpha_L^* = (0.32, 0.96, 0.975)$. Where $\alpha_U^*(1) = 0$ and $\alpha_U^*(2) = 0.001$, $\rho = 0.5$ and other settings are as Table 1.

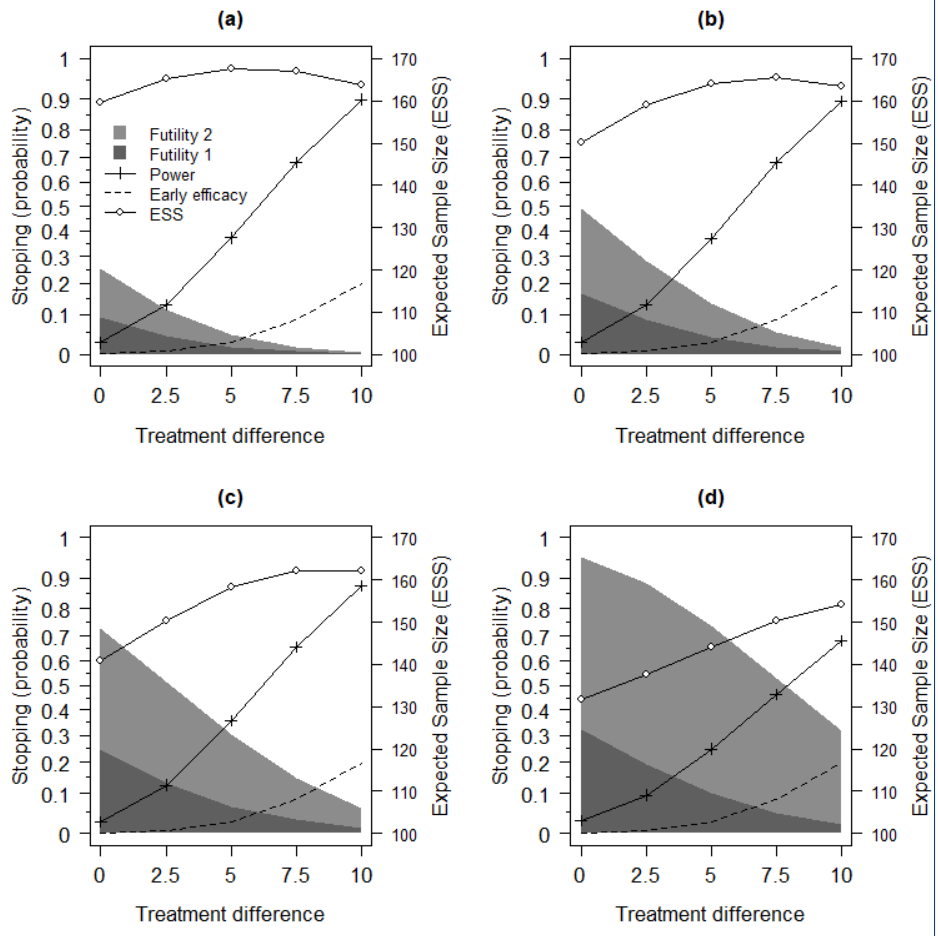


Figure 4 Design characteristics for three early looks. Estimated probabilities of stopping for *futility* and *efficacy* at the first, second and third looks, expected sample size (ESS) and overall study power, for effect sizes in range 0 to 10 for (a) $\alpha_L^* = (0.08, 0.16, 0.24, 0.975)$, (b) $\alpha_L^* = (0.16, 0.32, 0.48, 0.975)$, (c) $\alpha_L^* = (0.24, 0.48, 0.72, 0.975)$ and (d) $\alpha_L^* = (0.32, 0.64, 0.96, 0.975)$. Where $\alpha_U^*(1) = 0$, $\alpha_U^*(2) = 0$ and $\alpha_U^*(3) = 0.001$, $\rho = 0.5$ and other settings are as Table 1.

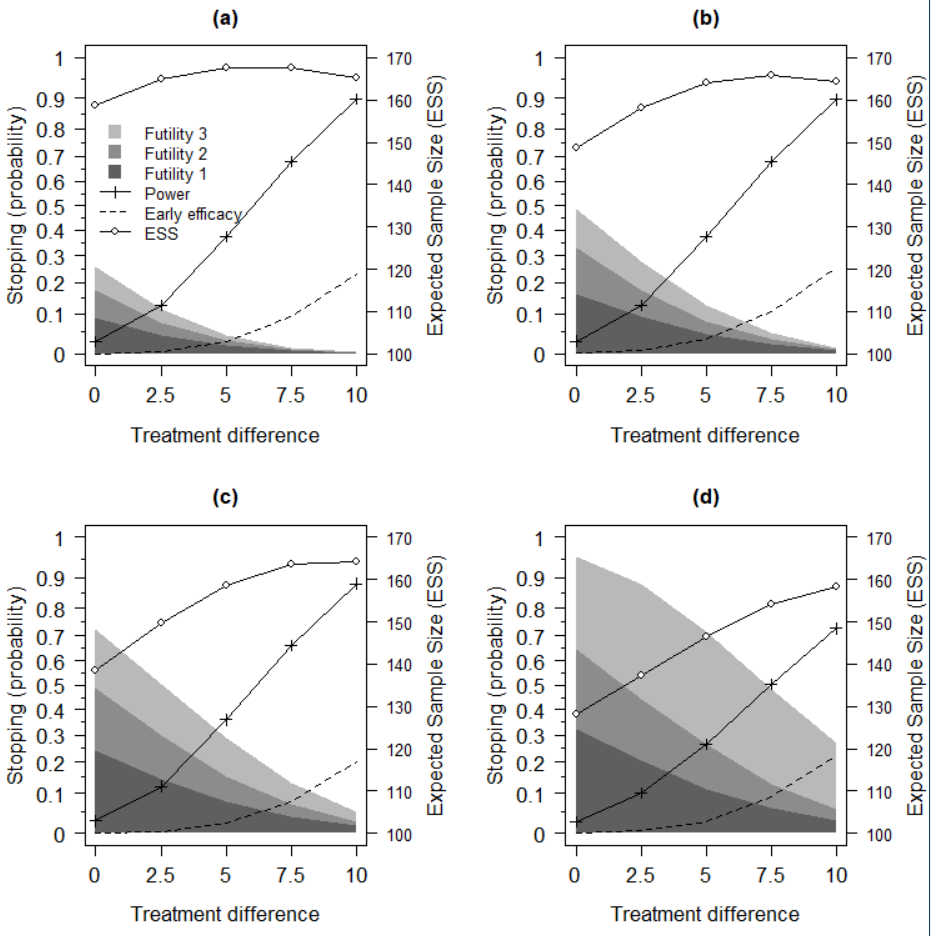
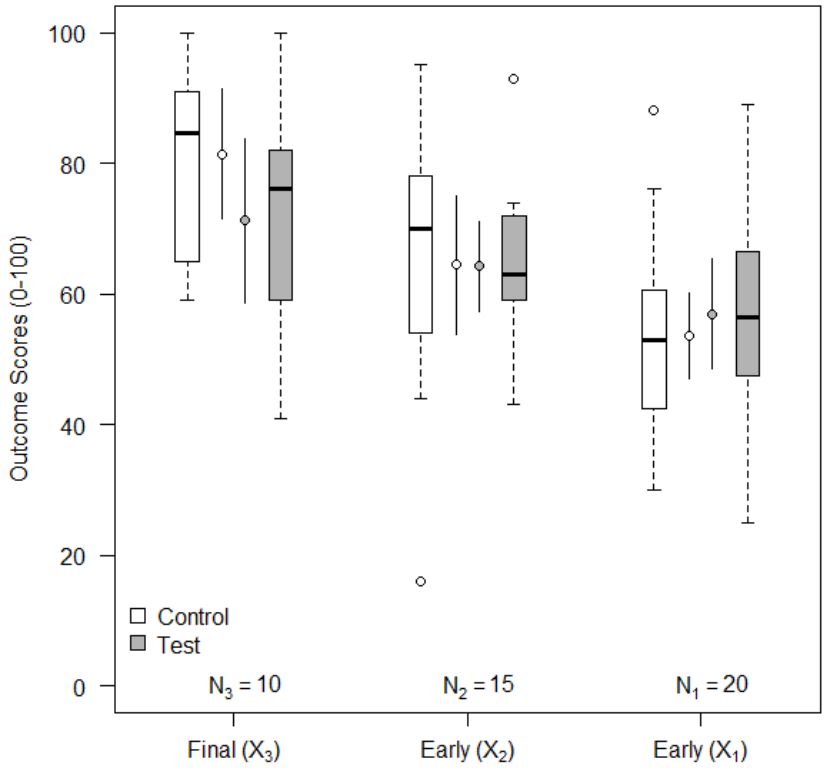


Figure 5 Outcome score data at the first look. Boxplots and means with 95% confidence intervals of early (X_1 and X_2) and final (X_3) outcome data by intervention group at the first interim analysis.



Supplementary material

Nick Parsons[1], Nigel Stallard[1], Helen Parsons[2], Philip Wells[2], Martin Underwood[2,3], James Mason[2] and Andrew Metcalfe[2,3]

B and var(B) for $K-1$ early outcomes

For a two-arm study, participants are randomized to either the control ($j=0$) or active intervention ($j=1$) arms, and at an interim analysis long-term (final) outcomes are available from N_K subjects and early (short-term) outcomes from $N_1 \dots N_{K-1}$ subjects in each arm of the study. Assuming equal numbers of subjects providing data for the earlier outcome X_{k-1} than the later outcome X_K, the effect of the test intervention on the long-term (primary) outcome X_K [1] is given by

$$B = \frac{1}{N_K}\left[\sum_{i=1}^{N_K}(X_{i1K} - X_{i0K}) + \sum_{k=1}^{K-1}\left[\rho_{kK}\frac{\sigma_K}{\sigma_k} \sum_{i=N_K+1}^{N_k}\left(X_{i1k} - X_{i0k} - \frac{1}{N_k}\sum_{m=1}^{N_k}(X_{m1k} - X_{m0k})\right)\right]\right], \tag{1}$$

with variance

$$\text{var}(B) = \frac{2\sigma_K^2}{N_K}\left[1 - \sum_{k=1}^{K-1}\left(\rho_{kK}^2\frac{N_k - N_K}{N_k}\right) + \sum_{k=1}^{K-2}\sum_{k'=k+1}^{K-1} 2\rho_{kK}\rho_{k'K}\rho_{kk'} \left(min(N_k, N_{k'})\frac{N_K}{N_k N_{k'}} + 1 - \frac{N_K}{N_k} - \frac{N_K}{N_{k'}}\right)\right]. \tag{2}$$

Estimates $\hat{B}$ and var$(\hat{B})$ follow from estimates of the correlations $\rho_{kk'}$, and standard deviations σ_k and σ_K, obtained from the appropriate regression models, using all available data.

B and var(B) for two early outcomes

For a two-arm study, participants are randomized to either the control ($j=0$) or active intervention ($j=1$) arms, and at an interim analysis long-term (final) outcomes are available from N_{j3} subjects and early (short-term) outcomes from N_{j1} and N_{j2} subjects in each arm of the study. Denoting the final outcome data for study participant i by X_{ij3}, and early outcomes by X_{ij1} and X_{ij2}, then following Galbraith and Marschner [1], the treatment effect B, which uses all the available early endpoint data, is given by

$$B = \eta_3 + \gamma_{13}\eta_1 + \gamma_{23}\eta_2,$$

where

$$\eta_3 = \frac{1}{N_{13}}\sum_{i=1}^{N_{13}} X_{i13} - \frac{1}{N_{03}}\sum_{i=1}^{N_{03}} X_{i03},$$

$$\eta_2 = \frac{1}{N_{13}}\sum_{i=N_{13}+1}^{N_{12}} X_{i12} - \frac{1}{N_{03}}\sum_{i=N_{03}+1}^{N_{02}} X_{i02} - \frac{N_{12}-N_{13}}{N_{13}}\left\{\frac{1}{N_{12}}\sum_{i=1}^{N_{12}} X_{i12}\right\} + \frac{N_{02}-N_{03}}{N_{03}}\left\{\frac{1}{N_{02}}\sum_{i=1}^{N_{02}} X_{i02}\right\},$$

$$\eta_1 = \frac{1}{N_{13}}\sum_{i=N_{13}+1}^{N_{12}} X_{i11} - \frac{1}{N_{03}}\sum_{i=N_{03}+1}^{N_{02}} X_{i01} - \frac{N_{11}-N_{13}}{N_{13}}\left\{\frac{1}{N_{11}}\sum_{i=1}^{N_{11}} X_{i11}\right\} + \frac{N_{01}-N_{03}}{N_{03}}\left\{\frac{1}{N_{01}}\sum_{i=1}^{N_{01}} X_{i01}\right\},$$

and γ_{13} and γ_{23} are the regressions of X_3 on X_1 and X_2 respectively, adjusted for the intervention effect. The variance of B is given by

$$\mathrm{var}(B) = \frac{\sigma_3^2(N_{03}+N_{13})}{N_{03}N_{13}}\left[1 - \rho_{13}^2\frac{N_{01}+N_{11}-N_{03}-N_{13}}{N_{01}+N_{11}} - \rho_{23}^2\frac{N_{02}+N_{12}-N_{03}-N_{13}}{N_{02}+N_{12}} + 2\rho_{13}\rho_{23}\rho_{12}\Big(1-\frac{N_{03}+N_{13}}{N_{02}+N_{12}}\Big)\right].$$

Estimates $\hat{B}$ and $\mathrm{var}(\hat{B})$ follow from estimates of the correlations ρ_{13}, ρ_{23} and ρ_{12}, regression coefficients γ_{13} and γ_{23} and standard deviations, σ_1, σ_2 and σ_3, obtained from the appropriate regression models, using all available data.

R code for worked example

The below code implements the worked example in the main text. Before running the `gsDesign` [2] package should be installed in R [3].

```
# planned early looks
N <- 30; N.1 <- c(20, 25); N.2 <- c(15, 20); N.3 <- c(10, 15)
# expected information
information <- function(N1, N2, N3, sigma.3 = 18, rho.1.2 = 0,
rho.1.3 = 0.5, rho.2.3 = 0.5){
var <- (2 * (sigma.3 ^ 2) / N3) * (1 - (rho.1.3 ^ 2) * ((N1 - N3)/N1) -
(rho.2.3 ^ 2) * ((N2 - N3)/N2) +
2 * rho.1.2 * rho.1.3 * rho.2.3 * (1 - N3/N2))
return(1/var)
}
# at look 1
i.look.1 <- information(N1 = N.1[1], N2 = N.2[1], N3 = N.3[1])
# at look 2
i.look.2 <- information(N1 = N.1[2], N2 = N.2[2], N3 = N.3[2])
# at end
i.end <- information(N1 = N, N2 = N, N3 = N)
```

```
v.info <- c(i.look.1, i.look.2, i.end) / i.end
v.info

# calculate boundaries
library(gsDesign)
# set alphas
alpha.star.u <- c(0.000, 0.001, 0.025)
alpha.star.l <- c(0.020, 0.600, 0.975)
# modify for entry into gsDesign
gs.alpha.star.u <- diff(c(0, alpha.star.u))
gs.alpha.star.l <- diff(c(0, alpha.star.l))
gs.bound <- gsBound(I = v.info, trueneg = gs.alpha.star.l,
falsepos = gs.alpha.star.u)
bounds <- rbind(lower = gs.bound$a, upper = gs.bound$b)
bounds

# data
study.data <- data.frame(X.1 = c(56,65,59,40,43,88,34,40,49,62,42,
49,54,76,54,30,52,57,52,70,50,57,39,41,19,82,82,71,37,68,64,47,53,
51,69,37,48,57,88,60,89,53,56,60,71,32,63,84,25,32,69,26,64,59,58,
46,28,37,71,64), X.2 = c(71,70,16,80,78,78,78,55,44,95,53,48,68,62,
70,28,57,52,91,52,67,50,73,55,59,62,67,63,34,82,60,53,58,46,64,73,
69,72,63,74,72,61,93,62,43,68,66,65,62,40,47,60,54,63,74,42,24,60,
100,62), X.3 = c(64,93,86,91,83,100,82,59,65,91,70,76,86,90,80,56,
88,81,64,48,80,80,100,86,78,84,82,91,73,72,78,41,59,84,100,49,82,
74,66,79,92,83,88,76,77,67,86,72,66,60,77,52,66,86,75,65,29,48,85,
97), treat = rep(c(0, 1), each = 30))

# calculate test statistic
# change look to 2 to get calculation at second look
look <- 1
# sample sizes at look
N1 <- N.1[look]; N2 <- N.2[look]; N3 <- N.3[look]

# effect estimates
eff.X <- function(x, xn, xvar, N3){
s.X <- c(x[x$treat == 0, xvar][1:xn], x[x$treat == 1, xvar][1:xn])
treat.X <- rep(seq(0, 1), rep(xn, 2))
reg.X <- lm(s.X ~ factor(treat.X))
sigma.X <- summary(reg.X)$sigma
if(xvar == "X.1" | xvar == "X.2"){
eff.X  <- sum(x[x$treat == 1, xvar][(N3 + 1):xn] -
x[x$treat == 0, xvar][(N3 + 1):xn]
- rep(reg.X$coef[2], (xn - N3)))} else {
eff.X <- as.numeric(reg.X$coef[2])
}
return(list(eff.X = eff.X, sigma.X = sigma.X))
}
eff.2 <- eff.X(x = study.data, xn = N2, xvar = "X.2", N3 = N3)
eff.1 <- eff.X(x = study.data, xn = N1, xvar = "X.1", N3 = N3)
eff.3 <- eff.X(x = study.data, xn = N3, xvar = "X.3", N3 = N3)

# regressions
reg.XY <- function(x, xvar = "X.1", yvar = "X.3", yn = N3){
X <- c(x[x$treat == 0, xvar][1:yn], x[x$treat == 1, xvar][1:yn])
Y <- c(x[x$treat == 0, yvar][1:yn], x[x$treat == 1, yvar][1:yn])
treat.Y <- rep(seq(0, 1), rep(yn, 2))
reg.X.Y <- lm(Y ~ factor(treat.Y) + X)
gamma.X.Y <- coef(reg.X.Y)[3]
return(as.numeric(gamma.X.Y))
}
gamma.1.3 <- reg.XY(x = study.data, xvar = "X.1", yvar = "X.3", yn = N3)
gamma.2.3 <- reg.XY(x = study.data, xvar = "X.2", yvar = "X.3", yn = N3)
gamma.1.2 <- reg.XY(x = study.data, xvar = "X.1", yvar = "X.2", yn = N2)
# full model
get.subdata <- function(x, yn){
sdata <- rbind(x[x$treat == 0,][1:yn,], x[x$treat == 1,][1:yn,])
return(sdata)
}
reg.1.2.3 <- lm(X.3 ~ factor(treat) + X.1 + X.2,
data = get.subdata(x = study.data, yn = N3))
```

```
# estimate B
B.hat <- eff.3$eff.X + (1 / N3) * eff.1$eff.X * gamma.1.3 +
(1 / N3) * eff.2$eff.X * gamma.2.3

# estimate correlations
C <- matrix(c(gamma.1.3 * (eff.1$sigma ^ 2),
gamma.2.3 * (eff.2$sigma ^ 2)), ncol = 2)
D <- matrix(c((eff.1$sigma ^ 2), gamma.1.2 * (eff.1$sigma ^ 2),
gamma.1.2 * (eff.1$sigma ^ 2), (eff.2$sigma ^ 2)), ncol = 2)
sigma.3 <- sqrt(as.numeric(summary(reg.1.2.3)$sigma ^ 2 +
C %*% solve(D) %*% t(C)))
rho.1.3 <- as.numeric(gamma.1.3 * eff.1$sigma / sigma.3)
rho.2.3 <- as.numeric(gamma.2.3 * eff.2$sigma / sigma.3)
rho.1.2 <- gamma.1.2 * eff.1$sigma / eff.2$sigma

# estimate variance
info.look <- information(N1 = N1, N2 = N2 , N3 = N3, sigma.3 = sigma.3,
rho.1.2 = rho.1.2, rho.1.3 = rho.1.3, rho.2.3 = rho.2.3)
vB.hat <- 1/info.look

# test statistic
z <- B.hat/sqrt(vB.hat)
z

# decision making
if(z < bounds["lower", look]){decision <- "STOP: Futility"}
if(z > bounds["upper", look]){decision <- "STOP: Efficacy"}
if(z > bounds["lower", look] &
z < bounds["upper", look]){decision <- "CONTINUE"}
decision
```

Author details
[1] Statistics and Epidemiology Unit, Warwick Medical School, University of Warwick, CV47AL Coventry, UK. [2] Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, CV47AL Coventry, UK. [3] University Hospital Coventry and Warwickshire, CV2 2DX Coventry, UK.

References

1. Galbraith, S., Marschner, I.C.: Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. Stat Med **22**(11), 1787–805 (2003)
2. Anderson, K.: gsDesign: Group Sequential Design. (2016). R package version 3.0-1. https://CRAN.R-project.org/package=gsDesign
3. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/