Received: 25 June 2021 | Revised: 20 January 2023 | Accepted: 18 March 2023

DOI: 10.1002/sim.9734

RESEARCH ARTICLE

Statistics in Medicine WILEY

Point estimation for adaptive trial designs II: Practical considerations and guidance

David S. Robertson[1] | Babak Choodari-Oskooei[2] | Munya Dimairo[3] | Laura Flight[3] | Philip Pallmann[4] | Thomas Jaki[1,5]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[2]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, UK

[3]School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

[4]Centre for Trials Research, Cardiff University, Cardiff, UK

[5]Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

Correspondence
David S. Robertson, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK.
Email: david.robertson@mrc-bsu.cam.ac.uk

Funding information
Biometrika Trust; Health and Care Research Wales; Medical Research Council, Grant/Award Numbers: MC_UU_00002/14, MC_UU_00004_09, MC_UU_12023_29; National Institute for Health Research, Grant/Award Numbers: BRC-1215-20014, NIHR-SRF-2015-08-001

In adaptive clinical trials, the conventional end-of-trial point estimate of a treatment effect is prone to bias, that is, a systematic tendency to deviate from its true value. As stated in recent FDA guidance on adaptive designs, it is desirable to report estimates of treatment effects that reduce or remove this bias. However, it may be unclear which of the available estimators are preferable, and their use remains rare in practice. This article is the second in a two-part series that studies the issue of bias in point estimation for adaptive trials. Part I provided a methodological review of approaches to remove or reduce the potential bias in point estimation for adaptive designs. In part II, we discuss how bias can affect standard estimators and assess the negative impact this can have. We review current practice for reporting point estimates and illustrate the computation of different estimators using a real adaptive trial example (including code), which we use as a basis for a simulation study. We show that while on average the values of these estimators can be similar, for a particular trial realization they can give noticeably different values for the estimated treatment effect. Finally, we propose guidelines for researchers around the choice of estimators and the reporting of estimates following an adaptive design. The issue of bias should be considered throughout the whole lifecycle of an adaptive design, with the estimation strategy prespecified in the statistical analysis plan. When available, unbiased or bias-reduced estimates are to be preferred.

KEYWORDS
adaptive design, bias-correction, conditional bias, point estimation, unconditional bias

1 | INTRODUCTION

Traditional clinical trials follow a design that is fixed in advance, with the data only analyzed after completion when all trial participants have been recruited and have accrued outcome data.[1] In contrast, adaptive designs allow for pre-planned modifications to the trial's course based on data accumulating within the trial.[2-4] Adding controlled flexibility to the trial design in this way, while still maintaining scientific rigour, can lead to advantages in terms of efficiency and ethics compared with a traditional fixed design.[2,5] The uptake of adaptive designs in practice is becoming increasingly common,[6-8] with the COVID-19 pandemic only accelerating their use.[9,10] We start by briefly highlighting some well-established types of trial adaptations below, which we will return to in this article before focusing attention to the question of treatment effect estimation at trial completion.

Early trial stopping: Group sequential designs

When monitoring the accumulating outcome data of a clinical trial, it can be beneficial to have the option of stopping the trial early for safety, futility (ie, lack of benefit), or efficacy as soon as sufficient evidence is reached to make a reliable conclusion. As continuous monitoring after every trial participant is often impractical, it is more feasible to monitor the data at (typically pre-specified) periodic intervals after a group of trial participants have accrued outcome data. This is known as a group sequential design, and at each interim analysis, the trial can be stopped early for futility or efficacy based on predefined stopping rules or boundaries.[11] Researchers can derive their own stopping boundaries or use standard ones such as the O'Brien-Fleming (OBF)[12] and Haybittle-Peto[13] stopping boundaries. Optimal stopping boundaries using various optimization criteria have also been proposed.[14,15]

Treatment selection: Multi-arm multi-stage (MAMS) designs

In some therapeutic areas, there may be several treatments or combinations of treatments awaiting evaluation in controlled clinical trials. One way of doing so efficiently is to use a MAMS design, where multiple experimental treatment arms are compared simultaneously against a single common control.[16] Similarly, to group sequential designs, MAMS designs allow for early stopping of recruitment to a treatment arm for efficacy or futility (which also allows early stopping of the whole trial). These pre-planned stopping rules can be chosen to find the single best treatment[17] or all promising treatments[18] to carry forward for further evaluation. More flexible stopping rules and adaptation methods also exist.[19-21] A variant of MAMS is the drop-the-loser design,[22-24] where a pre-determined number of experimental treatment arms (ie, the worst performing ones) are dropped at each stage, typically leaving a single treatment at the final analysis.

Population selection: Adaptive enrichment designs

There can sometimes be large uncertainty regarding which patients would benefit from a study treatment, combined with some information to suggest that patients with certain characteristics may benefit more than others. For example, in oncology, there is a recognition that tumors can have potentially large biological heterogeneity. This motivates characterizing and selecting sub-populations that are more likely to benefit from an experimental treatment.[25] Adaptive enrichment designs use interim analyses to decide which of the subpopulations should be recruited from for the remainder of the trial, where the subpopulations can be defined using biomarkers for example. Such designs can increase recruitment to the sub-populations that are estimated to receive the greatest benefit, and decrease (or stop) recruitment to the sub-populations that do not. A variety of decision rules have been proposed to select subpopulations in this way, from both Bayesian[25-27] and frequentist[28-30] perspectives.

Changing randomisation probabilities: Response-adaptive randomisation (RAR)

Traditional clinical trials use fixed randomization schemes, which do not change as a result of patients' response to treatment. Alternatively, the accruing response data can be used to change the randomization probabilities for allocating patients to treatment arms, which is known as response-adaptive randomization (RAR). A common motivation for doing so is to allocate more patients to a treatment that is estimated to be more effective during the trial, but RAR can also be used to target other experimental objectives such as increasing the power of a specific treatment comparison.[31] Many different types of RAR procedures have been proposed for various trial contexts.[31-33] RAR can also be applied in adaptive trials in combination with other adaptations such as treatment and (sub-) population selection.[34,35]

Changing sample sizes: Sample size re-estimation

When calculating the sample size for a trial, there may be substantial uncertainty around key design parameters (eg, the variance of the outcome). Sample size re-estimation (or reassessment or recalculation) designs aim to help ensure that the sample size for a trial is appropriate, by estimating design parameters at an interim analysis and using these to recalculate the sample size based on, for example, conditional power considerations.[36-38] This may be done in either a blinded or unblinded manner.[37,39] Sample size re-estimation can also be used in conjunction with other types of trial adaptations, such as group sequential designs.[40]

Further educational material on all of these adaptive designs can be found in Burnett et al,[5] Pallmann et al,[2] and the PANDA online resource (https://panda.shef.ac.uk/).

Regardless of the type of adaptive design, it is crucial that the integrity and validity of the trial is maintained.[41] Appropriate estimation of treatment effects is a key part of trial validity, as stated in the FDA guidance on adaptive designs[42](p8): "It is important that clinical trials produce sufficiently reliable treatment effect estimates to facilitate an evaluation of benefit-risk and to appropriately label new drugs (*sic* treatments), enabling the practice of evidence-based medicine". The issue with estimation after an adaptive clinical trial is that the conventional end-of-trial estimates can be prone to bias, which is defined as "a systematic tendency for the estimate of treatment effect to deviate from its true value".[42](p3) It is clear that (all else being equal, such as the variance) it is desirable to obtain unbiased point estimates of treatment effects in order to make reliable conclusions about the treatments in a trial. While equally important, the construction of related quantities for inference, such as confidence intervals or regions, is beyond the scope of this article so we signpost the interested reader to related literature.[43,44]

This article is the second in a two-part series that studies the issue of potential bias in point estimation for adaptive designs. In part I of the series,[45] we reviewed and compared current methods for unbiased and bias-reduced estimation of treatment effects after an adaptive clinical trial and critically discussed different approaches. In the current article (part II), we consider point estimation for adaptive designs from a practical perspective, and propose a set of guidelines for researchers around the choice of estimators and the reporting of estimates following an adaptive design. We first describe the problem of estimation bias in an adaptive design in more detail in Section 2, and review current practice in Section 3. We then provide an exemplary case study in Section 4 for a real adaptive trial, using different types of unbiased and bias-reduced estimators (with R code provided) as described in part I of this article series. We also include a simulation study and graphical representation of the sampling distribution of these different estimators. We conclude with guidance for researchers and discussion in Sections 5 and 6.

2 | THE PROBLEM OF ESTIMATION BIAS IN ADAPTIVE DESIGNS

The problem with using conventional estimators (ie, maximum likelihood estimators, MLEs) after an adaptive trial design is that these are prone to bias. This can be because of population or treatment selection that takes place following an interim analysis (see Bauer et al[46] for a detailed explanation of why selection process results in bias) or other types of adaptations, such as early stopping, that affect the sampling distribution of the estimator. The usual MLE is sometimes referred to as the “naive” estimator for the trial as it does not take into account the planned and realized trial adaptations.

As introduced in part I of this article series, different definitions of an unbiased estimator are relevant in our context, which we recapitulate below. We denote the population parameter of interest, the treatment effect, by θ and an estimator thereof by $\hat{\theta}$.

Mean-unbiased estimators

An estimator $\hat{\theta}$ is called *mean-unbiased* if its expected value is the same as the true value of the parameter of interest, that is, $E(\hat{\theta}) = \theta$.

Median-unbiased estimators

An estimator $\hat{\theta}$ is called *median-unbiased* if $P(\hat{\theta} < \theta) = P(\hat{\theta} > \theta)$ that is, if the probability of overestimation is the same as the probability of underestimation.

Conditionally and unconditionally unbiased estimators

An estimator is unconditionally unbiased (also known as marginally unbiased) if it is unbiased when averaged across all possible realizations of an adaptive trial. In contrast, an estimator is conditionally unbiased if it is unbiased only conditional on the occurrence of a subset of trial realizations. For example, in a drop-the-loser trial the interest will typically be on estimating the performance of the ultimately selected arm, motivating the use of a conditionally unbiased estimator (conditional on that arm being selected).

2.1 | The potential negative impacts of reporting biased estimates

Adaptive designs play an important role in health care decision-making, increasingly providing evidence of the clinical effectiveness of an intervention as well as key secondary outcomes such as health economic ones. Failing to account for biases in point estimates can result in incorrect decision-making, potentially wasting limited resources, preventing patients from receiving effective treatments, or exposing patients to unnecessary risks. In the following subsections, we consider some potential negative impacts of reporting biased estimates.

2.1.1 | Reporting biased primary outcomes

As highlighted by Dimairo et al,[3,4] the goal of clinical trials should be to provide reliable estimates of the treatment effect to inform accurate decision-making, but this can be compromised when an adaptive design is analyzed with inappropriate statistical methods. Clearly, reporting substantially biased estimates for a primary outcome measure following an adaptive design can result in poor decisions. However, other negative impacts include the results of adaptive designs being viewed with skepticism amongst stakeholders who are aware of potential biases but do not feel they have been adequately addressed by researchers.[47] This could impede the uptake of results from an adaptive trial design or discourage research teams from using these designs in practice.

A further concern is the potential for over- or underestimation of treatment effects to affect further research. In a phase II trial, for example, ineffective treatments with exaggerated effects may be wrongly selected for further investigations in phase III trials or potentially effective treatments may not be pursued further when their effects are underestimated.[48-52] However, Wang et al[53] and Goodman[54] suggest that group sequential trials that stop early do not produce materially biased estimates.

The consequences following biased estimates from a phase III trial could include treatments being made available to patients with an overstated benefit or treatments not being recommended for use in practice because of an understated benefit, see Briel et al[48] and Guyatt et al[55] for example. Both scenarios can have a detrimental impact on patients, especially in resource limited healthcare systems such as the National Health Service (NHS) in the UK, where resources spent on a treatment with overstated benefit removes funding for alternative treatments elsewhere in the system. Mueller et al[56] also argue that there are serious ethical problems when trialists fail to account for bias in estimates following an adaptive design, as this may violate the scientific validity of the research and social value when these estimates are used to inform clinical decision-making.

2.1.2 | Secondary clinical outcomes

Clinical trials often collect information about a number of key secondary outcomes that may also require adjustment in an adaptive design. If these secondary outcomes are strongly correlated with the primary outcome used to inform the adaptations to the trial they will also be vulnerable to bias.[57,58] This is highlighted in the FDA guidance on adaptive designs,[42] which states “It is widely understood that multiple analyses of the primary endpoint can […] lead to biased estimation of treatment effects on that endpoint. Less well appreciated, however, is that […] biased estimation can also apply to any endpoint correlated with the primary endpoint”.

As highlighted in the benefit-risk analysis literature, there can often be a trade-off between different outcomes when developing and evaluating an intervention.[59,60] In an example reported by Briel et al,[48] a trial assessing the effectiveness of vitamin E in premature newborns was stopped early based on an interim analysis of approximately half of the total number of participants planned at the start of the trial. This early analysis showed a reduction in intracranial hemorrhage.[61] However, a later evidence synthesis showed that this trial failed to identify that vitamin E increases the risk of sepsis.[62] Failing to accurately estimate treatment effects on key secondary endpoints could result in an intervention being adopted whose safety is overestimated or whose side effects are underestimated.

2.1.3 | Meta-analysis and evidence synthesis literature

Meta-analysis and evidence synthesis provide frameworks for combining the results from several studies[63] and are useful tools for providing an overall understanding of what the synthesized research has found.[64] In a review of 143 trials using adaptive designs that stopped early, Montori et al[65] found that few evidence syntheses and meta-analyses that included these trials considered the possible biases resulting from using these designs.

Several authors have argued that failing to account for adaptive designs in a meta-analysis or evidence synthesis can introduce bias.[66,67,49,50] Cameron et al[68] explored the impact of adaptive designs in a network meta-analysis. The authors considered three alternative methods to convert outcome data derived from an adaptive design to non-adaptive designs and found that failing to account for different study designs could bias estimates of effect in a network meta-analysis. Additionally, Walter et al[52] suggest that the estimate of treatment benefit can be calculated more accurately by applying weights to subgroups of studies with and without stopping rules.

However, there are several authors that suggest the biases are minimal[53,69-72] including Schou et al[73] who argue that removing truncated trials from a meta-analysis leads to substantial bias, whereas including these trials does not introduce large biases. The authors therefore recommend that all studies regardless of whether they stop early should be included in meta-analyses. Finally, Marschner et al[74] and Luchini et al[75] provide guidance on how sensitivity analyses may be conducted to explore the impact of trials that stopped early for benefit in a meta-analysis in line with CONSORT and GRADE[76] reporting checklists.

2.1.4 | Health economics

Increasingly clinical trials are designed with health economic objectives in mind, so that related outcomes are collected to inform a health economic analysis following the trial.[77] This may include clinical data on primary and secondary outcomes to inform parameters in a health economic model or costs and quality of life data collected directly from participants during the trial.[78]

Marschner and Schou[79] discuss the underestimation of treatment effects in sequential clinical trials when they do not stop early for benefit. The authors highlight the importance of an unbiased estimate of the treatment effect for cost-effectiveness analyses using a reanalysis of the GUSTO study.[80,81] They show that the treatment effect may have been underestimated and the experimental therapy appeared less cost-effective than it actually was. Flight[82] showed, using a simulation study, that when there are high levels of correlation between primary and health economic outcomes collected during a group sequential design, bias is introduced in the point estimates (and confidence intervals) of health economic outcomes. The levels of bias may be reduced in a model-based health economic analysis but this will depend on several factors such as the data structure, correlation, and adaptive design used.

A review by Flight et al[83] found no health economic analyses were adjusted following a group-sequential design. This potentially compromises decision-making if a decision to fund a treatment is based on biased estimates of cost-effectiveness. Additionally, patients may be penalized when a treatment is not funded based on an underestimate of cost-effectiveness, or resources may be wasted based on an overestimate of cost-effectiveness. Flight[82] extended the bias-adjusted maximum likelihood (ML) estimate approach proposed by Whitehead[84] to health economic outcomes and illustrated how this can reduce bias in a health economic analysis following an adaptive design.

2.2 | The magnitude of the problem

In this subsection, we discuss the extent of the bias in point estimates of treatment effects as a result of interim monitoring or data-dependent stopping of an adaptive design. This might be due to the pre-specified treatment selection criteria or other stopping rules, that is, lack-of-benefit (futility) and/or efficacy boundaries. More generally, a number of authors have discussed how the correlation between the MLE and the random design features in an adaptive design leads to bias in the MLE.[85-87] The random design features are features of the design (eg, the number of treatment arms, sample size, allocation ratio) that are determined by the accumulating trial data and are therefore considered a random variable. As an example that we expand on below, in a two-stage group sequential design the final sample size N is a random variable that is equal to N_1 if the trial stops at stage 1 or N_2 if the trial stops at stage 2. If $N = N_1$ then the MLE tends to be larger than the true treatment effect, whereas if $N = N_2$ then the MLE tends to be smaller than the true treatment effect (see Figure 1). This may be interpreted as correlation between the MLE and the random design, which leads to conditional bias in the MLE.

Indeed, it has previously been shown that the average treatment effect from a group sequential design is conditionally biased when the trial terminates early.[51,52,88] This conditional bias generally tends to be larger the “earlier” the selection happens, that is, when the decision to stop the treatment arm or continue to the next stage is based on a relatively small amount of information. By information, we mean statistical information which is driven by the number of participants in trials with continuous and binary outcomes, and the number of primary outcome events in trials with time-to-event outcomes. However, the degree of any potential bias will depend on the stopping rules, that is, how likely it is to stop the trial early, as well as the underlying treatment effect, as we now illustrate using results from Walter et al.[52]

Consider a two-stage group sequential design using one-sided OBF efficacy stopping boundaries (see Section 4 for a formal definition), where the interim analysis is conducted when 50% of the sample have provided outcome data (ie, 50% information fraction). We assume that the study outcome variable is normally distributed with known SD equal to one. Figure 1 shows the bias of the estimated treatment effect and the probability of early stopping as the true treatment effect μ varies from -0.5 to 1. The lines labeled “large”, “medium” and “small” correspond to studies with sample sizes of $N = 620$, 100, and 40, respectively (which give 80% power when $\alpha = 0.05$ to detect treatment effects of size $\mu = 0.2$, 0.5, and 0.8).

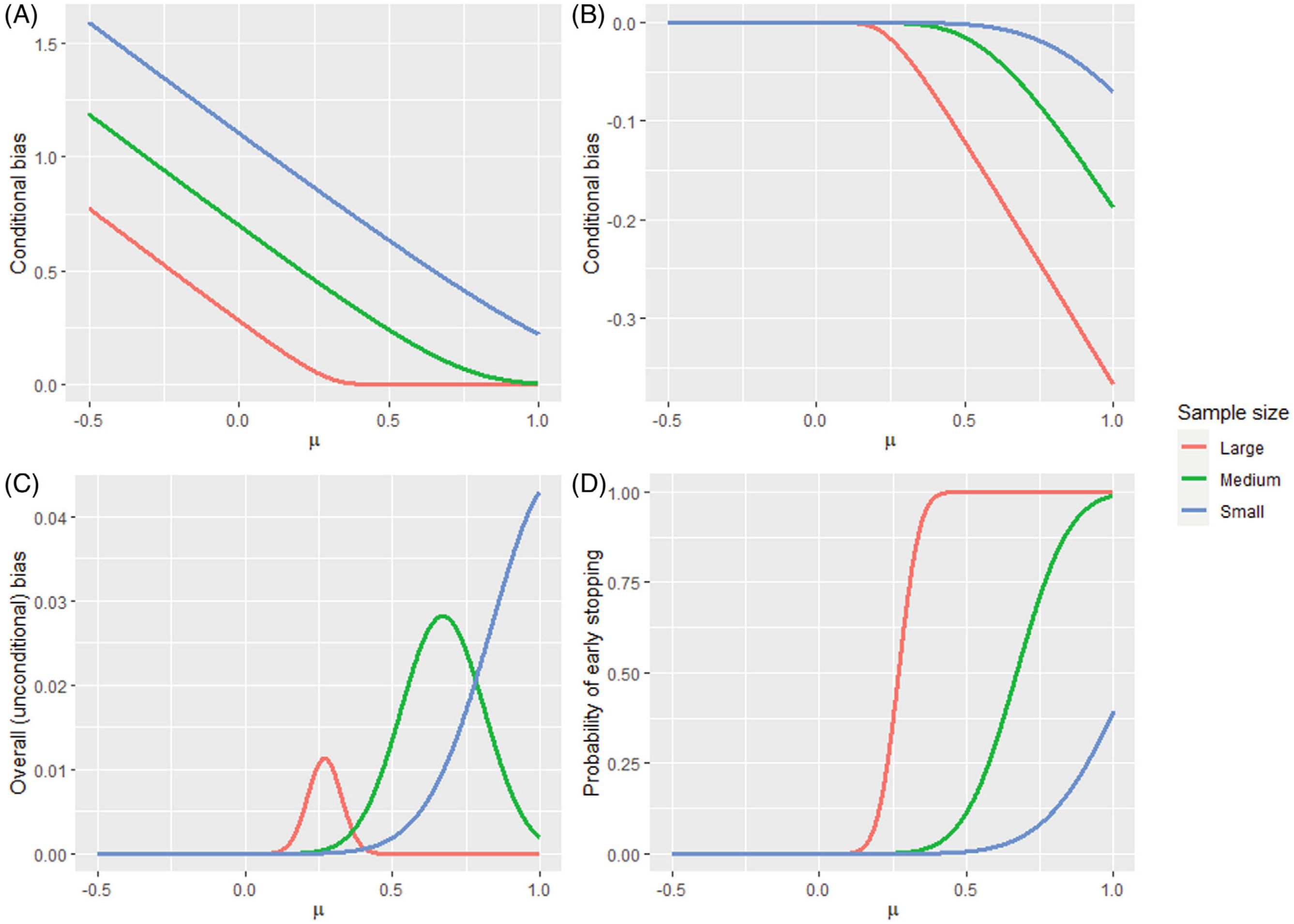

FIGURE 1 Bias of the average treatment effect and probability of early stopping for a two-stage group sequential design using one-sided O'Brien-Fleming efficacy stopping boundaries under different sample sizes (small, large, medium), with overall $\alpha = 0.05$. The interim analysis *P*-value threshold for efficacy is 0.0088. (A) The expected over-estimation in trial realizations that stop early for overwhelming efficacy (ie, conditional bias), (B) The expected under-estimation in trial realizations that do not stop early for overwhelming efficacy (ie, conditional bias at the final analysis), (C) the overall (unconditional) bias, and (D) the probability of early stopping.

Panel A shows that there can be a substantial positive conditional bias in the estimated treatment effect for trials that stop early for efficacy, with the magnitude of this bias increasing as the sample size decreases. Conversely, Panel B shows that there is a smaller negative conditional bias in the estimated treatment effect for trials that continue to the second stage (ie, do not stop early), with the magnitude of this bias increasing as either the trial sample size or true treatment effect increases. Panel C shows the overall (ie, unconditional) bias in the estimated treatment effect across all trials, which is small but positive and particularly noticeable for small sample sizes.

These results need to also be interpreted in light of Panel D, which shows that there is a very small chance of stopping for efficacy for small treatment effects when the sample size of the trial is itself small or medium. Hence, although there is a large positive conditional bias for such trials that stop early for efficacy (Panel A), and these results may then be used for recommending the adoption of a new treatment, such an event is very unlikely (Panel D). Similarly, although the negative conditional bias in trials that do not stop early can be substantial for large treatment effects (Panel B), the probability of such an event is negligible for large sample sizes (Panel D).

Previous empirical studies[89] showed that in designs with lack-of-benefit stopping boundaries the size of the selection bias for trials that reach the final stage is generally small. In fact, the bias is negligible if the experimental arm is truly effective. For trials that stopped early for lack of benefit, by definition a claim that the study treatment is better than the control is not made. Therefore, the fact that the treatment effect estimate is biased may be of less importance though results are useful in evidence synthesis. Furthermore, in designs that utilize an intermediate outcome for treatment selection it has been shown that this reduces the selection bias in the estimates of treatment effects in both selected and dropped treatment arms. In these designs, the degree of bias depends on the correlation between the intermediate and definitive outcome measures and this bias is markedly reduced by further patient follow-up and reanalysis at the planned 'end' of the trial.[89]

In drop-the-loser designs, where treatment selection is done based on the relative performance (eg, ranking) of research arms, the average treatment effect will be overestimated in the treatment arms that continue to the next stage, and will be underestimated in deselected treatment arms. In these designs, the degree of bias depends strongly on the true underlying treatment effects of the research arms, which is always unknown, the timing of treatment selection as well as the number of treatment arms to select from. Generally speaking though, in many scenarios there is a fundamental dilemma in that "the better the selection, the worse the bias".[46]

It has been shown that in drop-the-loser designs, the bias tends to be largest, and confidence intervals have incorrect coverage, where the underlying treatment effects of the treatment arms are equal, for example when all arms are under the global null or global alternative hypothesis.[24,90] More generally, bias will be smaller where the underlying effects differ amongst the treatment arms than when they are similar. In contrast, when one treatment arm has a distinctly larger treatment effect than those it is competing against, bias in the final average treatment effect is minimal for the arm which is performing best when the selection takes place. Moreover, the number of treatment arms to select from was also found to increase the degree of bias in the average treatment effects in pick-the-winner designs[91] (a MAMS variant). Finally, early stopping rules that are binding increase bias compared to a design with no early stopping rules, particularly under the global null hypothesis or where only one treatment arm had the target treatment effect.[46]

3 | REVIEW OF CURRENT PRACTICE

Given that there are a variety of potential negative impacts of reporting biased estimates following an adaptive design, we reviewed the relevant literature to understand how often methods for reducing or removing bias in point estimates (as described in part I of this article series) are used in practice. We focused on results reported from adaptive trials using the different types of adaptations described in Section 1.

3.1 | Search strategy

A review of group sequential trials was based on pre-existing reviews known to the authors by the May 30, 2022. For other adaptive designs, we systematically searched the MEDLINE database via PubMed (from Jan 1, 2000 to May 30, 2022) using the following search terms (Table 1).

3.2 | Results

Group sequential designs

For group sequential designs, Stevely et al[47] identified 68 trials that were published in leading medical journals, of which 13 (19%) were multi-arm trials. A total of 46/68 (68%) were stopped early, primarily either for efficacy (10/46, 22%) or futility (28/46, 61%). Of these trials, only 7% (3/46) disclosed the use of some form of bias correction. A subsequent review of 19 two-arm group sequential trials[92] in oncology that were stopped early found that none of these applied

TABLE 1 Search strategy for the initial database search of MEDLINE.

Type of adaptive design	Search terms
MAMS designs	[("multi-stage") OR ("multi stage") OR ("multi-arm multi-stage") OR ("multi arm multi stage") OR (two-stage) OR ("two stage") OR ("pick the winner") OR ("pick-the-loser") OR ("drop the loser") OR ("drop-the-loser") OR ("dose selection") OR ("seamless")]
RAR designs	[("response adaptive") OR ("response-adaptive") OR ("adaptive randomization") OR ("adaptive randomization") OR ("outcome adaptive") OR ("outcome-adaptive")]
Adaptive enrichment designs	[("adaptive enrichment") OR ("population enrichment") OR ("patient enrichment") OR ("enrichment design") OR ("biomarker-adaptive") OR ("biomarker adaptive") OR ("subgroup selection") OR ("subpopulation selection") OR ("enrichment")]

Abbreviations: MAMS, multi-arm multi-stage; RAR, response-adaptive randomisation.

TABLE 2 Summary of systematic search of adaptive designs from Jan 1, 2000 to May 30, 2022.

Type of adaptive design	Number of records screened	Number of randomized trials that reported results	Number of randomized trials that reported an unbiased or bias-adjusted estimate
MAMS	773	22	2
RAR	59	17	0
Adaptive enrichment	530	3	0

any bias correction to the estimated hazard ratios. Case studies also highlighted the routine lack of use of bias-corrected estimators in trials that are stopped early.[93] In summary, in trials that use group sequential designs, bias-adjusted methods are rarely used in practice and the implications are not well known.[47,92-94]

MAMS trials

A total of 765 records were retrieved and screened for MAMS trials. Only 14/765 (1.8%) reported results. These articles were supplemented by an additional 8 trials from related work[3] within the same search period. As a result, we reviewed 22 eligible MAMS trials; phase II (n = 10), phase II/III (n = 9), and phase III (n = 3). The vast majority of trials (95.5%, 21/22) had at least one treatment arm that was dropped early but the trial continued beyond the first interim analysis and 81.8% (18/22) used frequentist methods. Only 2/22 (9.1%) of the trials reported the use of a bias-adjusted point estimator of the treatment effect. One described the use of a bias-adjustment for the mean of the selected dose in stage 2 (which was then used to derive an adjusted t-test)[95] while the other used a uniformly minimum variance conditionally unbiased estimator.[96]

RAR designs

There were 59 records retrieved that used a RAR design; of which 22 were randomized trials. Of these 22, 17 were reporting interim (n = 4), both interim and final results (n = 1) and final results (n = 12); phase II (n = 10), phase I/II (n = 2), phase II/III (n = 1), phase III (n = 3) and phase IV (n = 1). Of the 5 that reported interim results, 2 were stopped early for futility, 2 had a treatment that graduated (ie, was declared successful) for further evaluation at phase III, and the remaining was stopped early for efficacy. Most of the trials (76.5%, 13/17) used Bayesian methods with the remaining using hybrid (frequentist and Bayesian, n = 2) and frequentist methods (n = 2). None of the trials used bias correction methods.

Adaptive enrichment trials

Of the 528 records screened, only 1 trial was an adaptive enrichment trial reporting results. There were an additional 2 known trials not retrieved in the search. Only 1 of these 3 trials reported interim results and was stopped early for futility. Enrichment was triggered in 1 trial. All 3 trials used frequentist methods and none used any bias correction methods.

A summary of the results of the systematic search is given in Table 2 As can be seen, across the adaptive designs considered, unbiased or bias-adjusted treatment effect estimates are currently rarely used and reported.

These results are in stark contrast with some of the recommendations for best practice that have been made around bias-adjusted analyses in widely-used guidance on adaptive designs (namely, the FDA guidance,[42] and the Adaptive designs CONSORT Extension[3,4](ACE)). The FDA guidance stresses the importance of using “methods for adjusting estimates to reduce or remove bias associated with adaptations and to improve on performance such as the mean squared error”, stating that “Such methods should be prospectively planned and used for reporting results when they are available”. In particular, for group sequential designs “a variety of methods exist to compute estimates and confidence intervals that appropriately adjust for the group sequential stopping rules … To ensure the scientific and statistical credibility of trial results and facilitate important benefit-risk considerations, an approach for calculating estimates and confidence intervals that appropriately accounts for the group sequential design should be prospectively planned and used for reporting results.” As for the CONSORT Extension, the guidance is much less prescriptive, but nevertheless recommends the discussion of “the implications of … potential bias and imprecision of the treatment effects if naïve estimation methods were used”. The relevant parts of both these guidance documents are quoted in full in Appendix A.1.

4 | CASE STUDY: GROUP SEQUENTIAL DESIGN

In this section, we illustrate how different types of unbiased and bias-reduced estimators (as reviewed in part I of this article series) can be used in practice for a group sequential design that uses OBF stopping boundaries, which we now briefly describe. In a group sequential design, participants are allocated to the treatments and the accumulating data are analyzed after each complete group of data becomes available. When using OBF efficacy boundaries, the nominal significance levels needed to reject the null hypothesis increases as the trial progresses, that is, it is more difficult to reject H_0 at earlier analyses. Given the standardized test statistics Z_k for group $k = 1, \dots, K$, the one-sided OBF boundaries and stopping rules take the following form[11]:

After group $k = 1, \dots, K-1$.

if $Z_k \geq C(K, \alpha)\sqrt{(K/k)}$ stop, reject H_0.

otherwise continue to group $k + 1$.

After group K.

if $Z_k \geq C(K, \alpha)$ stop, reject H_0.

otherwise stop, do not reject H_0.

Here, the values of $C(K, \alpha)$ are chosen to ensure that the overall type I error probability for the K stage trial is controlled at preset level α.

We use a group sequential design as our case study in order to illustrate the widest range of different estimators (both unconditional and conditional). As well, some adaptive designs (eg, certain types of MAMS designs) can be viewed as an extension of group sequential designs and therefore we can illustrate more general underlying principles of point estimation.

We consider the phase III MUSEC (multiple sclerosis and extract of cannabis) trial, as described by Bauer et al[97] and Zajicek et al.[98] This is an example of a two-stage group sequential design where the trial continued to the second stage (as the criterion for early stopping was not met). The MUSEC trial investigated the effect of a standardized oral cannabis extract (CE) on muscle stiffness for adult patients with stable multiple sclerosis. The primary endpoint was a binary outcome—whether or not a patient had "relief from muscle stiffness" after 12 weeks of treatment, based on a dichotomized 11-point category rating scale. A two-stage group sequential design with early stopping for superiority using the OBF boundary was used, with a pre-planned maximum total sample size of 400 subjects (200 per arm) and an unblinded interim analysis planned after 200 subjects (100 per arm) had completed the 12 week treatment course.

In the actual trial, an unblinded sample size re-estimation based on conditional power considerations[97] was also carried out at the interim analysis, which reduced the total planned sample size from 400 to 300. For the purpose of illustrating the calculation of a larger range of adjusted estimators, we ignore this sample size re-estimation in what follows. If we were to take into account the sample size re-estimation then the methods for adaptive group sequential designs would apply (see Section 6 of part I of this series), but only median-unbiased estimators are available in that setting. For more general guidelines around best practice in this context, see Section 5.

Table 3 summarizes the observed data from the trial at the interim and final analyses, as well as the standardized test statistics and the OBF efficacy stopping boundaries. As can be seen, at the interim analysis the boundary for early rejection of the null hypothesis (no difference in the proportion of subjects with relief from muscle stiffness between treatment arms) was almost reached, with the standardized test statistic being close to the stopping boundary.

TABLE 3 Observed data from the MUSEC trial at the interim and final analyses, with standardized test statistics and O'Brien-Fleming (OBF) group sequential boundary (one-sided, with early stopping only for superiority).

	Interim data		**Final data**	
	Placebo	**CE arm**	**Placebo**	**CE arm**
Number of subjects with relief from muscle stiffness	12	27	21	42
Total number of subjects	97	101	134	143
Standardized test statistic	2.540		2.718	
OBF boundary	2.797		1.977	

4.1 | Calculation of unbiased and bias-adjusted estimators

Using the observed data from the MUSEC trial, we now demonstrate how to calculate various unbiased and bias-adjusted estimators for the treatment difference, from both a conditional and unconditional perspective. More formally, letting p_{CE} and p_0 denote the response probability for patients on CE and the placebo respectively, we consider estimators of $\theta = p_{CE}$ - p_0. R code to obtain these estimators is provided in the Supporting Information.

The conventional end-of-trial estimator for the treatment difference, that is, the overall MLE, is given by $\hat{\theta}_{obs} = \hat{p}_{CE} - \hat{p}_0$, where $\hat{p}_{CE}$ and $\hat{p}_0$ are the observed proportions of successes on the CE and placebo arms respectively. In what follows, it is also useful for illustrative purposes to consider the MLE calculated just using the interim data (stage 1 data), denoted $\hat{\theta}_1$, as well as the MLE calculated just using the stage 2 data (ie, only the data from after the interim analysis), denoted $\hat{\theta}_2$. These estimators are inefficient (and potentially unethical) since they “discard” patient data, so we are not recommending that these are used in practice.

Unconditional perspective

From an unconditional perspective, we want to estimate θ regardless of the stage that the trial stops, and are interested in the bias as averaged over all possible stopping times, weighted by the respective stage-wise stopping probabilities. More formally, letting T be the random variable denoting the stage that the trial eventually stops, we define the unconditional bias of an estimator $\hat{\theta}$ as

$$bias(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta = \sum_{k=1}^{2} E_\theta[\hat{\theta}|T = k] \Pr_\theta [T = k] - \theta.$$

In the two-stage trial setting, Emerson[99] presented an analytical expression for this bias of the overall MLE, which depends on the unknown value of θ:

$$bias_\theta(\hat{\theta}) = \frac{I_2 - I_1}{I_2\sqrt{I_1}} \phi\left(e - \theta\sqrt{I_1}\right),$$

where e denotes the efficacy stopping boundary, I_1 and I_2 denote the (observed) information at stage 1 and stage 2 respectively and ϕ denotes the probability density function (pdf) of a standard normal distribution. The full definitions of the information I_1 and I_2 for our trial context are given in Appendix A.2, which depend on the number of observed successes. Following Whitehead,[84] we can use this expression to calculate an unconditional bias-corrected MLE $\tilde{\theta}$ (UBC-MLE), which is the solution of the equation $\tilde{\theta} = \hat{\theta}_{obs} - bias_{\tilde{\theta}}(\hat{\theta}_{obs})$.

Alternatively, the uniformly minimum variance unbiased estimator (UMVUE) can be calculated by using the Rao-Blackwell technique on the stage 1 MLE $\hat{\theta}_1$, which is unconditionally unbiased (see Section 4.1 of part I of this article series for further details). More formally, the UMVUE in our two-stage trial context is given by $E[\hat{\theta}_1 \mid (T, \hat{\theta}_{obs})]$, with the following closed-form expression when $T = 2$:

$$UMVUE = \hat{\theta}_{obs} - \frac{\sqrt{I_2 - I_1}}{\sqrt{I_1 I_1}} \frac{\phi\left(\frac{e - Z_2\sqrt{I_1/I_2}}{\sqrt{(I_2 - I_1)/I_2}}\right)}{\Phi\left(\frac{e - Z_2\sqrt{I_1/I_2}}{\sqrt{(I_2 - I_1)/I_2}}\right)},$$

where Z_2 is the (observed) overall standardized test statistic at stage 2, and Φ represents the cumulative distribution function (cdf) of a standard normal distribution.

A median unbiased estimator (MUE) can also be calculated, which depends on a choice of the ordering of the sample space with respect to evidence against the null hypothesis (see Section 4.2 of part I of this article series for further details). In what follows, we use stagewise ordering, which has desirable properties described by Jennison and Turnbull.[11] This allows the use of the *P*-value function $P(\theta)$ to find the MUE, which is the solution to the equation.

$P(\hat{\theta}_{MUE}) = 0.5$. The formula for the *P*-value function (for a trial that continues to the second stage) is as follows:

$$P(\theta) = \int_{-\infty}^{e} \int_{z_2}^{\infty} f_2\left((x_1 x_2), \left(\theta\sqrt{I_1}\theta\sqrt{I_2}\right), \begin{pmatrix} 1 & \sqrt{I_1/I_2} \\ \sqrt{I_1/I_2} & 1 \end{pmatrix}\right) dx_2 dx_1,$$

where $f_2\left((x_1 x_2),\mu,\Sigma\right)$ is the density of a bivariate normal distribution with mean μ and covariance matrix Σ evaluated at the vector (x_1, x_2). See also the R code provided in the Supporting Information.

Conditional perspective

From a conditional perspective, we are interested in estimation conditional on the trial continuing to stage 2. We define this conditional bias of an estimator $\widehat{\theta}$ as $E[\,\widehat{\theta} \mid T = 2] - \theta$. In the context of group sequential trials, as argued by several authors,[100-104] the conditional bias of an estimator is also an important consideration: given that the study has in fact stopped with $T = 2$, we can use this knowledge in our bias calculations. As well, while the unconditionally unbiased estimators are unbiased overall, they tend to overestimate the treatment effect when there is early stopping and underestimate the effect when the trial continues to the end. The authors see value in both the conditional and unconditional perspective. As the unconditional estimators tend to be biased once the stopping reason is known they do, however, have a slight preference for conditional estimators. Nonetheless, there is no consensus in the literature and we provide a few example quotations illustrating this in Appendix A.2.

We can calculate an analytical expression for the conditional bias of the overall MLE (ie, at the final stage), which is given below and again depends on the unknown true parameter θ:

$$\text{Conditional bias}_{\theta}(\widehat{\theta}) = -\frac{\sqrt{I_1}}{I_2}\,\frac{\phi\left(e - \theta\sqrt{I_1}\right)}{\Phi\left(e - \theta\sqrt{I_1}\right)}.$$

Using this expression, we can calculate a conditional bias-corrected MLE $\widetilde{\theta}_c$ (CBC-MLE), which is the solution of the equation $\widetilde{\theta_c} = \widehat{\theta}_{obs} - \text{conditional bias}_{\widetilde{\theta}_c}\left(\widehat{\theta}_{obs}\right)$. As well, the uniformly minimum variance conditionally unbiased estimator (UMVCUE) can be calculated by again using the Rao-Blackwell technique, but this time applied to the stage 2 MLE $\widehat{\theta}_2$ (ie, excluding the stage 1 data), which is conditionally unbiased. See Section 4 of part I of this article series for details. More formally, the UMVCUE is given by $E[\widehat{\theta}_2 \mid (T = 2, \widehat{\theta}_{obs})]$, resulting in a closed-form expression as follows:

$$UMVCUE = \widehat{\theta}_{obs} - w_1 \frac{\phi\left(w_2\left(\widehat{\theta}_{obs} - e/\sqrt{I_1}\right)\right)}{\Phi\left(w_2\left(\widehat{\theta}_{obs} - e/\sqrt{I_1}\right)\right)},$$

where $w_1 = \frac{1}{(I_2 - I_1)\sqrt{I_1^{-1} + (I_2 - I_1)^{-1}}}$, $w_2 = I_1\sqrt{I_1^{-1} + (I_2 - I_1)^{-1}}$.

Finally, a conditional MUE (CMUE), $\widehat{\theta}_{CMUE}$, can be calculated following Koopmeiners et al[105] and Grayling and Wason[106] It is defined as the value of θ that gives a conditional median equal to the observed overall MLE. More formally, it is the solution to the following equation:

$$0.5 = \int_{-\infty}^{\widehat{\theta}_{obs}} f(\widehat{\theta} \mid T = 2)\, d\widehat{\theta},$$

where $f(\widehat{\theta} \mid T = 2)$ is the conditional density of the overall MLE (conditional on continuing to stage 2), see Appendix A.2 for further details.

Table 4 gives the values of all of the estimators described above, calculated using the observed data and OBF stopping boundaries from the MUSEC trial. For illustration purposes, we also calculated the Standard Error (SE) for all estimators using a parametric bootstrap approach assuming that the true unknown difference in proportions (θ) is equal to 0.14, which should be treated with caution as they will vary depending on this key assumption—see Section 4.2 for a simulation study that gives the values of the SEs for different assumed values of θ. R code for calculating the SEs is given in the Supporting Information.

The overall MLE is 0.1370 (with a SE of 0.054), and is the comparator for all the other estimators in Table 4 since it is the conventional end-of-trial point estimate. Starting with the unconditionally unbiased and bias-adjusted estimators, the stage 1 MLE is slightly larger (0.1436), but this is based on only the stage 1 data and hence is slightly inefficient: it has an information fraction of 0.795 and a SE of 0.057. The MUE, UBC-MLE and UMVUE are all slightly lower than the MLE, although they are all within 0.01 in absolute terms (ie, within 7% in relative terms). This downward correction is intuitive – we would expect the MLE to overestimate the magnitude of θ averaged over the possible stopping times: if $\widehat{\theta}_1$

TABLE 4 Naive, unconditionally and conditionally unbiased/bias-adjusted estimates calculated using the observed data and O'Brien-Fleming efficacy stopping boundaries from the MUSEC trial.

Type of estimator	Estimator	Difference in proportions (SE)	Relative difference to overall MLE
MLE/naive	MLE (overall)	0.1370 (0.054)	–
Unconditionally unbiased/bias-adjusted	MLE (stage 1)	0.1436 (0.057)	+5%
	Median unbiased estimator (MUE)	0.1341 (0.054)	−2%
	UMVUE	0.1278 (0.054)	−7%
	Bias-corrected MLE (UBC-MLE)	0.1328 (0.055)	-3%
Conditionally unbiased/bias-adjusted	MLE (stage 2)	0.1139 (0.111)	−17%
	Conditional MUE (CMUE)	0.1851 (0.080)	+35%
	UMVCUE	0.1724 (0.071)	+26%
	Bias-corrected MLE (CBC-MLE)	0.1909 (0.073)	+39%

Note: Standard errors (SEs) are calculated using a parametric bootstrap approach with 10^5 replicates, assuming that the true difference in proportions is equal to 0.14.

is sufficiently larger than θ, the trial stops with $T = 1$ and the MLE equals $\hat{\theta}_1$, whereas if $\hat{\theta}_1$ is lower than θ by a similar amount, the trial can continue, allowing the stage 2 data to reduce the negative bias of the overall MLE. The SEs of these estimators are all very similar to that of the MLE under the assumption of a true difference in proportions of 0.14, reflecting the small corrections to the MLE. Finally, we see that MUE > UBC-MLE > UMVUE, which reflects the fact that the MUE is not mean-unbiased, and the UBC-MLE will also be expected to have residual mean bias as it is not exactly mean-unbiased (see Section 4.2 for simulation results).

Moving on to the conditionally unbiased and bias-adjusted estimators, the stage 2 MLE is substantially lower (0.1139) than the overall MLE (and indeed any of the other estimators conditional or unconditional). However, the information fraction for stage 2 was only 0.205 ignoring the 0.795 from stage 1, and hence this estimator has a substantially higher variability with a (conditional) SE of 0.111. The CMUE, CBC-MLE, and UMVCUE are noticeably larger than the overall MLE (in relative terms 35%, 39%, and 26% larger respectively). An upward correction is intuitive from a conditional perspective: there is downward selection pressure on the stage 1 MLE $\hat{\theta}_1$, since if $\hat{\theta}_1$ is sufficiently larger than θ then the trial does not continue to stage 2. Given that the stage 1 MLE was almost large enough for the standardized test statistic to cross the OBF stopping boundary (note that on the test statistic scale, the OBF stopping boundary at stage 1 was 0.1581), the relatively large correction to the overall MLE is not too surprising. These three estimators have substantially lower (conditional) SE than the stage 2 MLE, since they are utilizing all of the trial data. However the conditional SEs are larger than the unconditional ones. This is a general property of conditional estimators: by conditioning, the information that is contained in the statistic that is conditioned on (in this case, the stopping stage) is lost. Finally, we see that both the CMUE and CBC-MLE are larger than the UMVCUE, which again reflects the residual mean (conditional) bias in the CMUE and CBC-MLE (see Section 4.2 for simulation results).

As pointed out by an anonymous reviewer, the observed stage at which the trial stops (or more generally, the observed study design) will imply something about the treatment effect which can then be taken into account. In our case study, the study proceeds to the second stage, which implies that the MLE is on average conditionally biased downwards[101] (see Panel B in Figure 1). Therefore, it arguably is undesirable to apply an unconditional adjustment that adjusts the MLE further downwards. Instead, a conditional approach may make more sense because it will adjust the MLE upwards to take account of the fact that the MLE has a conditional negative bias (see Panel B in Figure 1). These trends are all reflected in the results presented in Table 4. See also the discussion on the conditional vs unconditional perspective in Section 5.1.

In summary, for the MUSEC trial data, the use of different estimators can give noticeably different values for the estimated treatment effect, particularly when considering a conditional vs unconditional perspective. This could influence the interpretation of the trial results in certain cases, and highlights the importance of pre-specifying which estimator(s) will be reported following an adaptive design. The choice of estimator(s) will depend on what the researchers wish to achieve regarding the estimand in question. There will be pros and cons for the different estimators, one key example being the bias-variance trade-off. We explore these issues further in Section 5. We also note that there is a strong link between design and estimation—the estimated values above depend on the design of the trial, and would be different if (eg,) the design had also included futility stopping boundaries.

4.2 | Simulation study

Since the point estimates presented above represent one realization of the trial data given the trial design, in this subsection we carry out a simulation study to investigate the performance of the estimators under different scenarios. We stress however that (unlike when calculating the standard errors of the estimates) we have *not* used the assumed unknown value for the underlying treatment effect (θ) to calculate the unbiased and bias-adjusted estimates in Table 4. As can be seen from the formulae in Section 4.1, these estimators do not depend explicitly on θ, but only on the observed data and efficacy stopping boundary (in this case). Hence, the point estimates presented in the third column of Table 4 would not change under different values of θ assuming that the realized trial data remained the same.

To demonstrate the bias-variance properties of the estimators when averaged over *many* trial realizations following the two-stage design of the MUSEC trial, we ran simulations under different values of θ (0.10, 0.14, and 0.18). To do so, we used the (asymptotic) canonical joint distribution of the standardized test statistics at each stage,[11] which is bivariate normal—for further details, see the R code included in the Supporting Information. For further comprehensive simulation results comparing the different estimators in the context of two-stage group sequential trials, we refer the reader to Grayling and Wason.[106]

For each value of θ, we simulated 10^5 trial replicates and calculated the mean values of the point estimators as well as their standard errors across the trial replicates. Note that the unconditional estimators are all equal to the stage 1 MLE for the trial realizations that stop at the interim analysis. This is by definition for the overall MLE (since this is the MLE calculated at the stage the trial stops at) as well as the MUE and the UMVUE, while the UBC-MLE would depend on the unknown value of I_2 and hence we simply set the UBC-MLE equal to the observed stage 1 MLE. As for the conditional estimators, these are all conditional on the trial continuing to stage 2 (and so are calculated using trial realizations that continue to stage 2, see the R code for further details). The simulation results are displayed in Table 5 below.

The results for the mean values of the point estimators are what we would expect: on average, the unbiased estimators are unbiased whereas the naïve and bias-adjusted estimators have a (small) positive bias due to the early stopping for efficacy. The overall MLE has the largest mean bias out of the unconditional estimators, although this is less than 0.005 in absolute terms across the different values of θ. The MUE and UBC-MLE have a residual positive bias (less than 0.003 in absolute terms), reflecting how they are not mean unbiased. Looking at the conditional estimators, the CMUE and CBC-MLE have a larger positive bias (up to 0.014 in absolute terms) while having a similar SE to the UMVCUE, and so would likely not be recommended for use in this trial context.

TABLE 5 Simulation results showing the mean values of the point estimators and the corresponding standard errors (SE) under different assumed values of θ. There were 10^5 trial replicates for each value of θ.

		Difference in proportions (SE)		
Type of estimator	**Estimator**	**$\theta = 0.10$**	**$\theta = 0.14$**	**$\theta = 0.18$**
MLE/naive	MLE (overall)	0.103 (0.054)	0.144 (0.054)	0.184 (0.053)
Unconditionally unbiased/bias-adjusted	MLE (stage 1)	0.100 (0.057)	0.140 (0.057)	0.180 (0.057)
	Median unbiased estimator (MUE)	0.101 (0.053)	0.142 (0.054)	0.182 (0.054)
	UMVUE	0.100 (0.052)	0.140 (0.054)	0.180 (0.055)
	Bias-corrected MLE (UBC-MLE)	0.101 (0.054)	0.142 (0.055)	0.183 (0.054)
Conditionally unbiased/bias-adjusted	MLE (stage 2)	0.100 (0.111)	0.140 (0.111)	0.180 (0.111)
	Conditional MUE (CMUE)	0.115 (0.083)	0.152 (0.080)	0.190 (0.081)
	UMVCUE	0.100 (0.062)	0.140 (0.071)	0.179 (0.080)
	Bias-corrected MLE (CBC-MLE)	0.111 (0.067)	0.154 (0.073)	0.194 (0.078)

Note: The probability of stopping at stage 1 was 0.15, 0.37 and 0.65 for $\theta = 0.10$, 0.14 and 0.18, respectively.

TABLE 6 Simulation results showing the mean values of the unconditional point estimators and the corresponding standard errors (SE) under different assumed values of θ, separated by trial replicates that stop at the interim analysis and those that continue to stage 2. There were 10^5 trial replicates in total for each value of θ.

		Difference in proportions (SE)		
	Estimator	**$\theta = 0.10$**	**$\theta = 0.14$**	**$\theta = 0.18$**
Trial stops early at the interim analysis	MLE (stage 1)	0.188 (0.025)	0.197 (0.031)	0.212 (0.038)
Trial continues to stage 2	MLE (overall)	0.087 (0.043)	0.113 (0.038)	0.132 (0.033)
	MLE (stage 1)	0.084 (0.045)	0.106 (0.038)	0.120 (0.030)
	Median unbiased estimator (MUE)	0.086 (0.041)	0.109 (0.034)	0.126 (0.027)
	UMVUE	0.084 (0.039)	0.106 (0.030)	0.120 (0.023)
	Bias-corrected MLE (UBC-MLE)	0.085 (0.041)	0.110 (0.036)	0.128 (0.032)

Note: The probability of stopping at stage 1 was 0.15, 0.37, and 0.65 for $\theta = 0.10$, 0.14, and 0.18, respectively.

Tables 4 and 5 demonstrate that while on average, the different point estimators will be close together, for a particular trial realization, the estimates may be quite different. The differences between the estimators we see in Table 4 are a consequence of the observed data for the MUSEC trial, which were quite "extreme" in the sense that the stage 1 MLE was very close to the stopping boundary and substantially larger than the stage 2 MLE.

As for the SEs, for the unconditional estimators, the stage 1 MLE has the highest SE, reflecting how it only uses the stage 1 data (with an information fraction of 0.795). The other unconditional estimators have very similar SEs, which change little as θ increases. For the conditional estimators, the stage 2 MLE has a substantially higher SE since it only uses the stage 2 data (with an information fraction of only 0.205). The UMVCUE and CBC-MLE have similar SEs (with the CMUE having a slightly larger SE); however, their SEs increase as θ increases. This reflects how the stage 1 and stage 2 data will be expected to have a larger discrepancy as θ increases, since the stage 1 MLE will have to be below the efficacy stopping boundary of 0.1581 in order for the trial to continue to the second stage.

Like in Section 2.2, it is informative to also report separate means for the unconditional estimators for the trial replicates that stop early at the interim analysis and those that continue to stage 2. Table 6 shows these mean values of the unconditional point estimators as well as their corresponding standard errors across 10^5 trial replicates. We see that for the trial replicates that stop at the interim analysis, the stage 1 MLE has a substantial positive conditional bias, particularly for $\theta = 0.10$ (but note that the probability of stopping in this case is only 0.15). Conversely, for the trial replicates that continue to stage 2, all of the unconditional estimators are conditionally biased, with a noticeable negative conditional bias across all three values of θ. These results need to be interpreted in light of the probability of stopping at stage 1, which was 0.15, 0.37 and 0.65 for $\theta = 0.10$, 0.14, and 0.18, respectively. Overall, the results again demonstrate that even if estimators are unconditionally unbiased, they may have considerable conditional bias.

Apart from summarizing the mean values and SEs, it is also useful to look at the whole sampling distribution of the point estimators. Figure 2 shows these sampling distributions, assuming $\theta = 0.14$ and with 10^5 trial replicates. For the unconditional estimators (except for the stage 1 MLE), the distributions are a mixture from the trial replicates that stopped in stage 1 and those that continued to stage 2 (recall that for those trial replicates that stopped in stage 1, all the unconditional estimators are equal to the stage 1 MLE). It is interesting to note that the sampling distribution of the MUE is substantially smoother than those of all the other estimators (ignoring the stage 1 MLE), particularly compared with the UMVUE and the overall MLE. As for the conditional estimators, the stage 2 MLE has a substantially wider sampling distribution than the others, reflecting how it uses less information. Meanwhile, the sampling distributions of the UMVCUE, CMUE and the CBC-MLE appear quite similar.

5 | GUIDANCE: BEST PRACTICES FOR POINT ESTIMATION IN ADAPTIVE DESIGNS

In this section, we give guidance on the choice of estimators and the reporting of estimates for adaptive designs. This builds on the relevant parts of the FDA guidance for adaptive designs[42] and the ACE.[3,4] The issue of estimation and potential bias

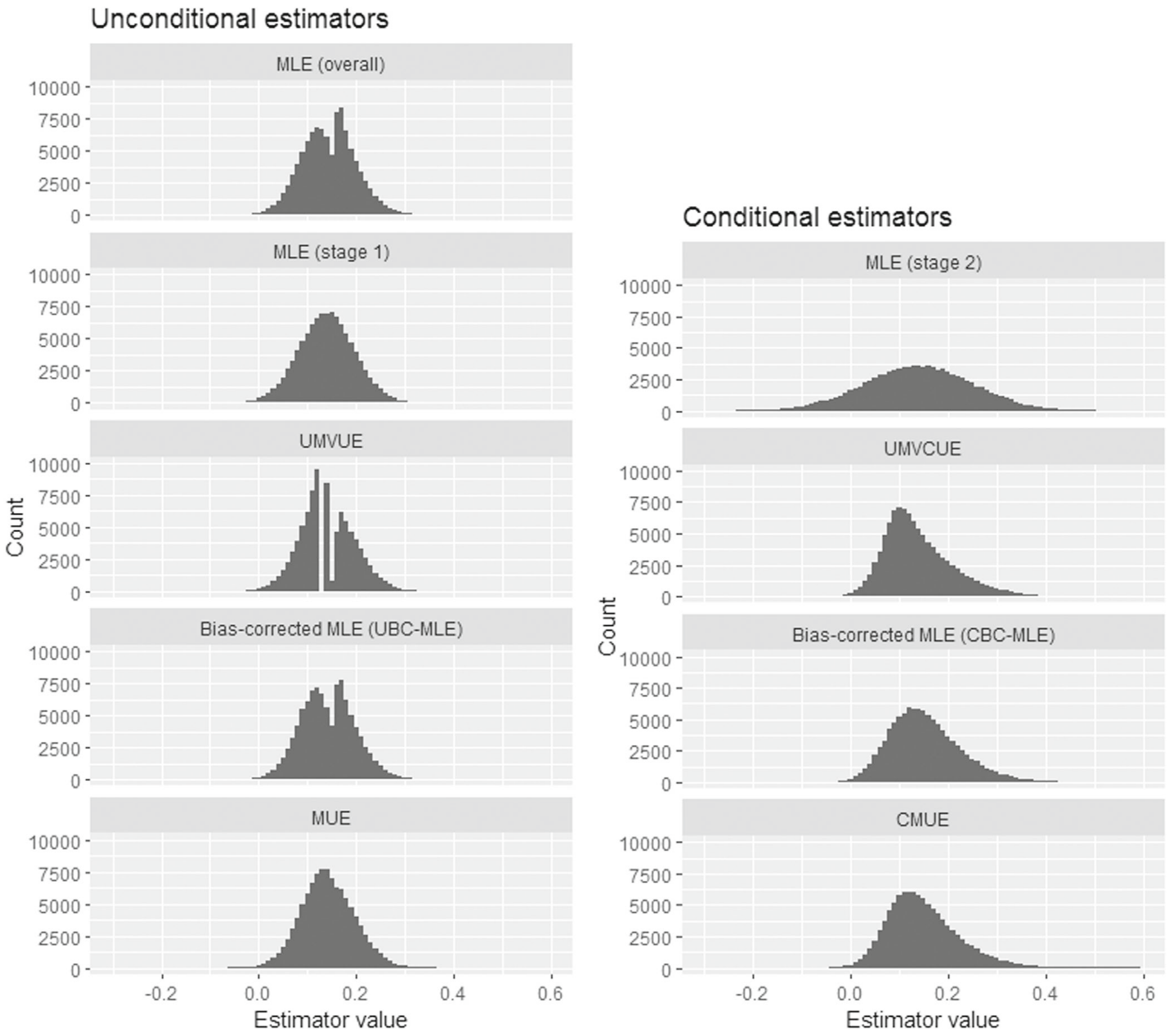

FIGURE 2 Sampling distributions of the point estimates from 10^5 trial replicates, assuming that $\theta = 0.14$. CMUE, conditional median unbiased estimator; MLE, maximum likelihood estimator; MUE, median unbiased estimator; UMVCUE, uniformly minimum variance conditionally unbiased estimator; UMVUE, uniformly minimum variance unbiased estimator.

should be considered throughout the whole lifecycle of an adaptive trial, from the planning stage to the final reporting and interpretation of the results. Indeed, the design and analysis of an adaptive trial are closely linked, and one should not be considered without the other. In what follows, our main focus is on the confirmatory setting where analyses are fully pre-specified, but some of the principles can also apply to more exploratory settings, particularly around the choice of estimators and the final reporting of trial results.

5.1 | Planning stage

The context, aims and design of an adaptive trial should all inform the analysis strategy used, which includes the choice of estimators. These decisions should not only be left to trial statisticians, but also be discussed with other trial investigators to ensure that it is consistent with what they want to achieve. Firstly, it is necessary to decide on what exactly is to be estimated (ie, the estimands of interest; see also Appendix A.2). Secondly, the desired characteristics of potential estimators should be decided. Two key considerations are as follows:

- *Conditional vs unconditional perspective*: The choice of whether to look at the conditional or unconditional bias of an estimator will depend on the trial design. For example, in a drop-the-losers trial where only a single candidate treatment is taken forward to the final stage, a conditional perspective reflects the interest being primarily in estimating the effect of the successful candidate. On the other hand, for group sequential trials, the unconditional perspective is recognized as being an important consideration (see Appendix A.2). As seen in the simulation study in Section 4.2, it can be the case that these different point estimators are on average similar over repeated realizations of the trial, but for a single realization are markedly different. As well, the standard errors of the conditional estimators can be larger than those for the unconditional estimators. More generally, (as pointed out by an anonymous reviewer) in situations where the observed design (see Section 2.2) implies a directionality to the MLE bias, the conditional estimation that takes this directionality into account may be preferable. Stopping a group sequential trial for efficacy at the first interim analysis (over-estimate) or at the final analysis (under-estimate) are examples of this situation.

  A new perspective on the question of conditional vs unconditional inference has recently been provided by Marschner.[87] This work presents a unifying formulation of adaptive designs and a general approach to their analysis, which is based on the partitioning of the overall unconditional information into its two component sources. More precisely, the unconditional likelihood can be expressed as the product of the "design likelihood" (ie, the information contained in the realized design) and the 'conditional likelihood' (conditioned on the realized design). Rather than advocating for or against unconditional inference over conditional inference in general, the framework allows for the exploration of the extent to which conditional bias is likely to be present within a given sample (using meta-analysis techniques). For further details, we refer the reader to Marschner.[87]

- *Bias-variance trade-off*: Typically there will be a trade-off between the bias and variance of different estimators. Depending on the context and aims of the trial, different relative importance may be given to the two. For example, in a phase II trial where a precise estimate of the treatment effect is needed to inform a follow-up confirmatory study, the variance of an estimator may be of greater concern, whereas in a definitive phase III trial an unbiased estimate of treatment effect is key for real-world decision-making, as discussed in Section 2.1 One proposal given in the literature is to use the mean squared error as a way of encompassing both bias and variance.

Potentially different criteria will be needed for different outcomes, for example, when considering primary and secondary outcomes, which may then lead to using different estimators for different outcomes. As well, in some trial settings such as multi-arm trials (and drop-the-loser designs) where more than one arm reaches the final stage, the bias of each arm could be considered separately, but there may also be interest in calculating for example, the average bias at the across all arms that are selected. In any case, once criteria for assessing estimators have been decided, the next step is to find potential estimators that can be used for the trial design in question. Part I of this article series is a starting point to find relevant methodological literature and code for implementation.

For more commonly used adaptive designs, a review of the literature may be sufficient to compare the bias and variance of different estimators. Otherwise, we would recommend conducting simulations to explore the bias and variance of potential estimators given the adaptive trial design. In either case, we recommend assessing the estimators across a range of plausible parameter values and design scenarios, taking into account important factors such as the probability of early stopping or reaching the final stage of a trial. More generally, any simulations should follow the relevant FDA guidelines regarding simulation reports for adaptive designs[42(pp28-29)], see also guidance by Mayer et al.[107]

The simulation-based approach can also be used when there are no proposed alternatives to the standard MLE for the trial design under consideration. Even in this setting, we would still encourage an exploration of the bias properties of the MLE. If there is a potential bias of a non-negligible magnitude, then this can impact how the results of the trial are reported (see Section 5.4).

5.2 | Pre-specification of analyses

The statistical analysis plan (SAP) and health economic analysis plan (HEAP) should include a description of the estimators that are planned to be used to estimate treatment effects of interest when reporting the results of the trial, and a justification of the choice of estimators based on the investigations conducted during the planning stage. This reflects the FDA guidance,[42(p28)] which states that there should be "prespecification of the statistical methods that will be used to […] estimate treatment effects … " and "evaluation and discussion of the design … which should typically include […] bias of treatment effect estimates". The trial statistician and health economist should work together to develop plans that are complementary to both their analyses.

When available, unbiased or bias-reduced estimators should be used and (in line with the ACE guidance[3,4]) reported alongside the standard MLE. In settings where multiple adjusted estimators are available and are of interest, one adjusted estimator should be designated the “primary” adjusted estimator for the final reporting of results, with the others included as sensitivity or supplementary analyses (depending on the estimand of interest[108,109]). This is to aid clarity in the interpretation of the trial results, and to avoid “cherry-picking” the estimator that gives the most favorable treatment effect estimate. Similarly, when only one adjusted estimate is reported alongside the standard MLE, it should be made clear which one is the ‘primary’ result. More generally, guidelines for adaptive designs should have a clear requirement to consider bias and bias-adjustment when analyzing trial results.

As an example of what this looks like in practice, we point the reader to the TAO (Treatment of Acute Coronary Syndromes with Otamixaban) trial as described by Steg et al,[110] particularly their appendix B, section 9. The authors consider the MLE and a median-unbiased estimator (MUE), and explore the bias and MSE via simulations and conclude that the MUE has a “consistently smaller” bias with no “noticeable difference in terms of MSE”. Therefore, they propose to use the MUE as the point estimator in their trial.

We have deliberately avoided making recommendations on the most appropriate adjustment method because the most appropriate choice of estimator depends heavily on the context and goals of the trial, as well as the type of adaptive design (and trial adaptations) in question. In addition, given that estimation for adaptive designs is an ongoing research area, there is a risk that any recommendations may become outdated. However, for some adaptive designs, such as the group sequential design presented in our case study in Section 4, it is possible to provide stronger guidance (see the discussion in Section 4.1 as an example).

5.3 | Data monitoring committees (DMCs)

When presenting interim results to DMCs, the issue of potential bias should also be considered. We would recommend that the sensitivity of the standard MLE (based on the interim data) to potential bias should be reported, for example based on simulations conducted during the planning stage. As recommended by Zhang et al[102] and Shimura et al,[92] when unbiased or bias-reduced estimators are available, these should also be presented to the DMC, as an additional tool for appropriately considering potential bias in the decision-making process of whether to stop a trial early (or to perform other trial adaptations such as modifying the sample size).

5.4 | Reporting results for a completed trial

When reporting results following an adaptive design, there should be a clear description of the “statistical methods used to estimate measures of treatment effects”.[3(p16)] Hence, when unbiased or bias-adjusted estimators are used, this should be made clear, along with any underlying assumptions made to calculate them (eg, being unbiased conditional on the observed stopping time). As reflected in the ACE guidance,[3(p16)] “when conventional or naive estimators derived from fixed design methods are used, it should be clearly stated” as well.

The FDA guidance on adaptive designs[42(p30)] states that “treatment effect estimates should adequately take the design into account”. Hence, we reiterate that adjusted estimates taking the trial design into account are to be preferred, if available. The FDA guidance goes on to say that “if naive estimates such as unadjusted sample means are used, the extent of bias should be evaluated, and estimates should be presented with appropriate cautions regarding their interpretation”.[42(p30)] Similarly, the ACE guidelines encourage researchers to discuss “Potential bias and imprecision of the treatment effects if naive estimation methods were used”.

These discussions would naturally link back to the planning stage literature review and/or simulations (which could potentially be updated in light of the trial results and any unplanned adaptations that took place), taking into account important factors such as the probability of early stopping and plausible values of the unknown true treatment effect. For example, if the potential bias of the MLE is likely to be negligible, this would be a reassuring statement to make. On the other hand, in a setting where no adjusted estimators currently exist in the literature (eg, for trials which combine multiple trial adaptations together) and there is the potential for non-negligible bias in the MLE, a statement flagging up this potential concern would allow appropriate caution to be taken when using the point estimate to inform clinical or policy decisions, future studies or meta-analyses.

As discussed in Section 5.4, it should be specified in advance (ie, in the SAP for a confirmatory study) which estimator will be used for the primary analysis and which (if any) estimator(s) will be used as a sensitivity analysis. If an unbiased or bias-adjusted estimator is reported, then it is useful to look at the similarity with the MLE as part of the sensitivity analysis. If the two estimates are very close to each other, then it is reassuring that the trial results appear to be somewhat robust to the estimation strategy used. Conversely, if the two estimates are substantially different, then this may indicate that more care is needed in interpreting the trial results and when using the point estimates for decision-making or further research. However, we caution that the observed difference between the MLE and an unbiased (or bias-adjusted) estimate is not necessarily a precise measure of the actual bias in the observed MLE. Firstly, the bias of the MLE depends on the true underlying treatment effect, which is unknown. Secondly, an unbiased estimator is only unbiased on average, and not necessarily in any particular trial realization. To this end, a potentially more transparent way of reporting results is to show the plausible extent of bias of the MLE across a practically reasonable range of the true treatment effect, in addition to considering the corresponding probabilities of stopping (or reaching the final stage). Again this would build on the planning stage review and simulations.

Finally, the reporting of appropriate measures of uncertainty for the estimators, such as confidence or credible intervals, is also important. If methods exist for constructing confidence intervals associated with the adjusted estimator, then clearly these can be reported. However, for many adjusted estimators it is not clear how to construct valid confidence intervals, and hence the "standard" confidence interval for the MLE may be the only one available.

6 | DISCUSSION

In this article, we have critically assessed how bias can affect standard estimators of treatment effects in adaptive designs and the negative effects this can have. As discussed in part I of this article series,[45] there is a growing body of methodological literature proposing unbiased and bias-adjusted estimators for a wide variety of adaptive trial designs. However, as shown in this article, there has been little uptake of adjusted estimators in practice, with the vast majority of trials continuing to only report the MLE (if indeed it is made clear which estimation method is being used at all). There are a variety of reasons why this may be the case. Firstly, there is a common belief that the bias of the MLE will typically be negligible in realistic trial scenarios. This assumption is sometimes made without supporting evidence such as simulation studies for a variety of trial contexts. In theory, the bias can be very large in certain scenarios[111]. However, as discussed throughout this article, the magnitude of the bias depends on the type of design and whether we are interested in conditional or unconditional bias. Hence, this issue needs to be carefully considered for the specific trial in question. Secondly, there is perhaps a lack of awareness of the range of different unbiased and bias-adjusted estimators that exist in the methodological literature. Linked with this, statistical software and code to easily calculate adjusted estimators is relatively sparse (see also Grayling and Wheeler[112]), which is an obstacle to the uptake of methods in practice even if they exist. It also remains the case that for more complex or novel adaptive designs, adjusted estimators may not exist (see part I of this article series, particularly Section 6).

It is our hope that this article series will encourage the increased use and reporting of adjusted estimators in practice for trial settings where these are available. As described in our guidance section, estimation issues should be considered in the design stage of an adaptive trial. Bowden and Wason[113] give a good example of how this can be done in a principled way for two-stage trials with a binary endpoint. More generally, the estimation strategy should take the design of the trial into account, which motivates the use of adjusted estimators. In terms of trial reporting, statements about the potential bias of the reported estimates can indicate where more care is needed in interpretation of the results and the use of the point estimates for further research including evidence synthesis and health economic analyses.

Finally, to improve the uptake of unbiased and bias-adjusted estimators in practice, there is the need for the further development of user-friendly software and code to allow straightforward calculation of trial results and to aid in simulations. Ideally, the calculation of adjusted estimators could be added to existing widely-used software for adaptive trial design and analysis. Otherwise, there is scope for stand-alone software packages or code (such as that provided for our case study) focusing on estimation after adaptive designs, particularly with simulation studies in mind.

ACKNOWLEDGEMENTS

This research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). This report is independent research supported by the National Institute for Health Research (Prof Jaki's Senior Research Fellowship, NIHR-SRF-2015-08-001). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHCS). T Jaki and DS Robertson received funding from the UK Medical Research Council (MC_UU_00002/14). DS Robertson also received funding from the Biometrika Trust. B Choodari-Oskooei was supported by the MRC grant (MC_UU_00004_09 and MC_UU_12023_29). The Centre for Trials Research receives infrastructure funding from Health and Care Research Wales and Cancer Research UK.

DATA AVAILABILITY STATEMENT

All of the data that support the findings of this study are available within the paper and Supplementary Information. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

ORCID

David S. Robertson https://orcid.org/0000-0001-6207-0416
Babak Choodari-Oskooei https://orcid.org/0000-0001-7679-5899
Munya Dimairo https://orcid.org/0000-0002-9311-6920
Philip Pallmann https://orcid.org/0000-0001-8274-9696
Thomas Jaki https://orcid.org/0000-0002-1096-188X

REFERENCES

1. Friedman F, Furberg C, DeMets DL. *Fundamentals of Clinical Trials*. Vol 4. New York, NY: Springer; 2010.
2. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med*. 2018;16(1):29. doi:10.1186/s12916-018-1017-7
3. Dimairo M, Pallmann P, Wason J, et al. The adaptive designs CONSORT extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. *BMJ*. 2020;369:m115. doi:10.1136/bmj.m115
4. Dimairo M, Pallmann P, Wason J, et al. The adaptive designs CONSORT extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. *Trials*. 2020;21(1):528. doi:10.1186/s13063-020-04334-x
5. Burnett T, Mozgunov P, Pallmann P, Villar SS, Wheeler GM, Jaki T. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC Med*. 2020;18(1):352. doi:10.1186/s12916-020-01808-2
6. Hatfield I, Allison A, Flight L, Julious SA, Dimairo M. Adaptive designs undertaken in clinical research: a review of registered clinical trials. *Trials*. 2016;17(1):150. doi:10.1186/s13063-016-1273-9
7. Bhatt DL, Mehta C. Adaptive designs for clinical trials. *N Engl J Med*. 2016;375(1):65-74. doi:10.1056/NEJMra1510061
8. Bothwell LE, Avorn J, Khan NF, Kesselheim AS. Adaptive design clinical trials: a review of the literature and ClinicalTrials.Gov. *BMJ Open*. 2018;8(2):e018320. doi:10.1136/bmjopen-2017-018320
9. Stallard N, Hampson L, Benda N, et al. Efficient adaptive designs for clinical trials of interventions for COVID-19. *Stat Biopharm Res*. 2020;12(4):483-497. doi:10.1080/19466315.2020.1790415
10. Kunz CU, Jörgens S, Bretz F, et al. Clinical trials impacted by the COVID-19 pandemic: adaptive designs to the rescue? *Stat Biopharm Res*. 2020;12(4):461-477. doi:10.1080/19466315.2020.1799857
11. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Baton Rouge, FL: Chapman & Hall/CRC; 2000.
12. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549. doi:10.2307/2530245
13. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol*. 1971;44(526):793-797. doi:10.1259/0007-1285-44-526-793
14. Wason JMS, Mander AP, Thompson SG. Optimal multistage designs for randomised clinical trials with continuous outcomes. *Stat Med*. 2012;31(4):301-312. doi:10.1002/sim.4421
15. Wason JMS. OptGS: an R package for finding near-optimal group-sequential designs. *J Stat Softw*. 2015;66(2):1-13. doi:10.18637/jss.v066.i02
16. Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *Clin Investig*. 2015;5(4):393-399. doi:10.4155/cli.15.13
17. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med*. 2003;22(5):689-703. doi:10.1002/sim.1362
18. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*. 2012;99(2):494-501. doi:10.1093/biomet/ass002
19. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999;18(14):1833-1848.
20. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biom J*. 2006;48(4):623-634. doi:10.1002/bimj.200510232
21. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biom J*. 2006;48(4):635-643. doi:10.1002/bimj.200510231
22. Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*. 1989;45(2):537-547.
23. Sampson AR, Sill MW. Drop-the-losers design: normal case. *Biom J*. 2005;47(3):257-268, discussion 269-281. doi:10.1002/bimj.200410119

24. Sill MW, Sampson AR. Drop-the-losers design: binomial case. *Comput Stat Data Anal*. 2009;53(3):586-595. doi:10.1016/j.csda.2008.07.031
25. Thall PF. Adaptive enrichment designs in clinical trials. *Annu Rev Stat Appl*. 2021;8(1):393-411. doi:10.1146/annurev-statistics-040720-032818
26. Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med*. 2009;28(10):1445-1463. doi:10.1002/sim.3559
27. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Stat Methods Med Res*. 2019;28(7):2096-2111. doi:10.1177/0962280217747312
28. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Stat Med*. 2013;32(16):2695-2714. doi:10.1002/sim.5738
29. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics*. 2013;14(4):613-625. doi:10.1093/biostatistics/kxt010
30. Götte H, Donica M, Mordenti G. Improving probabilities of correct interim decision in population enrichment designs. *J Biopharm Stat*. 2015;25(5):1020-1038. doi:10.1080/10543406.2014.929583
31. Robertson DS, Lee KM, Lopez-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. *Stat Sci*. 2020. Accessed, 2021. http://arxiv.org/abs/2005.00564.
32. Villar SS, Robertson DS, Rosenberger WF. The temptation of overgeneralizing response-adaptive randomization. *Clin Infect Dis*. 2020;73:e842. doi:10.1093/cid/ciaa1027
33. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice, 2nd Edition*. Hoboken, NJ: John Wiley & Sons, Inc; 2016. doi:10.1002/9781118742112
34. Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L. I-SPY 2: an adaptive breast cancer trial Design in the Setting of Neoadjuvant chemotherapy. *Clin Pharmacol Ther*. 2009;86(1):97-100. doi:10.1038/clpt.2009.68
35. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov*. 2011;1(1):44-53. doi:10.1158/2159-8274.CD-10-0010
36. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biom J Biom Z*. 2006;48(4):537-555. doi:10.1002/bimj.200510238
37. Chuang-Stein C, Anderson K, Gallo P, Collins S. Sample size Reestimation: a review and recommendations. *Drug Inf J*. 2006;40(4):475-484. doi:10.1177/216847900604000413
38. Proschan MA. Sample size re-estimation in clinical trials. *Biom J Biom Z*. 2009;51(2):348-357. doi:10.1002/bimj.200800266
39. Pritchett YL, Menon S, Marchenko O, et al. Sample size Re-estimation designs In confirmatory clinical trials—current state, statistical considerations, and practical guidance. *Stat Biopharm Res*. 2015;7(4):309-321. doi:10.1080/19466315.2015.1098564
40. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001;57(3):886-891. doi:10.1111/j.0006-341x.2001.00886.x
41. Chow SC, Chang M, Pong A. Statistical consideration of adaptive methods in clinical development. *J Biopharm Stat*. 2005;15(4):575-591. doi:10.1081/BIP-200062277
42. U.S. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. 2019, https://www.fda.gov/media/78495/download
43. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Berlin Heidelberg: Springer; 2016.
44. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika*. 2013;100(4):985-996. doi:10.1093/biomet/ast035
45. Robertson DS, Choodari-Oskooei B, Dimairo M, Flight L, Pallmann P, Jaki T. Point estimation for adaptive trial designs I: a methodological review. *Stat Med*. 2023;42(2):122-145. doi:10.1002/sim.9605
46. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias—two hostile brothers. *Stat Med*. 2009;29:1-13. doi:10.1002/sim.3716
47. Stevely A, Dimairo M, Todd S, et al. An investigation of the shortcomings of the CONSORT 2010 statement for the reporting of group sequential randomised controlled trials: a methodological systematic review. *PloS One*. 2015;10(11):e0141104. doi:10.1371/journal.pone.0141104
48. Briel M, Bassler D, Wang AT, Guyatt GH, Montori VM. The dangers of stopping a trial too early. *J Bone Joint Surg Am*. 2012;94(Suppl 1):56-60. doi:10.2106/JBJS.K.01412
49. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-1187. doi:10.1001/jama.2010.310
50. Bassler D, Montori VM, Briel M, et al. Reflections on meta-analyses involving trials stopped early for benefit: is there a problem and if so, what is it? *Stat Methods Med Res*. 2013;22(2):159-168. doi:10.1177/0962280211432211
51. Walter SD, Han H, Briel M, Guyatt GH. Quantifying the bias in the estimated treatment effect in randomized trials having interim analyses and a rule for early stopping for futility. *Stat Med*. 2017;36(9):1506-1518. doi:10.1002/sim.7242
52. Walter SD, Guyatt GH, Bassler D, Briel M, Ramsay T, Han HD. Randomised trials with provision for early stopping for benefit (or harm): the impact on the estimated treatment effect. *Stat Med*. 2019;38(14):2524-2543. doi:10.1002/sim.8142
53. Wang H, Rosner GL, Goodman SN. Quantifying over-estimation in early stopped clinical trials and the "freezing effect" on subsequent research. *Clin Trials*. 2016;13(6):621-631. doi:10.1177/1740774516649595
54. Goodman SN. Stopping trials for efficacy: an almost unbiased view. *Clin Trials*. 2009;6(2):133-135. doi:10.1177/1740774509103609
55. Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early. *BMJ*. 2012;344:e3863. doi:10.1136/bmj.e3863

56. Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH. Ethical issues in stopping randomized trials early because of apparent benefit. *Ann Intern Med*. 2007;146(12):878-881. doi:10.7326/0003-4819-146-12-200706190-00009
57. Whitehead J. Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics*. 1986;42(3):461-471.
58. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Rev. 2nd ed: Wiley & Sons; 1997.
59. Hughes D, Waddingham E, Mt-Isa S, et al. Recommendations for benefit-risk assessment methodologies and visual representations. *Pharmacoepidemiol Drug Saf*. 2016;25(3):251-262. doi:10.1002/pds.3958
60. Juhaeri J. Benefit-risk evaluation: the past, present and future. *Ther Adv Drug Saf*. 2019;10:1-10. doi:10.1177/2042098619871180
61. Fish WH, Cohen M, Franzek D, Williams JM, Lemons JA. Effect of intramuscular vitamin E on mortality and intracranial hemorrhage in neonates of 1000 grams or less. *Pediatrics*. 1990;85(4):578-584.
62. Brion LP, Bell EF, Raghuveer TS. Vitamin E supplementation for prevention of morbidity and mortality in preterm infants. *Cochrane Database Syst Rev*. 2003;2010:1-258. doi:10.1002/14651858.CD003665
63. Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Chichester, West Sussex: John Wiley & Sons; 2002.
64. Gough D, Davies P, Jamtvedt G, et al. Evidence synthesis international (ESI): position statement. *Syst Rev*. 2020;9(1):155. doi:10.1186/s13643-020-01415-5
65. Montori VM, Devereaux PJ, Adhikari NKJ, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294(17):2203. doi:10.1001/jama.294.17.2203
66. Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics*. 1992;48(1):41-53.
67. Bassler D, Ferreira-Gonzalez I, Briel M, et al. Systematic reviewers neglect bias that results from trials stopped early for benefit. *J Clin Epidemiol*. 2007;60(9):869-873. doi:10.1016/j.jclinepi.2006.12.006
68. Cameron C, Ewara E, Wilson FR, et al. The importance of considering differences in study Design in Network Meta-analysis: an application using anti-tumor necrosis factor drugs for ulcerative colitis. *Med Decis Making*. 2017;37(8):894-904. doi:10.1177/0272989X17711933
69. Todd S. Incorporation of sequential trials into a fixed effects meta-analysis. *Stat Med*. 1997;16(24):2915-2925. doi:10.1002/(sici)1097-0258(19971230)16:24<2915::aid-sim688>3.0.co;2-0
70. Goodman SN. Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Ann Intern Med*. 2007;146(12):882-887. doi:10.7326/0003-4819-146-12-200706190-00010
71. Goodman SN. Systematic reviews are not biased by results from trials stopped early for benefit. *J Clin Epidemiol*. 2008;61(1):95-96. doi:10.1016/j.jclinepi.2007.06.012
72. Goodman S, Berry D, Wittes J. Bias and trials stopped early for benefit. *JAMA*. 2010;304(2):157-159. doi:10.1001/jama.2010.931
73. Schou IM, Marschner IC. Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Stat Med*. 2013;32(28):4859-4874. doi:10.1002/sim.5893
74. Marschner IC, Askie LM, Schou IM. Sensitivity analyses assessing the impact of early stopping on systematic reviews: recommendations for interpreting guidelines. *Res Synth Methods*. 2020;11(2):287-300. doi:10.1002/jrsm.1394
75. Luchini C, Veronese N, Nottegar A, et al. Assessing the quality of studies in meta-research: review/guidelines on the most important quality assessment tools. *Pharm Stat*. 2021;20(1):185-195. doi:10.1002/pst.2068
76. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407-415. doi:10.1016/j.jclinepi.2010.07.017
77. Mitchell JM, Patterson JA. The inclusion of economic endpoints as outcomes in clinical trials reported to ClinicalTrials.Gov. *J Manag Care Spec Pharm*. 2020;26(4):386-393. doi:10.18553/jmcp.2020.26.4.386
78. Drummond M. *Methods for the Economic Evaluation of Health Care Programmes*. 4th ed. Oxford, UK: Oxford University Press; 2015.
79. Marschner IC, Schou IM. Underestimation of treatment effects in sequentially monitored clinical trials that did not stop early for benefit. *Stat Methods Med Res*. 2019;28(10-11):3027-3041. doi:10.1177/0962280218795320
80. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med*. 1993;329(10):673-682. doi:10.1056/NEJM199309023291001
81. Mark DB, Hlatky MA, Califf RM, et al. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med*. 1995;332(21):1418-1424. doi:10.1056/NEJM199505253322106
82. Flight L. *The Use of Health Economics in the Design and Analysis of Adaptive Clinical Trials*, PhD Thesis. University of Sheffield; 2020. http://etheses.whiterose.ac.uk/27924/
83. Flight L, Arshad F, Barnsley R, et al. A review of clinical trials with an adaptive design and health economic analysis. *Value Health J Int Soc Pharmacoeconom Outcomes Res*. 2019;22(4):391-398. doi:10.1016/j.jval.2018.11.008
84. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*. 1986;73(3):573-581. doi:10.1093/biomet/73.3.573
85. Bowden J, Trippa L. Unbiased estimation for response adaptive clinical trials. *Stat Methods Med Res*. 2017;26(5):2376-2388. doi:10.1177/0962280215597716
86. Flournoy N, Oron AP. Bias induced by adaptive dose-finding designs. *J Appl Stat*. 2020;47(13-15):2431-2442. doi:10.1080/02664763.2019.1649375
87. Marschner IC. A general framework for the analysis of adaptive experiments. *Stat Sci*. 2021;36(3):465-492. doi:10.1214/20-STS803
88. Emerson SS, Banks PLC. Interpretation of a leukemia trial stopped early. *Case Studies in Biometry*. New York, NY: Wiley; 1994.
89. Choodari-Oskooei B, Parmar MK, Royston P, Bowden J. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. *Trials*. 2013;14(1):23. doi:10.1186/1745-6215-14-23

90. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Stat Med*. 2013;32(10):1677-1690. doi:10.1002/sim.5463
91. Brückner M, Titman A, Jaki T. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. *Stat Med*. 2017;36(20):3137-3153. doi:10.1002/sim.7367
92. Shimura M, Nomura S, Wakabayashi M, Maruo K, Gosho M. Assessment of Hazard ratios in oncology clinical trials terminated early for superiority: a systematic review. *JAMA Netw Open*. 2020;3(6):e208633. doi:10.1001/jamanetworkopen.2020.8633
93. Wittes J. Stopping a trial early–and then what? *Clin Trials*. 2012;9(6):714-720. doi:10.1177/1740774512454600
94. Fernandes RM, van der Lee JH, Offringa M. A systematic review of the reporting of data monitoring Committees' roles, interim analysis and early termination in pediatric clinical trials. *BMC Pediatr*. 2009;9:77. doi:10.1186/1471-2431-9-77
95. Kaufmann P, Thompson JLP, Levy G, et al. Phase II trial of CoQ10 for ALS finds insufficient evidence to justify phase III. *Ann Neurol*. 2009;66(2):235-244. doi:10.1002/ana.21743
96. Barker KL, Newman M, Stallard N, et al. Physiotherapy rehabilitation for osteoporotic vertebral fracture—a randomised controlled trial and economic evaluation (PROVE trial). *Osteoporos Int*. 2020;31(2):277-289. doi:10.1007/s00198-019-05133-0
97. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med*. 2016;35(3):325-347. doi:10.1002/sim.6472
98. Zajicek JP, Hobart JC, Slade A, Barnes D, Mattison PG, MUSEC Research Group. Multiple sclerosis and extract of cannabis: results of the MUSEC trial. *J Neurol Neurosurg Psychiatry*. 2012;83(11):1125-1132. doi:10.1136/jnnp-2012-302468
99. Emerson SS. *Parameter Estimation Following Group Sequential Hypothesis Testing*, PhD Thesis. *University of Washington*; 1988.
100. Troendle JF, Yu KF. Conditional estimation following a group sequential clinical trial. *Commun Stat Theory Methods*. 1999;28(7):1617-1634. doi:10.1080/03610929908832376
101. Fan XF, DeMets DL, Lan KKG. Conditional bias of point estimates following a group sequential test. *J Biopharm Stat*. 2004;14(2):505-530. doi:10.1081/bip-120037195
102. Zhang JJ, Blumenthal GM, He K, Tang S, Cortazar P, Sridhara R. Overestimation of the effect size in group sequential trials. *Clin Cancer Res*. 2012;18(18):4872-4876. doi:10.1158/1078-0432.CCR-11-3118
103. Schönbrodt FD, Wagenmakers EJ. Bayes factor design analysis: planning for compelling evidence. *Psychon Bull Rev*. 2018;25(1):128-142. doi:10.3758/s13423-017-1230-y
104. Shimura M, Gosho M, Hirakawa A. Comparison of conditional bias-adjusted estimators for interim analysis in clinical trials with survival data: comparison of conditional bias-adjusted estimators for survival data. *Stat Med*. 2017;36(13):2067-2080. doi:10.1002/sim.7258
105. Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Stat Med*. 2012;31(5):420-435. doi:10.1002/sim.4430
106. Grayling MJ, Wason JM. Point estimation following a two-stage group sequential trial. *Stat Methods Med Res*. 2022;32:287-304. doi:10.1177/09622802221137745
107. Mayer C, Perevozskaya I, Leonov S, et al. Simulation practices for adaptive trial designs in drug and device development. *Stat Biopharm Res*. 2019;11(4):325-335. doi:10.1080/19466315.2018.1560359
108. Collignon O, Schiel A, Burman C, Rufibach K, Posch M, Bretz F. Estimands and complex innovative designs. *Clin Pharmacol Ther*. 2022;112:1183-1190. doi:10.1002/cpt.2575
109. Okwuokenye M, Peace KE. Adaptive design and the Estimand framework. *Ann Biostat Biom Appl*. 2019;1(5):1-4. doi:10.33552/ABBA.2019.01.000524
110. Steg PG, Mehta SR, Pollack CV, et al. Design and rationale of the treatment of acute coronary syndromes with otamixaban trial: a double-blind triple-dummy 2-stage randomized trial comparing otamixaban to unfractionated heparin and eptifibatide in non-ST-segment elevation acute coronary syndromes with a planned early invasive strategy. *Am Heart J*. 2012;164(6):817-824.e13. doi:10.1016/j.ahj.2012.10.001
111. Graf AC, Gutjahr G, Brannath W. Precision of maximum likelihood estimation in adaptive designs. *Stat Med*. 2016;35(6):922-941. doi:10.1002/sim.6761
112. Grayling MJ, Wheeler GM. A review of available software for adaptive clinical trial design. *Clin Trials*. 2020;17(3):323-331. doi:10.1177/1740774520906398
113. Bowden J, Wason J. Identifying combined design and analysis procedures in two-stage trials with a binary end point. *Stat Med*. 2012;31(29):3874-3884. doi:10.1002/sim.5468

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Robertson DS, Choodari-Oskooei B, Dimairo M, Flight L, Pallmann P, Jaki T. Point estimation for adaptive trial designs II: Practical considerations and guidance. *Statistics in Medicine*. 2023;1-25. doi: 10.1002/sim.9734

APPENDIX A

A.1 Guidance on bias-adjusted analyses for adaptive designs

A.1.1 FDA: Adaptive designs for clinical trials of drugs and biologics[42]

"Adaptive designs require specific analytical methods to avoid increasing the chance of erroneous conclusions and introducing bias in estimates. For complex adaptive designs, such methods may not be readily available, and simulations are often critical"—page 6.

"It is important that clinical trials produce sufficiently reliable treatment effect estimates to facilitate an evaluation of benefit-risk and to appropriately label new drugs, enabling the practice of evidence-based medicine. Some adaptive design features can lead to statistical bias in the estimation of treatment effects and related quantities. For example, each of the two cases of Type I error probability inflation mentioned in section III.A. above has a potential for biased estimates. Specifically, a conventional end-of-trial treatment effect estimate such as a sample mean that does not take the adaptations into account would tend to overestimate the true population treatment effect. This is true not only for the primary endpoint which formed the basis of the adaptations, but also for secondary endpoints correlated with the primary endpoint. Furthermore, confidence intervals for the primary and secondary endpoints may not have correct coverage probabilities for the true treatment effects.

For some designs there are known methods for adjusting estimates to reduce or remove bias associated with adaptations and to improve performance on measures such as the mean squared error (eg, Jennison and Turnbull 1999; Wassmer and Brannath 2016). Such methods should be prospectively planned and used for reporting results when they are available. Biased estimation in adaptive design is currently a less well-studied phenomenon than Type I error probability inflation, however, and methods may not be available for other designs. For these other designs, the extent of bias in estimates should be evaluated, and treatment effect estimates and associated confidence intervals should be presented with appropriate cautions regarding their interpretation."—page 8.

"Finally, conventional fixed sample estimates of the treatment effect such as the sample mean tend to be biased toward greater effects than the true value when a group sequential design is used. Similarly, confidence intervals do not have the desired nominal coverage probabilities. Therefore, a variety of methods exist to compute estimates and confidence intervals that appropriately adjust for the group sequential stopping rules (Jennison and Turnbull 1999). To ensure the scientific and statistical credibility of trial results and facilitate important benefit-risk considerations, an approach for calculating estimates and confidence intervals that appropriately accounts for the group sequential design should be prospectively planned and used for reporting results."—pages 12-13.

"Consider group sequential designs: It is widely understood that multiple analyses of the primary endpoint can inflate the Type I error probability and lead to biased estimation of treatment effects on that endpoint. Less well appreciated, however, is that Type I error probability inflation and biased estimation can also apply to any endpoint correlated with the primary endpoint (Hung et al. 2007)."—page 22.

[Documentation Prior to Conducting an Adaptive Trial] "Evaluation and discussion of the design operating characteristics, which should typically include Type I error probability; power; expected, minimum, and maximum sample size; bias of treatment effect estimates; and coverage of confidence intervals. Such evaluations might be achieved through analytical calculations and/or computer simulations. If operating characteristics are evaluated analytically, appropriate details (eg, literature references or proofs) for the methodology should be submitted."—page 28.

"Appropriate reporting of the adaptive design and trial results … For example, the trial summary should describe the adaptive design utilized. In addition, treatment effect estimates should adequately take the design into account, or if naive estimates such as unadjusted sample means are used, the extent of bias should be evaluated, and estimates should be presented with appropriate cautions regarding their interpretation."—page 30

A.1.2 The adaptive designs CONSORT extension (ACE) statement[3,4]

"A goal of every trial is to provide reliable estimates of the treatment effect for assessing benefits and risks to reach correct conclusions. Several statistical issues may arise when using an AD depending on its type and the scope of adaptations, the adaptive decision-making criteria and whether frequentist or Bayesian methods are used to design and analyses the trial. Conventional estimates of treatment effect based on fixed design methods may be unreliable when applied to ADs (eg, may exaggerate the patient benefit). Precision around the estimated treatment effects may be incorrect (eg, the width of confidence intervals may be incorrect). Other methods available to summarize the level of evidence in hypothesis testing (eg, *P*-values) may give different answers. Some factors and conditions that influence the magnitude of estimation bias have been investigated and there are circumstances when it may not be of concern. Secondary analyses (eg, health economic evaluation) may also be affected if appropriate adjustments are not made. Cameron et al discuss methodological challenges in performing network meta-analysis when combining evidence from randomized trials with ADs and fixed designs. Statistical methods for estimating the treatment effect and its precision exist for some ADs and implementation tools are being developed. However, these methods are rarely used or reported and the implications are unclear. Debate and research on inference for some ADs with complex adaptations is ongoing. In addition to statistical methods for comparing outcomes between groups (item 12a), we specifically encourage authors to clearly describe statistical methods used to estimate measures of treatment effects with associated uncertainty (eg, confidence or credible intervals) and *P*-value (when appropriate); referencing relevant literature is sufficient. When conventional or naïve estimators derived from fixed design methods are used, it should be clearly stated. In situations where statistical simulations were used to either explore the extent of bias in estimation of the treatment effects … or operating characteristics, it is good practice to mention this and provide supporting evidence (item 24c)."—page 16.

"For AD randomized trials, further discussion should include the implications of: …

- Potential bias and imprecision of the treatment effects if naïve estimation methods were used;" —page 21.

"For some AD randomized trials, methods to derive statistical properties analytically may not be available. Thus, it becomes necessary to perform simulations under a wide range of plausible scenarios to investigate the operating characteristics of the design (item 7a), impact on estimation bias (item 12b), and appropriateness and consequences of decision-making criteria and rules. In such cases, we encourage authors to reference accessible material used for this purpose (eg, simulation protocol and report, or published related material). Furthermore, it is good scientific practice to reference software, programs or code used for this task to facilitate reproducible research."—page 24.

A.2 Case study: Group sequential design

A.2.1 Definition of the information at stages 1 and 2

At stage k ($k = 1,2$), let $\widetilde{p_k}$ denote the pooled estimate of the mean overall success probability, that is, the total number of observed successes divided by the total number of subjects. Then the observed information I_k is given by

$$I_k = \frac{1}{\widetilde{p_k}\left(1-\widetilde{p_k}\right)\left(1/n_{0k}+1/n_{CEk}\right)},$$

where n_{0k} and n_{CEk} are the number of subjects on the placebo and CE arms, respectively, at stage k.

A.2.2 Definition of the conditional density of $\hat{\theta}$

The conditional density of $\hat{\theta}$, conditional on continuing to stage 2, is given by the following expression:

$$f(\hat{\theta}|T=2) = \frac{1-\Phi\left(\frac{c/\sqrt{I_1}-\hat{\theta}}{1/I_1-1/I_2}\right)}{1-\Phi\left(c-\theta\sqrt{I_1}\right)}\frac{\exp\left[-\frac{I_2}{2}(\hat{\theta}-\theta)^2\right]}{\sqrt{2\pi/I_2}},$$

where c is the stopping boundary at stage 1, that is, the trial stops at stage 1 if $Z_1 \geq c$.

A.2.3 Conditional vs unconditional perspectives

Below we give some quotations from the literature focusing on the issue of the conditional vs unconditional perspective in the context of group sequential designs.

Troendle and Yu[100](pp1617-1618):

"Suppose a group sequential clinical trial is undertaken to determine the effect of an experimental drug on the state of a certain disease. Now suppose it is known that the trial was stopped at the first interim analysis because of treatment efficacy, and that the estimated treatment effect was X_1-Y_1, the difference in sample means from the two groups … Is X_1-Y_1 a reasonable estimate of the effect size? Although X_1-Y_1 is unbiased, the general estimator X_T-Y_T, where T is the random stopping time, is known to be biased. Recently, an unbiased estimator … and an essentially unbiased estimator … have been developed for this problem. However, as will be shown later, these methods remain unbiased by overestimating the effect when there is early stopping while underestimating the effect when the trial stops later. The overall effect is an unbiased estimator, but does that leave the scientist, who knows $T = 1$ any happier? We propose conditioning on the stopping time in a group sequential trial to reduce the discrepancy between the conditional expectation of the estimator and the parameter value."

Fan et al[101](pp506-507):

"We also note that the bias referred to is the marginal or overall bias [ie, the unconditional bias]. As much as the importance of the marginal bias, sometimes we will also face the question of what the potential bias is given the fact that the study is already stopped at this time, especially when it is a very early interim stage. To answer this question, we feel it is more relevant to investigate the bias conditioning on the actual stopping time. In this article, we focus on the angle of the conditional bias and in the meanwhile also keep in mind the marginal bias."

"The conditional method is not meant to replace the unconditional methods because these two methods are developed to address different issues. Instead it is proposed as an addition and alternative means that we can take advantage of when the conditional bias is more concerning, rather than a replacement to the unconditional methods."

Zhang et al[102](p4876):

"Although this article focuses on the bias conditional on the observed stopping time, we also recognize the importance of the marginal or unconditional bias … Evaluation of the unconditional bias is particularly helpful in the trial design stage; however, there is also value in assessing the potential bias given that the trial has already stopped (conditional bias), especially on the basis of a very early interim analysis. Fan and colleagues found that the conditional bias may be quite serious, even in situations in which the unconditional bias is acceptable[101] Most of the available adjustment methods focus on the unconditional bias, which has little effect on the conditional bias.

Schoenbrot and Wagenmakers[103](p140):

"Although sequential designs have negligible unconditional bias, it may nevertheless be desirable to provide a principled 'correction' for the conditional bias at early terminations, in particular when the effect size of a single study is evaluated."

Shimura et al[104](p2068):

"A reduction in conditional bias is as important as a reduction in overall bias because, in practice, researchers can only obtain an estimate that is conditional on the stopping stages."