Bayesian adaptive designs for multi-arm trials

Ryan, Elizabeth; Lamb, Sarah E; Williamson, Esther; Gates, Simon

Bayesian adaptive designs for multi-arm trials: an orthopaedic case study

Elizabeth G Ryan[1*], Sarah E Lamb[2,3], Esther Williamson[2], and Simon Gates[1]

[*]Corresponding author. **Email:** E.G.Ryan@bham.ac.uk

[1]Cancer Research UK Clinical Trials Unit, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK B15 2TT

[2]Centre for Rehabilitation Research, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences (NDORMS), Botnar Research Centre, University of Oxford, Oxford, UK OX3 7LD

[3]College of Medicine and Health, University of Exeter, Exeter, UK EX1 2LU

Abstract

Background: Bayesian adaptive designs can be more efficient than traditional methods for multi-arm randomised controlled trials. The aim of this work was to demonstrate how Bayesian adaptive designs can be constructed for multi-arm phase III clinical trials and assess potential benefits that these designs offer.

Methods: We constructed several alternative Bayesian adaptive designs for the Collaborative Ankle Support Trial (CAST), which was a randomised controlled trial that compared four treatments for severe ankle sprain. These incorporated response adaptive randomisation, arm dropping, and early stopping for efficacy or futility. We studied the Bayesian designs' operating characteristics via simulation. We then virtually re-executed the trial by implementing the Bayesian adaptive designs using patient data sampled from the CAST study to demonstrate the practical applicability of the designs.

Results: We constructed five Bayesian adaptive designs, each of which had high power and recruited fewer patients on average than the original design's target sample size. The virtual executions showed that most of the Bayesian designs would have led to trials that declared superiority of one of the interventions over the control. Bayesian adaptive designs with RAR or arm dropping were more likely to allocate patients to better performing arms at each interim analysis. Similar estimates and conclusions were obtained from the Bayesian adaptive designs as from the original trial.

Conclusions: Using CAST as an example, this case study found that Bayesian adaptive designs can be constructed for phase III multi-arm trials using clinically relevant decision criteria. These designs demonstrated that they can potentially generate earlier results and allocate more patients to better-performing arms. We recommend the wider use of Bayesian adaptive approaches in phase III clinical trials.

Trial registration: CAST study registration ISRCTN, ISRCTN37807450. Registered 25 April 2003, retrospectively registered. http://www.isrctn.com/ISRCTN37807450

Keywords: Bayesian adaptive design; interim analysis; multi-arm trial; response adaptive randomisation; arm dropping; monitoring; orthopaedic; emergency medicine; randomised controlled trials; phase III

Background

The traditional phase III trial design involves randomising patients to one of two arms, often with equal probability of allocation. The sample size is calculated using frequentist methods, which involve assuming a particular treatment effect and type I error rate to achieve a particular level of power. Phase III trials generally require large sample sizes, have long duration, and many are declared “unsuccessful” due to a perceived lack of difference between treatment arms [1]. For decades, statisticians have been developing more efficient methods for designing clinical trials, yet the majority of trials continue to use traditional methods.

Adaptive trial designs have the potential to allow trials to answer their questions more efficiently, particularly for multi-arm trials, by enabling key operational components to be altered based on analyses of accumulated data. All possible decisions and adaptations must be specified before the trial commences, as well as the decision criteria. Potential adaptations in multi-arm trials include: stopping early for high probability of efficacy or futility; arm dropping; altering the randomisation probabilities between arms, known as outcome or response adaptive randomisation.

When implementing response adaptive randomisation (RAR), the probability of being assigned to each treatment arm is not fixed and may be altered at each interim analysis based on the accrued outcome data. For instance, the probability of being assigned to an arm could increase when the accumulated outcome data suggest that the treatment arm is superior, and thus maximises the number of patients receiving the better treatment. Arm dropping may be performed in multi-arm trials to remove an arm that does not appear to be effective. There is no globally optimal method for patient allocation in multi-arm trials and the choice of method depends on the aims and setting of the trial, as some allocation methods may be more practical than others. It is also advantageous to have planned interim analyses so that if the treatment effect is large and there is a high probability of claiming superiority, or conversely, if the treatment effect is very small or non-existent, then the trial can be stopped early.

Further advantages can be gained by designing a trial within the Bayesian framework. The Bayesian approach allows previous information on the treatment effect or response to be incorporated into the design via the prior distribution. The prior is updated as data are observed in the trial to become a posterior distribution. The posterior distribution is calculated at each interim analysis to incorporate current information. The posterior distribution provides probabilistic statements about the values of various measures of interest, such as the treatment effect, adverse event rates, or arm with the maximum response. For instance, one could obtain from the posterior distribution the probability that the relative risk is less than 1. The posterior may be used to drive the decisions at the interim analyses, such as whether to stop early for efficacy or drop an arm. The prior and posterior distributions also account for uncertainty in the unknown values of the measures of interest.

Bayesian adaptive designs have often been used in early phase trials, but there are few published phase III trials that have used a Bayesian adaptive approach from the design phase (e.g., [2-4]). In this work we will explore how Bayesian adaptive designs could be constructed for an emergency medicine (orthopaedic) multi-arm trial and examine the potential benefits that these designs may offer.

Methods

Case Study

The Collaborative Ankle Support Trial (CAST; [5-7]) was a pragmatic, individually randomised controlled trial (RCT) that compared the effectiveness of three types of mechanical ankle support with tubular bandage (control) for patients with severe ankle sprains. The three interventions were: Aircast® ankle brace, Bledsoe® boot, and below-knee cast. Patients above 16 years with an acute severe ankle sprain who were unable to bear weight, but had no fracture, were recruited from eight emergency departments in England. The primary outcome was the quality of ankle function at 12 weeks post-randomisation as measured by the foot- and ankle-related quality of life subscale (QoL) of the Foot and Ankle Outcome Score (FAOS; [8]). The FAOS QoL subscale ranges from 0 (extreme symptoms) to 100 (no symptoms). Randomisation occurred 2-3 days after the initial visit to the emergency department at a follow up clinical visit.

The CAST study was designed using traditional/frequentist methods and had a fixed design. A target sample size of 643 patients was required to provide more than 90% power to detect an absolute difference of 8-10 in the FAOS QoL, assuming a type I error rate of 5% and 20% loss to follow up. The minimal clinically important difference (MCID) in the FAOS QoL subscale was specified as a change between 8 and 10. The aim of this trial was to identify the best arm for treatment of severe ankle sprains to assist in recovery.

A revised sample size was calculated by the Data Monitoring Committee (DMC) after 100 participants were recruited and an estimated target of 480-520 participants provided at least 80% power to detect the MCID, assuming a type I error rate of 5% [7].

The CAST study randomised 584 patients: 144 to tubular bandage, 149 to Bledsoe® boot, 149 to Aircast® brace, and 142 to below-knee cast. At 12-weeks post-randomisation, the FAOS QoL was estimated to be 53.5 (95% CI 48.4, 58.6) for the tubular bandage arm. Clinically important benefits were found at 12 weeks in the FAOS QoL with the below-knee cast compared to the tubular bandage (mean difference 8.7; 95% CI 2.4, 15.0), and with the Aircast® brace compared to the tubular bandage (mean difference 8; 95% CI 1.8, 14.2). The Bledsoe® boot did not offer a clinically important difference over the tubular bandage (mean difference 6.1; 95% CI 0, 12.3). These estimates were adjusted for baseline FAOS QoL (standardised using the median as the centre), as well as age and sex.

Potential adaptations for Bayesian designs

In our Bayesian adaptive designs we want to quickly identify the best performing intervention arm. A secondary aim is to deliver the best therapy to patients within the trial. Our designs will reward better performing arms and remove poorly performing arms. The Bayesian adaptive designs were constructed as one-sided superiority studies as we were interested in demonstrating improvement over control.

To achieve this, the following types of adaptations will be explored: RAR, arm dropping and early stopping for either efficacy or lack of benefit (futility). Below we describe how these adaptive features have been incorporated into the Bayesian designs, as well as the rules with which these adaptations could be implemented. The rules for implementing these adaptations were determined based on the input of clinicians, criteria used in previous studies (e.g., [9, 10]) and the results of simulations which explored a range of clinically relevant values. Decision thresholds were also chosen to optimise probability of trial success, average number of patients randomised, and the proportion of patients randomised to the best therapy. Stopping boundaries were also chosen to ensure the simulated one-sided type I error rate was <2.5%.

The Bayesian adaptive designs were constructed by a statistician (EGR) who was independent of CAST and was blind to the data and results of the trial until the designs' operating characteristics had been simulated. The designs were constructed using the CAST protocol and discussions were held with CAST investigators (SEL and EW) to derive the design parameters and determine how the adaptive features could be incorporated to ensure the designs were practically feasible.

Interim analysis schedules and candidate designs

We investigated a range of interim analysis schedules where adaptations could be performed every 50, 100 or 200 patients due for their primary outcome assessment (12 weeks post-randomisation). We note that operationally, fewer interim analyses are typically preferred. We found that performing RAR or arm dropping more frequently increased the probability of trial success and decreased the sample size (results not shown), and so we only present the adaptive designs that performed RAR or arm dropping every 50 patients. Assessment of early stopping for efficacy or futility was performed every 200 patients due for their primary outcome assessment in each adaptive design. This was performed less frequently than RAR/arm dropping to control the type I error and reduce operational complexity, particularly for the monitoring committees who may not need to meet for randomisation probability updates or arm dropping decisions. A fixed Bayesian design was also investigated for comparative purposes. The Bayesian designs explored are described in Table 1.

Table 1. Bayesian adaptive designs explored for CAST study

Design	Interim analysis schedule[a]	Arm allocation[b]	Control allocation	Early stopping
1	None	1:1:1:1	Equal to other arms	None
2	Every 200 patients	1:1:1:1	Equal to other arms	Efficacy or futility every 200 patients
3	Every 50 patients	Arm dropping assessed at each interim analysis	Equal to other arms	Efficacy or futility every 200 patients
4	Every 50 patients	RAR at each analysis	Matched to best intervention arm	Efficacy or futility every 200 patients
5	Every 50 patients	RAR at each analysis	Fixed at 40%	Efficacy or futility every 200 patients
6	Every 50 patients	RAR at each analysis	No designated control; tubular bandage is treated as an intervention arm	Efficacy or futility every 200 patients

[a]At number patients due for primary outcome follow up (at 12 weeks post-randomisation); [b] RAR = Response adaptive randomisation

Response adaptive randomisation (RAR)

Prior to the first interim analysis, equal randomisation (ER) was used. At each interim analysis the randomisation probabilities were updated to be proportional to the posterior probability that the arm was the best intervention arm. The randomisation probabilities were then adjusted to sum to one. Enrolment was suspended to arms that had a randomisation probability < 0.1. The suspended arm(s) could re-enter the randomisation allocation at later interim analyses if the randomisation probabilities crossed above the threshold.

We explored designs that employed different approaches for control arm allocation in RAR. First we simulated trials in which the control allocation was matched to the intervention arm with the highest probability of allocation. This maximises the power for the comparison of the best arm to the control. We then assumed a fixed control allocation of approximately 40%, which may be preferred for logistical reasons. Various fixed allocations for the control were explored via simulation and the allocation of 40% was chosen based on the resulting operating characteristics (results not shown). A similar optimal control allocation was found in [11] for fixed designs. Finally, we explored a design in which the "control" arm (tubular bandage) allocation varied according to its probability of being the best arm. In this design, all arms were considered as "interventions" and recruitment to the tubular bandage arm could be suspended if it had a low probability of being the best arm (as for the other arms).

Arm dropping

We also investigated the use of permanent arm dropping, where an arm could be dropped if it had a low posterior probability (<10%) of being the best arm at an interim analysis. In the arm dropping designs, the control arm could not be dropped, but any intervention arm could be dropped. If an arm was dropped, the block size was reduced, but the overall maximum sample size was kept the same. Equal allocation was used for the remaining arms.

Early stopping for efficacy or futility

Early stopping for efficacy and futility was assessed at interim analyses performed when 200, 400 and 600 patients were due for their primary outcome assessment visit (12 weeks post-randomisation) in all adaptive designs.

For most of the adaptive designs explored (Designs 2-5, Table 1), we allowed early stopping for efficacy if there was a fairly large posterior probability of there being a MCID of 8 between the best intervention arm and the tubular bandage in the primary outcome (equation 1) and if there was a high probability (>90%) that the arm is the best arm (equation 2), i.e.,

$$\Pr\left(\theta_{Best} - \theta_{tubular\ bandage}\ > 8\right) > S_i \qquad (1)$$

$$\text{and } \Pr(Best_t) > 0.9 \qquad (2)$$

where θ_{Best} and $\theta_{tubular\ bandage}$ are the estimates of the FAOS QoL scores at 12 weeks for the best intervention arm and the tubular bandage, respectively; and S_i is the stopping boundary for efficacy at interim analysis *i* for the comparison of the best arm to the tubular bandage. $\Pr(Best_t)$ is the probability that arm *t* (*t* = boot, brace, below-knee cast) is the best arm.

Both criteria (1) and (2) must be met for the trial to stop early for efficacy. The S_i values used were 0.75, 0.7, and 0.6 for interim analyses performed at 200, 400 and 600 patients due for their primary outcome visit, respectively. These values were used for Designs 2-5 (Table 1). The stopping boundaries were chosen to ensure acceptable power and were clinically relevant values.

We also defined success criteria for the trial at the final analysis to enable the type I error and power to be calculated and compared across the designs. At the final analysis, the trial was declared successful for Designs 1-5 if:

$$\Pr\left(\theta_{Best} - \theta_{tubular\ bandage}\ > 8\right) > 0.5 \qquad (3)$$

If this criterion was not met then the trial was declared unsuccessful.

For Designs 2-5, early stopping for statistical futility was based on having a small posterior probability that the best arm is better than the tubular bandage, i.e.,

$$\Pr\left(\theta_{Best} > \theta_{tubular\ bandage}\right) < 0.05 \qquad (4)$$

Design 6 (Table 1) used RAR where allocation to the tubular bandage arm could vary according to its probability of being the best arm. This design focussed on identifying the best arm overall with a high probability rather than looking for a MCID between intervention arms and tubular bandage. Therefore, early stopping for efficacy or futility was based on the probability of being the best arm, evaluated at the best arm. That is, $\Pr(Best_{t=best\ arm})$. If this probability was <0.1 at interim analyses performed at 200, 400 or 600 patients, then the trial was stopped early for futility. If this probability was >0.975 at 200 patients, >0.95 at 400 patients, or >0.925 at 600 patients, then the trial was stopped early for efficacy. The trial was deemed to be successful at the final analysis if this probability was >0.9.

Simulation Settings

Simulations of the designs were performed in FACTS (version 6.2 [12]) so that the operating characteristics of each design could be studied. We used a recruitment rate of 5 patients/week and assumed it took 12 weeks to reach this recruitment rate. We also explored recruitment rates of 25 and 56 patients/week (assuming it took 12 weeks to reach these recruitment rates; see Appendix). We used the same dropout rate that the original study design assumed (20%). The posterior distribution was estimated for each treatment arm, and the FAOS QoL estimates at 12 weeks were adjusted for the baseline scores. Details on the model and priors used are given in Additional File 1.

Prior to the start of the CAST study there was uncertainty regarding the effect size and FAOS QoL values, and so we simulated a range of different true effect size scenarios for each design. The different scenarios explored for the primary outcome in each arm are given in Table 2.

Table 2. Scenarios explored for Bayesian designs

Scenario	Control/tubular bandage FAOS QoL	Boot FAOS QoL	Brace FAOS QoL	Below-knee Cast FAOS QoL
Null	50	50	50	50
One works, 10 more	50	50	50	60
One works, 5 more	50	50	50	55
Better, Best	50	55	60	65
One worse, others work	50	45	55	60
All work, two similar	50	55	60	60

We simulated 10,000 trials for each scenario in Table 2 for each design. The type I error was estimated using the proportion of simulations that incorrectly declared the trial to be successful when no difference was present in the true primary outcome scores (null scenario above). The power was calculated as the proportion of simulations that declared the trial to be successful, when at least one treatment was superior in the true FAOS QoL score.

We wanted to accurately estimate the response of the arm that was chosen to be the best. Some studies have shown that RAR can lead to a larger estimation bias compared to ER (e.g., [13]). To quantify bias in the estimates of the best arm responses, we use the mean square error (MSE) of estimation, where the expectation is taken over the space of successful trials.

Virtual Re-execution of Designs

A virtual re-execution of the CAST study was performed by implementing the Bayesian designs using the CAST data to illustrate the application and potential benefits of the Bayesian adaptive designs on a real-world trial. We maintained the original enrolment dates for the CAST patients in the re-execution. Since Designs 3-6 incorporated arm dropping or RAR every 50 patients, the required allocations for these designs are unlikely to match the allocations that actually occurred in the CAST data. Therefore, at each interim analysis we used the updated randomisation probabilities to obtain allocations for the next 50 patients and then randomly sampled (with replacement) a CAST patient for the re-execution dataset that had a matching treatment allocation and was randomised into the original CAST study within ±6 weeks of the re-execution enrolment date. To avoid bias, for each design the trial was virtually re-executed 1000 times by drawing data from the CAST dataset and performing the interim analyses. A flow diagram of the re-sampling and interim analysis process for Designs 3-6 is given in Figure 1. Further details are given in Additional File 1.

Designs 1 and 2 had fixed arm allocation probabilities throughout the trial, and so we could use the actual CAST data in the virtual executions of these designs without the need for re-sampling. The results of these analyses are displayed in Additional File 2. We also used a simplified version of the process described in Figure 1 to resample many datasets from the CAST data to virtually execute Designs 1 and 2 so that their results were more comparable to those from Designs 3-6. This also enabled us to examine potential gains in efficiency over a range of datasets.

Since the CAST study only recruited 584 patients we were unable to perform all planned interim analyses. The last interim analysis occurred at 400 patients for Design 2 and 550 patients for Designs 3-6. The final analysis occurred once follow-up data had been collected for the 584 patients. The re-executions were performed in R (version 3.5.0; R Foundation for Statistical Computing) and the rjags package [14] was used to perform the Bayesian analyses. We used a similar approach to [15] to perform the virtual re-executions and re-sampling of patients.

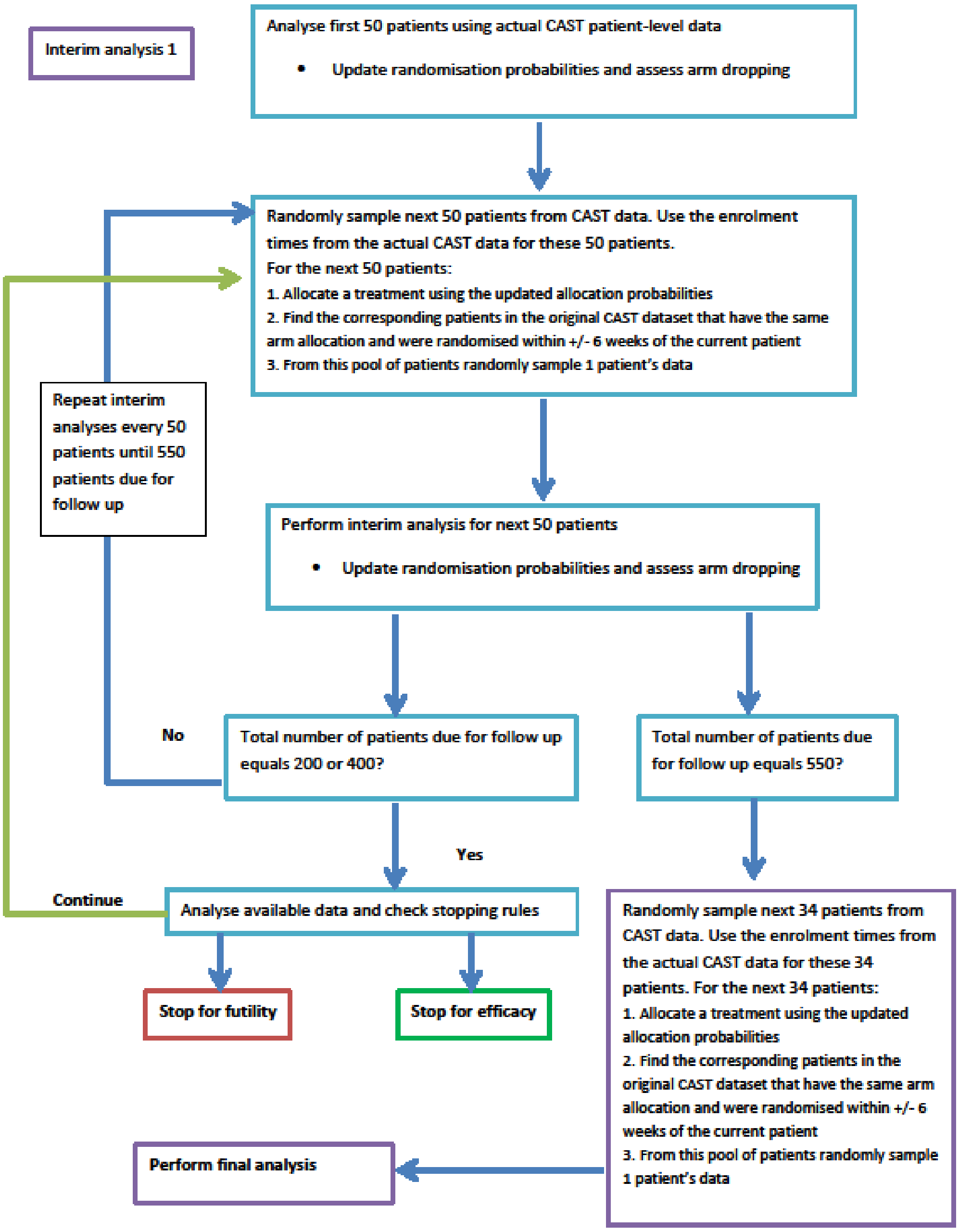
Interim analysis 1
Analyse first 50 patients using actual CAST patient-level data
• Update randomisation probabilities and assess arm dropping
Randomly sample next 50 patients from CAST data. Use the enrolment times from the actual CAST data for these 50 patients.
For the next 50 patients:
1. Allocate a treatment using the updated allocation probabilities
2. Find the corresponding patients in the original CAST dataset that have the same arm allocation and were randomised within +/- 6 weeks of the current patient
3. From this pool of patients randomly sample 1 patient's data
Repeat interim analyses every 50 patients until 550 patients due for follow up
Perform interim analysis for next 50 patients
• Update randomisation probabilities and assess arm dropping
No
Total number of patients due for follow up equals 200 or 400?
Total number of patients due for follow up equals 550?
Yes
Continue
Analyse available data and check stopping rules
Stop for futility
Stop for efficacy
Randomly sample next 34 patients from CAST data. Use the enrolment times from the actual CAST data for these 34 patients. For the next 34 patients:
1. Allocate a treatment using the updated allocation probabilities
2. Find the corresponding patients in the original CAST dataset that have the same arm allocation and were randomised within +/- 6 weeks of the current patient
3. From this pool of patients randomly sample 1 patient's data
Perform final analysis

Figure 1. Flow diagram showing the process for the virtual re-execution of Designs 3-6. Response adaptive randomisation or arm dropping was performed every 50 patients until the final analysis (at N=584). Early stopping for efficacy or futility was assessed every 200 patients. The process depicted in this figure was repeated 1000 times.

Results

Operating characteristics for Bayesian designs

Select operating characteristics for the Bayesian designs are presented in Table 3 and Figure 2. Further operating characteristics are given in Additional File 2. Boxplots of the distribution of the allocations to the control and best arm for each scenario across the 10, 000 simulations are presented in Figure 3. The effect of using a faster recruitment rate is summarised in Additional File 3.

The Bayesian adaptive designs generally offered a decreased average sample size and increased power/probability of trial success across the scenarios explored, compared to the Bayesian fixed design. There was little variation in the sample size for the null scenario across the Bayesian designs, and the simulated type I error was approximately 0 for the majority of designs under this scenario. There was also little variation in the sample size and probability of having a successful trial when a difference of 5 was assumed between the tubular bandage and the best arm for each design, apart from Bayesian Design 6, which had a slightly smaller average sample size and a higher probability of having a successful trial.

Bayesian Design 6 had the lowest average sample size and highest probability of success for the “One works, 10 more” and “One worse, others work” scenarios; the other adaptive designs had similar average sample sizes and probability of trial success and also offered improvement over the fixed design for these scenarios. All designs had high probability of success for the “Better, Best” scenario and the average sample size was reduced by incorporating interim analyses (Designs 2-6). Bayesian Design 6 did not perform as well as the other designs for the “All work, two similar” scenario in terms of the probability of trial success since one arm was not clearly superior. This decreased the probability of being the best arm and so the trial was deemed to be “unsuccessful” in 30% of the simulated trials.

We also simulated a scenario where all the intervention arms were inferior to the tubular bandage arm (FAOS 50, 45, 45, and 45 for tubular bandage, boot, brace, and below-knee cast, respectively). In Designs 1-5, all of the simulated trials were declared to be unsuccessful at the final analysis for this scenario and 41.72-58.91% of the simulated trials stopped early for futility (Designs 2-5). For this scenario Design 6 had similar results to the “One arm works, 5 more” scenario since it did not consider the tubular bandage to be a control arm here and had one arm superior by an FAOS of 5.

The probability of having a successful trial and average sample sizes were similar between Bayesian Designs 3-5 across the scenarios. Design 6 had the lowest average sample size and highest probabilities of success when a treatment effect was present and one arm was clearly superior. Design 6 did not perform as well as the other adaptive designs when two arms offered a similar improvement in the FAOS QoL. Design 6 had different objectives and decision criteria to the other Bayesian designs, and so care should be taken when choosing a preferred design since the designs are tailored to the aims of the investigators.

Due to the lack of successful trials in the null and “one arm works, 5 more” scenarios for the majority of designs, the MSE was not calculated for these scenarios. The adaptive designs tended to have slightly higher MSE than the fixed design, apart from Design 6 which had lower MSE. RAR and arm dropping designs had lower MSE compared to the design that just had early stopping for efficacy or futility (Design 2).

For each design, the correct selection of the best arm was made in approximately all of the simulated trials, where at least one arm was superior to control (data not shown). From Table 3 and Figure 3, it can be seen that, on average, more allocations were given to the best arm under designs that incorporated RAR or arm dropping when at least one arm was superior. Equal allocation to the treatment arms was achieved in the null scenario for these designs. Design 6 tended to allocate the highest proportion of patients to the best arm, followed by Design 5. RAR with control matched to best arm (Design 4) tended to have similar allocations to the design that used permanent arm dropping (Design 3). The designs with RAR or arm dropping (Designs 3-6) had a fairly large variation in their allocations to the best arm and the control, and were quite often skewed in their distribution. For Design 3, the proportion of arm drops was low for the best arm and high for the other arms (Additional File 2).

[Table 3 here]

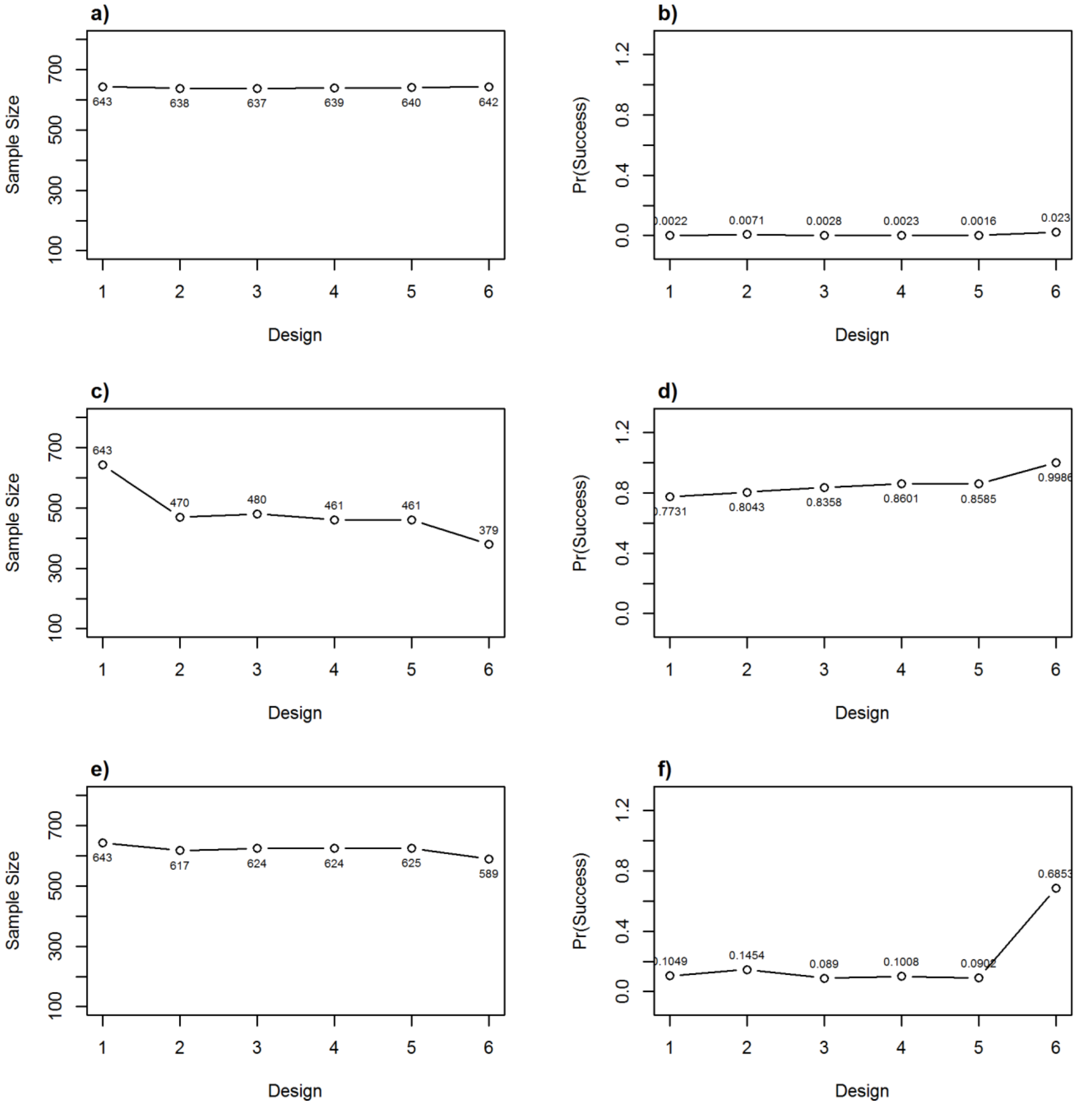

Figure 2. Average sample sizes (left column) and probability of trial success (right column) for each design. Each row represents a different scenario: a) and b) "Null" scenario; c) and d) "One works, 10 more"; e) and f) "One works, 5 more"; g) and h) "Better, Best"; i) and j) "One worse, others work"; k) and l) "All work, two similar". The type I error is represented

in Figure 2(b). The power is given in Figures 2 (d), (f), (h), (j), (l).

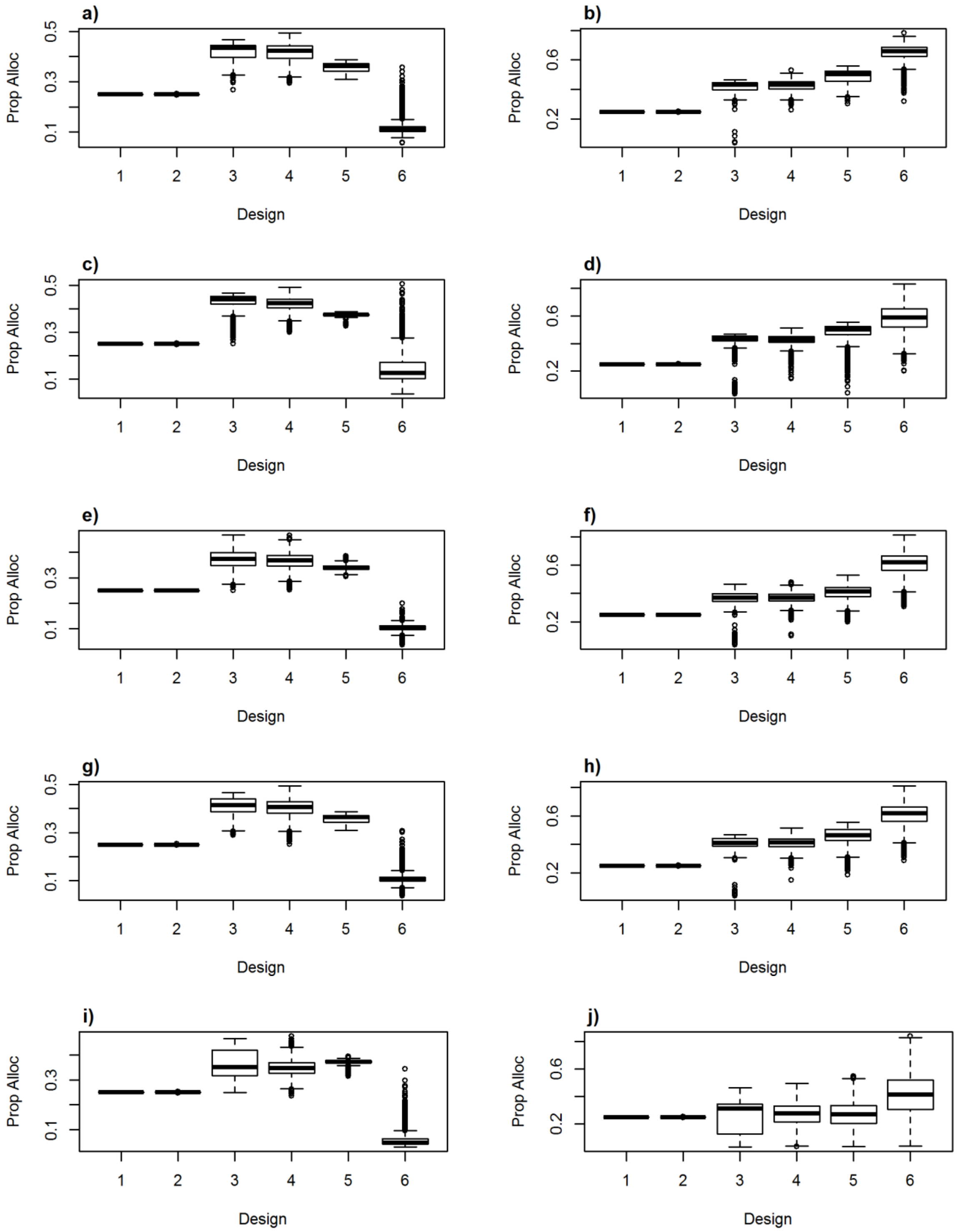

Figure 3. Allocations across 10,000 simulated trials for tubular bandage (left column) and best arm (right column). Each design is represented on the x-axis in each figure and each row represents a scenario: a) and b) “One works, 10 more”; c) and d) “One works, 5 more”; e) and f) “Better, Best”; g) and h) “One worse, others work”; i) and j) “All work, two similar”

Virtual Re-execution of designs

Table 4 presents a summary of the virtual re-execution of the CAST study under each Bayesian design across the 1000 trials that resampled the CAST study data.

[Table 4 here]

The results of the re-executions show that the Bayesian adaptive designs recommended early stopping for efficacy in 7.6-25.9% of trial re-executions, with the most frequent early stopping occurring in Design 2 which had fixed allocations and only allowed for early stopping of the trial. None of the trial re-executions recommended early stopping for futility, since all of the interventions performed better than the tubular bandage. At the final analysis for Designs 1-5, 83.5-89.4% of the trials were declared successful. Design 6, where decisions were based on the probability of being the best arm, had a low proportion (23%) of trials that were declared successful at the final analysis. This is due to the fact that the brace and below-knee cast had similar primary outcome scores and both performed well compared to the other arms. Thus, one arm was not often declared superior with a high probability. For each of the Bayesian designs, the below-knee cast was most frequently declared the best arm at the final analysis in the re-executions and thus had the same conclusion as the original trial.

The medians of the posterior estimates for the treatment effects over the 1000 re-executions were generally similar to the original frequentist analysis estimates. Designs 4 and 5 (RAR with control allocation matched to best arm and RAR with fixed control allocation, respectively) had slightly lower estimates of the mean difference between Bledsoe boot and tubular bandage. Design 6 had slightly higher estimates of the mean difference between the ankle brace and tubular bandage, and also between the below-knee cast and tubular bandage. One should also bear in mind that the re-executions were performed on re-sampled data from the original dataset, and so the estimates are likely to vary slightly.

Further summaries of the results and randomisation allocations at each interim analysis for each adaptive design are given in Additional File 4. These results show that the randomisation probabilities differed between Bayesian designs 4-6 at each interim analysis, and that these RAR designs often had quite different allocations to the CAST study, depending on which arm was "the best" at that interim analysis.

Discussion

Summary

In this study we have demonstrated how Bayesian adaptive designs can be constructed for Phase III multi-arm RCTs. Using an orthopaedic trial as a case study, we outlined the process involved in constructing the designs, described the adaptive schemes and stopping rules employed, and demonstrated the designs' behaviour through their operating characteristics across a range of scenarios. We also performed virtual executions of the Bayesian designs using data from the CAST study to demonstrate the decisions that would be made using the Bayesian designs and the trial data. Through use of the Bayesian adaptive approach we were able to make decisions about whether to stop the trial early based on the probability of having a MCID, update the randomisation allocations according to the probability of being the best arm, and suspend recruitment to arms that had a low probability of being the best.

Based on the operating characteristics, the use of Bayesian adaptive designs for this case study generally increased the power and decreased average sample size compared to a fixed design. Use of RAR or arm dropping offered increased power compared to an adaptive design that only allowed for early stopping when it was assumed that one arm offered a MCID. All designs had low type I error and high probabilities to detect a MCID in at least one arm, when it was assumed that a MCID was truly present. The correct selection of the best arm was made in approximately all of the simulated trials, where at least one arm was superior to control. Use of RAR or arm dropping produced simulated trials that gave more allocations to the best arm when at least one arm was superior. Equal allocation occurred when the arms had approximately the same primary outcome scores.

Design 6, whose decisions were made based on the probability of being the best arm, typically produced the lowest average sample size and highest power when one arm was assumed to be superior. However, this design did not perform well when two arms showed a similar improvement compared to the other arms. Design 6 had different objectives and decision criteria to the other Bayesian designs, and so care should be taken when choosing a preferred design since the designs are tailored to the aims of the investigators.

The virtual executions of the Bayesian designs using the CAST data showed that early stopping for efficacy only occurred in a small proportion of trials and that no trials stopped early for futility. At the final analysis >80% of the trials were declared successful in the 1000 executions of Designs 1-5. When Design 6 was executed 1000 times using the trial data, only 23% of the trials were declared successful at the final analysis, since both the brace and below-knee cast performed similarly well and a “best arm” was not declared with a high probability. A benefit of Design 6 was that the tubular bandage arm, which was the control arm in the other designs, had smaller allocation probabilities which allowed more allocations to better performing arms. The below-knee cast was most often declared the best arm at the final analysis in the re-executions, and so the Bayesian designs led to the same conclusion as the original trial.

The decisions made at the interim and final analyses of the Bayesian designs were driven by the primary outcome. We did not incorporate other outcomes and are not intending that the conclusions generated in this re-execution be used to inform clinical practice or to alter the conclusions of the original study.

Limitations

Adaptive designs have great promise for producing trials with better operating characteristics, but present a number of practical challenges. Korn and Freidlin [16] provide a summary of some of the advantages and disadvantages of different adaptive design elements.

Adaptive designs require a larger amount of work to build and evaluate potential designs, compared to fixed designs, and may take more effort to obtain approval from review boards. Adaptive designs can also be more complicated to implement. Performance of the interim analyses and making the required adaptations is dependent on being able to collect, enter, clean and analyse data in a timely manner, and alter the randomisation system with ease. This requires the trial management team, statisticians, programming teams and trial treatment providers/intervention suppliers to be responsive to changes that need to be made. These rapid changes may not be possible in all trial settings.

The interim analyses also need to be adequately spaced to allow time for DMCs and Trial Steering Committees (TSCs) to meet. Statistically, more frequent interim analyses generally produce better operating characteristics for designs that use RAR or arm dropping (e.g., [17]), but frequent interim analyses may not always be practical. The DMC/TSC may not necessarily need to meet for every interim analysis, e.g., for RAR adaptations, but would need to meet for stopping decisions.

The types of adaptations that can be made to multi-arm trials are situation-dependent. RAR presents difficulties in being able to anticipate and arrange for the delivery of treatments. The original CAST study design, which had fixed allocations, allowed the supply of treatment arms (including the supply of staffing) to be planned more easily than a design with RAR would. RAR may not always be possible due to restrictions on resources for delivering the treatments or delays in collecting the primary outcome data. Closure of arms may be practically easier to achieve. Whilst early stopping of trials may have benefits for funding agencies, academic trial investigators often do not wish to terminate trials early, due to potential loss of research income and staff retention. Changes in funding models are likely to be required to fully take advantage of innovation in trial design, such as a minimum study time funded with a mechanism to release funding if full study time is required. Additionally, trials that stop early may have little information on the long-term effects of treatment or on secondary outcomes, and are likely to produce less precise estimates of the treatment effects.

The use of RAR remains controversial and some of its properties are not well understood by clinicians. RAR has its greatest potential in multi-arm trials but has limited usefulness in two-armed trials [18, 19]. Adaptive designs are more susceptible to changes in patient population over time. Designs with RAR are robust to moderate changes in population (see [20]), but adaptive designs are not appropriate if the patient population changes dramatically during the trial. When evaluating adaptive designs, simulation is required to illustrate the operating characteristics and potential benefits, and investigate potential biases introduced by each adaptive feature.

Fairly short follow-up times are required for adaptive designs to offer improved efficiency. Adaptive designs are difficult to implement for very fast recruitment rates, particularly for studies that have relatively longer follow-up periods, since less information will be available at each interim analysis. This poses difficulties for phase III trials since the primary outcome is often based on long-term measures, and it may be difficult to design adaptive trials without extending the time frame of recruitment to allow for the interim analyses and potential adaptations to occur. Thus there may be a trade-off in reduced sample size but increased recruitment time (at a slower recruitment rate) for some adaptive trial design contexts.

In this work we virtually executed each of the proposed Bayesian designs using trial data to illustrate their practical applicability. However, in reality, one design would have been chosen and implemented, depending on its operating characteristics, practical restraints and the aims of the trial. When virtually executing the designs that incorporated arm dropping or RAR, resampling from the original trial data was required to obtain the required randomisation allocations. This may lead to an underestimation of the uncertainty in the results [10]. We addressed this by re-executing the

CAST study 1000 times and resampled patients within each trial. If different datasets had been used, different conclusions may have been obtained using these designs.

We did not simulate the decision making process of a DMC/TSC. We have assumed that the decision making process was driven by the primary outcome, but the DMC/TSC would also examine safety data and any relevant external evidence. Whilst the role of these committees is to ensure that the study protocol is accurately followed, they may also need to make deviations to ensure patient safety. For example, RAR may recommend increasing the allocation probability to an arm that has a higher rate of adverse events – an event that was not accounted for in the RAR algorithm. Alterations to the previously defined adaptations can lead to unknown operating characteristics.

The Bayesian adaptive designs were constructed as one-sided superiority studies, whereas the original CAST study was a two-sided trial. We were interested in demonstrating improvement over a much cheaper control, and felt that a DMC would be unlikely to continue enrolment into a poorly performing comparator just to show it is worse. Under most of our Bayesian adaptive designs, if an intervention arm performed poorly it would be dropped or have a very low probability of allocation. Harm may or may not be reflected in the FAOS QoL score, but the DMC could intervene if any arms were causing harm.

The designs presented here are situation specific and have been tailored to the clinical situation and aims of the CAST study. The definition of a successful trial and the level of sufficient evidence required to make decisions will differ between researchers and stakeholders, and will depend on the consequences of the actions that may be taken. The designs and findings from this work will not generalise to all phase III randomised controlled trials, but similar approaches can be used to construct Bayesian adaptive designs. We recommend that simulations are used to study the impact of each type of adaptive component on the operating characteristics when constructing Bayesian adaptive designs for multi-arm trials.

Conclusions

To enable phase III trials to achieve their aims, more efficient methods are required. Innovation in clinical trial design is extremely important as it can potentially improve the efficiency, quality of knowledge gained, cost and safety of clinical trials. In this work we have demonstrated how Bayesian adaptive trials can be designed and implemented for multi-arm phase III trials. Using a published example from orthopaedic medicine, we highlight some of the benefits of these designs, particularly for multi-arm trials.

List of abbreviations

RAR: response adaptive randomisation; **CAST**: Collaborative Ankle Support Trial; **RCT:** randomised controlled trial; **QoL**: quality of life; **FAOS**: Foot and Ankle Outcome Score; **MCID**: minimal clinically important difference; **DMC:** Data Monitoring Committee; **ER**: equal randomisation; **FACTS:** Fixed and Adaptive Clinical Trial Simulator; **MSE:** mean square error; **TSC:** Trial Steering Committee

Declarations

Ethics approval and consent to participate: The CAST study protocol was approved by the Northern and Yorkshire multicentre research ethics committee (MREC/2/3/9), and the research and ethics committees for each collaborating hospital. All individuals provided written informed consent. Additional ethics approval was not sought to perform this secondary analysis.

Consent for publication: Not applicable

Availability of data and materials: The data used in this study were generated as part of the CAST study. Requests to share individual, de-identified participant data, aggregated data, data dictionaries, and other study documents from this study should be sent to the CAST CI (SE Lamb).

Data-sharing requests will be assessed on their individual merits. Other documents relating to this secondary analysis may be available on request from the lead researcher (EG Ryan). Requests for documents will be assessed on their individual merits.

Competing interests: The authors declare that they have no competing interests.

Funding: EGR, SEL and SG were supported by a Medical Research Council (MRC) Methodology Research Grant (Grant number: MR/N028287/1) during the conduct of this research. The original CAST study was funded by the National Co-ordinating Centre for Health Technology Assessment (NCCHTA) Programme (project number 01/14/10). SG is supported as a NIHR Senior Investigator.

Authors' contributions: EGR constructed the Bayesian adaptive designs, ran the simulations of the designs, performed the virtual re-executions and drafted the manuscript; SG directed the research; SEL was a Chief Investigator of the CAST study; SEL and EW were involved in running the original CAST study and provided feedback on the Bayesian adaptive designs; all authors discussed and commented on the manuscript, and approved the final version of the manuscript.

Acknowledgements: The authors would like to thank the CAST team for their work on the original trial. The authors would also like to thank Kert Viele from Berry Consultants for his advice on running the Bayesian adaptive designs in FACTS.

Additional Files

Additional file 1 – Additional information on models, priors and re-execution process

Additional file 2 – Additional Operating Characteristics

Additional file 3 – Faster recruitment rates

Additional file 4 – Virtual execution of Bayesian designs

References

1. Arrowsmith J. Trial watch: phase III and submission failures: 2007-2010. Nat Rev Drug Discov. 2011; 10(2): 87
2. Shah PL, Slebos D-J, Cardoso PFG, Cetti E, Voelker K, Levine B, et al. Bronchoscopic lung-volume reduction with Exhale airway stents for emphysema (EASE trial): randomized, sham-controlled, multicentre trial. Lancet. 2011; 378: 997-1005.
3. Reardon MJ, Van Mieghem NM, Popma JJ, Kleiman NS, Søndergaard L, Mumtaz M, et al. Aortic-Valve Replacement in Intermediate-Risk Patients. N Engl J Med. 2017; 376(14): 1321-1331.
4. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct. N Engl J Med. 2018; 378: 11-21
5. Lamb SE, Nakash RA, Withers EJ, Clark M, Marsh JL, Wilson S, et al. Clinical and cost effectiveness of mechanical support for severe ankle sprains: design of a randomised controlled trial in the emergency department . BMC Musculoskelet Disord. 2005; 6:1
6. Cooke MW, Marsh JL, Clark M, Nakash R, Jarvis RM, Hutton JL, et al., on behalf of the CAST trial group. Treatment of severe ankle sprain: a pragmatic randomised controlled trial comparing the clinical effectiveness and cost-effectiveness of three types of mechanical ankle support with tubular bandage. The CAST trial. Health Technol Assess 2009; 13(13).
7. Lamb SE, Marsh JL, Hutton JL, Nakash R, Cooke MW; Collaborative Ankle Support Trial (CAST Group). Mechanical supports for acute, severe ankle sprain: a pragmatic, multicentre, randomised controlled trial. Lancet. 2009; 373(9663): 575-81.
8. Roos E, Brandsson S, Karlsson J. Validation of the foot and ankle outcome score for ankle ligament reconstruction. Foot Ankle Int. 2001; 22: 788–94.
9. Connor JT, Elm JJ, Broglio KR; ESETT and ADAPT-IT Investigators. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. J Clin Epidemiol. 2013; 66: S130-S137.
10. Connor JT, Luce BR, Broglio KR, Ishak KJ, Mullins CD, Vanness DJ, et al. Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study. Clin Trials. 2013; 10(5): 807-827.

11. Dunnett C. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. J Amer Stat Assoc. 1955; 50: 1096-121.
12. Fixed and Adaptive Clinical Trial Simulator (FACTS). Version 6.2. 2018 Berry Consultants, Austin TX.
13. Thall P, Fox P, Wathen, J. Statistical Controversies in Clinical Research: Scientific and Ethical Problems with Adaptive Randomization in Comparative Clinical Trials. Ann Oncol. 2015; 26:1621-8.
14. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March 2003. p. 20–2.
15. Luce BR, Connor JT, Broglio KR, Mullins CD, Ishak KJ, Saunders E, et al. Using Bayesian Adaptive Trial Designs for Comparative Effectiveness Research: A Virtual Trial Execution. Ann Intern Med. 2016; 165:431-438.
16. Korn EL and Freidlin B. Adaptive Clinical Trials: Advantages and Disadvantages of Various Adaptive Design Elements. J Natl Cancer Inst. 2017; 109(6): djx013
17. Jiang Y, Zhao W, Durkalski-Mauldin V. Impact of adaptation algorithm, timing, and stopping boundaries on the performance of Bayesian response adaptive randomization in confirmative trials with a binary endpoint. Contemp Clin Trials. 2017*;* 62:114-120.
18. Berry DA. Adaptive clinical trials: The promise and the caution. J Clin Oncol. 2011; 29(6): 606-09
19. Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? J Clin Oncol. 2011; 29: 771-776
20. Cook JD. The Effect of Population Drift on Adaptively Randomized Trials. UT MD Anderson Cancer Centre Department of Biostatistics Working Paper Series. Working Paper 39. 2007

Table 3. Operating characteristics for Bayesian designs for CAST study

Design[a]	Proportion stopping early for efficacy	Proportion stopping early for futility	MSE	Mean proportion allocated to control	Mean proportion allocated to boot	Mean proportion allocated to brace	Mean proportion allocated to below-knee cast
Design 1 Fixed design							
Null (50, 50, 50, 50)	NA	NA	NA	0.25	0.25	0.25	0.25
One arm works, 10 more (50, 50, 50, 60)	NA	NA	2.77	0.25	0.25	0.25	0.25
One arm works, 5 more (50, 50, 50, 55)	NA	NA	NA	0.25	0.25	0.25	0.25
Better best (50, 55, 60, 65)	NA	NA	3.29	0.25	0.25	0.25	0.25
One worse, others work (50, 45, 55, 60)	NA	NA	2.96	0.25	0.25	0.25	0.25
All work, two similar (50, 55, 60, 60)	NA	NA	3.39	0.25	0.25	0.25	0.25
Design 2 Interim analysis every 200 patients, early stopping for efficacy or futility							
Null (50, 50, 50, 50)	0.0063	0.0063	NA	0.25	0.25	0.25	0.25
One arm works, 10 more (50, 50, 50, 60)	0.732	0.732	5.03	0.25	0.25	0.25	0.25
One arm works, 5 more (50, 50, 50, 55)	0.1091	0.1091	NA	0.25	0.25	0.25	0.25
Better best (50, 55, 60, 65)	0.7953	0.7953	5.11	0.25	0.25	0.25	0.25
One worse, others work (50, 45, 55, 60)	0.6341	0.6341	5.10	0.25	0.25	0.25	0.25
All work, two similar (50, 55, 60, 60)	0.2701	0.2701	5.24	0.25	0.25	0.25	0.25
Design 3 Arm dropping every 50 patients							
Null (50, 50, 50, 50)	0.0025	0.0025	NA	0.36	0.21	0.21	0.21
One arm works, 10 more (50, 50, 50, 60)	0.6919	0.6919	3.68	0.40	0.11	0.11	0.39

One arm works, 5 more (50, 50, 50, 55)	0.0624	0.0624	NA	0.39	0.13	0.13	0.35
Better best (50, 55, 60, 65)	0.6843	0.6843	4.16	0.37	0.10	0.19	0.34
One worse, others work (50, 45, 55, 60)	0.6123	0.6123	3.86	0.38	0.07	0.18	0.36
All work, two similar (50, 55, 60, 60)	0.2692	0.2692	3.87	0.36	0.11	0.27	0.26
Design 4 RAR every 50 patients, control matched to best arm							
Null (50, 50, 50, 50)	0.0022	0.0022	NA	0.33	0.22	0.22	0.22
One arm works, 10 more (50, 50, 50, 60)	0.796	0.796	3.56	0.39	0.11	0.11	0.39
One arm works, 5 more (50, 50, 50, 55)	0.0733	0.0733	NA	0.37	0.14	0.14	0.35
Better best (50, 55, 60, 65)	0.8177	0.8177	4.05	0.36	0.11	0.19	0.35
One worse, others work (50, 45, 55, 60)	0.6872	0.6872	3.67	0.38	0.07	0.18	0.37
All work, two similar (50, 55, 60, 60)	0.2744	0.2744	3.73	0.35	0.12	0.27	0.27
Design 5 RAR every 50 patients, control fixed allocation of 40%							
Null (50, 50, 50, 50)	0.0015	0.0015	NA	0.37	0.21	0.21	0.21
One arm works, 10 more (50, 50, 50, 60)	0.7909	0.7909	3.29	0.36	0.10	0.10	0.44
One arm works, 5 more (50, 50, 50, 55)	0.0677	0.0677	NA	0.37	0.13	0.13	0.37
Better best (50, 55, 60, 65)	0.8069	0.8069	3.86	0.36	0.10	0.18	0.37
One worse, others work (50, 45, 55, 60)	0.6856	0.6856	3.43	0.36	0.07	0.17	0.40
All work, two similar (50, 55, 60, 60)	0.2744	0.2744	3.54	0.37	0.10	0.26	0.27
Design 6 RAR every 50 patients, no control arm							
Null (50, 50, 50, 50)	0.0117	0.0117	NA	0.25	0.25	0.25	0.25

One arm works, 10 more (50, 50, 50, 60)	0.9972	0.9972	2.34	0.13	0.13	0.13	0.61
One arm works, 5 more (50, 50, 50, 55)	0.5654	0.5654	NA	0.15	0.15	0.15	0.54
Better best (50, 55, 60, 65)	0.8982	0.8982	1.95	0.07	0.10	0.22	0.61
One worse, others work (50, 45, 55, 60)	0.8972	0.8972	1.95	0.10	0.07	0.22	0.61
All work, two similar (50, 55, 60, 60)	0.5493	0.5493	2.93	0.06	0.12	0.41	0.41

[a] Each row represents a different scenario for each design where the assumed FAOS QoL score is given in brackets as tubular bandage, boot, brace and below-knee cast score

Table 4. Summary of re-executions of CAST study using each Bayesian design

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6
Proportion stopping for efficacy at 200 patients	NA	0.216	0.148	0.166	0.147	0.072
Proportion stopping for efficacy at 400 patients	NA	0.043	0.011	0.017	0.011	0.004
Proportion stopping for futility at 200 patients	NA	0	0	0	0	0
Proportion stopping for futility at 400 patients	NA	0	0	0	0	0
Proportion re-executions declared successful at final analysis	0.855	0.894	0.835	0.865	0.877	0.23
Proportion re-executions tubular bandage (control) declared best at final analysis	0	0	0.001	0	0	0
Proportion re-executions boot declared best at final analysis	0.054	0.057	0.085	0.036	0.021	0.007
Proportion re-executions brace declared best at final analysis	0.437	0.402	0.43	0.451	0.481	0.432
Proportion re-executions below-knee cast declared best at final analysis	0.509	0.541	0.484	0.513	0.498	0.561

Median (IQR) of the posterior mean estimates for tubular bandage	54.25 (52.70, 55.68)	53.72 (51.90, 55.46)	54.40 (52.99, 55.74)	53.91 (52.52, 55.30)	53.97 (52.64, 55.33)	52.49 (51.68, 52.96)
Median (IQR) of the posterior estimates of the difference in means between boot and tubular bandage	5.60 (3.65, 7.48)	6.00 (4.02, 8.25)	5.65 (3.75, 7.56)	4.77 (2.42, 6.84)	4.85 (2.58, 7.05)	6.42 (3.98, 8.15)
Median (IQR) of the posterior estimates of the difference in means between brace and tubular bandage	8.60 (6,52, 10.63)	8.66 (6.67, 10.89)	7.62 (4.81, 10.22)	8.48 (5.65, 10.71)	8.67 (5.99, 10.73)	9.64 (6.01, 11.66)
Median (IQR) of the posterior estimates of the difference in means between below knee cast and tubular bandage	8.70 (6.86, 10.91)	9.69 (7.22, 13.29)	8.06 (5.44, 10.53)	8.79 (6.57, 11.39),	8.68 (6.58, 11.27)	10.57 (8.69, 11.78)