RESEARCH ARTICLE

Point estimation for adaptive trial designs I: A methodological review

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, 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

Recent FDA guidance on adaptive clinical trial designs defines bias as "a systematic tendency for the estimate of treatment effect to deviate from its true value," and states that it is desirable to obtain and report estimates of treatment effects that reduce or remove this bias. The conventional end-of-trial point estimates of the treatment effects are prone to bias in many adaptive designs, because they do not take into account the potential and realized trial adaptations. While much of the methodological developments on adaptive designs have tended to focus on control of type I error rates and power considerations, in contrast the question of biased estimation has received relatively less attention. This article is the first in a two-part series that studies the issue of potential bias in point estimation for adaptive trials. Part I provides a comprehensive review of the methods to remove or reduce the potential bias in point estimation of treatment effects for adaptive designs, while part II illustrates how to implement these in practice and proposes a set of guidelines for trial statisticians. The methods reviewed in this article can be broadly classified into unbiased and bias-reduced estimation, and we also provide a classification of estimators by the type of adaptive design. We compare the proposed methods, highlight available software and code, and discuss potential methodological gaps in the literature.

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

1 | INTRODUCTION

Adaptive clinical trials allow for preplanned opportunities to alter the course of the trial on the basis of accruing information.[1-3] This may include changes such as increasing the recruitment target (in sample size re-estimation designs), selecting the most promising treatment arms (in multi-arm multi-stage designs) or patient subpopulations (in population enrichment designs), shifting the randomization ratio toward more promising arms (in adaptive randomization designs), or terminating recruitment early for clear evidence of benefit or lack thereof (in group sequential designs).[4] Despite some additional complexities when implementing such trials,[5] they are increasingly being used in practice due

to their attractive features, adding flexibility to a trial design while maintaining scientific rigour.[6-8] The challenges of the COVID-19 pandemic have also recently accelerated their use.[9,10]

To date most of the research undertaken has focused on the design of such studies and the associated question of maintaining desirable operating characteristics related to hypothesis testing (namely, type I error and power) rather than estimation. General methods on the basis of P-value combination[11,12] and conditional error functions[13,14] have been proposed that are applicable to a wide range of adaptive designs as well as specialized methods for specific designs such as multi-arm multi-stage (MAMS) designs,[15,16] adaptive enrichment designs,[17,18] response-adaptive randomization (RAR) designs,[19,20] and sample size re-estimation[21,22] (see Section 1.1 for definitions of the different adaptive designs considered in this article).

In contrast, the question of estimation of treatment effects in an adaptive clinical trial has received comparatively less attention, as reflected in the recent FDA guidance on adaptive designs[23] which states "Biased estimation in adaptive design is currently a less well-studied phenomenon than Type I error probability inflation." This article is the first in a two-part series that studies the issue of potential bias in point estimation for adaptive designs. In the current article (part I), we review methods for unbiased and bias-reduced estimation of treatment effects after an adaptive clinical trial and critically discuss different approaches. In part II, we consider point estimation for adaptive designs from a practical perspective, including a set of guidelines for best practice.[24]

While equally important, the construction of related quantities for inference, such as confidence intervals or regions, is beyond the scope of this series so we signpost the interested reader to related literature.[25,26] Moreover, our focus is on bias in the statistical sense, and we do not consider issues such as operational bias in adaptive designs[2,3,27] or the impact of publication bias[28]; see also recent reference articles[29,30] for discussion of other types of biases in randomized controlled trials.

The structure of this article is as follows. After defining different types of adaptive designs in Section 1.1, we introduce key concepts and the different definitions of bias in Section 2. We describe the search strategy used for a systematic review of the methodological literature in Section 3. We present a summary of the results of the review grouped by the type of adaptive design in Section 4. Sections 5 and 6 then provide detailed descriptions and explanations of the different types of point estimators, divided into methods for *unbiased estimation* (which aims to completely remove bias) in Section 5 and *bias-reduced estimation* (which aims to reduce the magnitude of the bias, but not necessarily to eliminate it) in Section 6. In the Supplementary Information, we provide an annotated bibliography giving further details of all the papers included in the systematic review. The overall conclusions are discussed in Section 7.

1.1 | Glossary of adaptive designs

Table 1 gives definitions of the different types of adaptive designs considered in this article, including the different terminologies used for them. These definitions are based on those provided by Pallmann et al[1] and Burnett et al,[4] who note that there can be ambiguous terminology used in the literature. Adaptive designs can combine multiple adaptive features within a single trial, see Dimairo et al[2] for examples.

2 | ESTIMATION BIAS IN ADAPTIVE DESIGNS

The issue with estimation after an adaptive trial is that traditional maximum likelihood (ML) estimators tend to be biased either because of some selection that took place following an interim analysis (see Bauer et al[31] for a detailed explanation of why selection results in bias) or other mechanisms utilized in an adaptive design, such as early stopping, which might affect the sampling distribution of the estimator (these depend on the nature of the design). For this reason, the usual ML estimator (MLE) is sometimes referred to as the "naive" estimator for the trial.

Before considering the issue of estimation further it is worth clarifying what we mean by a biased estimator. The FDA guidance on adaptive designs[23] defines bias as "a systematic tendency for the estimate of treatment effect to deviate from its true value," and states that "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." It is clear that (all else being equal) it is desirable to obtain estimators of treatment effects that are unbiased in order to make reliable conclusions about the effects of study treatments.

TABLE 1 Definitions of the different types of adaptive designs considered in this article

Design	a.k.a.	Definition
Group sequential	Multi-stage design	Allows for early stopping of the trial for safety, futility, or efficacy
		Group sequential designs can either be two-arm or single-arm trials, with Simon's two-stage design being a common example of the latter. For the generalization to multiple treatment arms, see "Multi-arm multi-stage designs"
		When combined with the option of sample size re-estimation and/or adjustment of the stopping boundaries during the trial, these designs are sometimes referred to as "adaptive group sequential"
Sample size re-estimation	Sample size reassessment/recalculation/adaptations	Adjustment of the sample size, in either a blinded or unblinded fashion. This is typically done to help ensure the trial achieves the desired power
		Sample size re-estimation is sometimes combined with other trial adaptations, such as group sequential testing
Multi-arm multi-stage (MAMS)	Pick-the-winner/drop-the-loser/treatment selection	Compares multiple treatments, doses, durations or combinations to a common control, allowing for early stopping of treatment arms for efficacy or futility. In addition, this can include interim treatment selection of the most promising treatment(s) for further investigation in the subsequent stages
		Variations of MAMS designs include "drop-the-loser" trials, where inferior treatment arms are dropped from the trial (with the control group typically retained). "Seamless" trials (such as phase II/III trials) are also closely related, where the selection and confirmatory phases of development are combined into one trial. "Platform" trials can also be viewed as an extension of MAMS designs, by allowing new treatment arms to be added to the trial over time
Response-adaptive randomization (RAR)	Outcome-adaptive randomization	Shifts allocation ratios toward more promising or informative treatment(s) based on the accumulating patient response data. RAR is sometimes combined with options for early stopping (eg, for futility or efficacy)
Adaptive enrichment	Population enrichment/patient enrichment/adaptive population enrichment	Allows for selection of a target population during the trial, typically using predefined patient subpopulations based on biomarker information. For example, one class of adaptive enrichment designs allows the eligibility criteria to be adaptively updated during the trial based on the estimated treatment-biomarker interactions, restricting entry in the subsequent trial stage(s) to patients most likely to benefit from the new treatment

While it is relatively easy to define statistical bias, different definitions of an unbiased estimator are relevant in our context. To introduce these, let us denote the population parameter of interest, the treatment effect, by θ and an estimator thereof by $\widehat{\theta}$.

2.1 | Mean-unbiased estimators

An estimator $\widehat{\theta}$ is called *mean-unbiased* if its expected value is the same as the true value of the parameter of interest, that is, $E(\widehat{\theta}) = \theta$. This is the most commonly used definition of unbiasedness.

2.2 | Median-unbiased estimators

An estimator $\widehat{\theta}$ is called *median-unbiased* if $P(\widehat{\theta} < \theta) = P(\widehat{\theta} > \theta)$, that is if the probability of overestimation is the same as the probability of underestimation. Note that for symmetric sampling distributions of $\widehat{\theta}$, a median-unbiased estimator is also mean-unbiased.

2.3 | Conditionally and unconditionally unbiased estimators

A further distinction of unbiasedness in estimators refers to whether they are *conditionally* or *unconditionally* unbiased. In our context, 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, one might be interested in an estimator only conditional on a particular arm being selected at an interim analysis; as such, the focus becomes on a conditional unbiased estimator. We discuss the issue of conditional vs unconditional bias further throughout the rest of this article.

2.4 | Bias and mean squared error

When considering estimation after an adaptive trial one is often faced with the following conundrum: precise estimators can potentially be biased while unbiased estimators tend to be less precise, which reflects the classical bias-variance trade-off. This means that in many instances the mean squared error (MSE), a measure of precision defined as $E((\widehat{\theta} - \theta)^2)$, of a biased estimator is often smaller than the MSE of an unbiased estimator. This makes it challenging to find the "best" estimator that fits a particular trial design.

2.5 | The focus of our review

Since bias as defined above is an expectation (or probability) taken over possible realizations of data, it is inherently a frequentist concept. However, we can still evaluate the frequentist bias of a Bayesian point estimator, such as the posterior mean. Similarly, it is possible to use frequentist point estimators in the analysis of Bayesian trial designs (ie, where the adaptations and decision rules are driven by Bayesian analyses). Hence we consider both Bayesian point estimators and Bayesian adaptive designs in this article. However, we restrict our attention to phase II and III trial designs, since phase III trial results have a direct impact on health policy and the adoption of new treatments for wider public use, while phase II trial results directly influence whether further research of a treatment in phase III is required.

3 | METHODOLOGICAL REVIEW: SEARCH STRATEGY AND PAPER SELECTION

We conducted a database search of Scopus on July 13, 2022 of all available papers up to that date. We used a "title, abstract, keywords" search, with the predefined search terms for different categories of estimator given in Table 2.

TABLE 2 Search strategy used for the initial database search of Scopus

Category of estimator	Search terms
Unbiased	((((unbiased OR ((conditionally OR mean OR median OR "uniformly minimum variance" OR "uniformly minimum variance conditionally") AND unbiased)) AND estim*) OR "median adjusted" OR umvcue OR umvue) AND ("adaptive design" OR "adaptive trial" OR "adaptive clinical trial" OR "group-sequential"))
Bias-reduced	(((("bias-reduced" OR "bias-adjusted" OR "conditional moment" OR "empirical Bayes" OR shrinkage) AND estimat*) OR ((corrected OR adjusted OR conditional) AND (mle OR "maximum likelihood"))) AND ("adaptive design" OR "adaptive trial" OR "adaptive clinical trial" OR "group-sequential")
Resampling-based methods	(((bootstrap OR nonparametric OR non-parametric OR resampling OR "distribution free" OR distribution-free OR jackknife OR "Monte Carlo") AND estim*) AND ("adaptive design" OR "adaptive trial" OR "adaptive clinical trial" OR "group-sequential"))
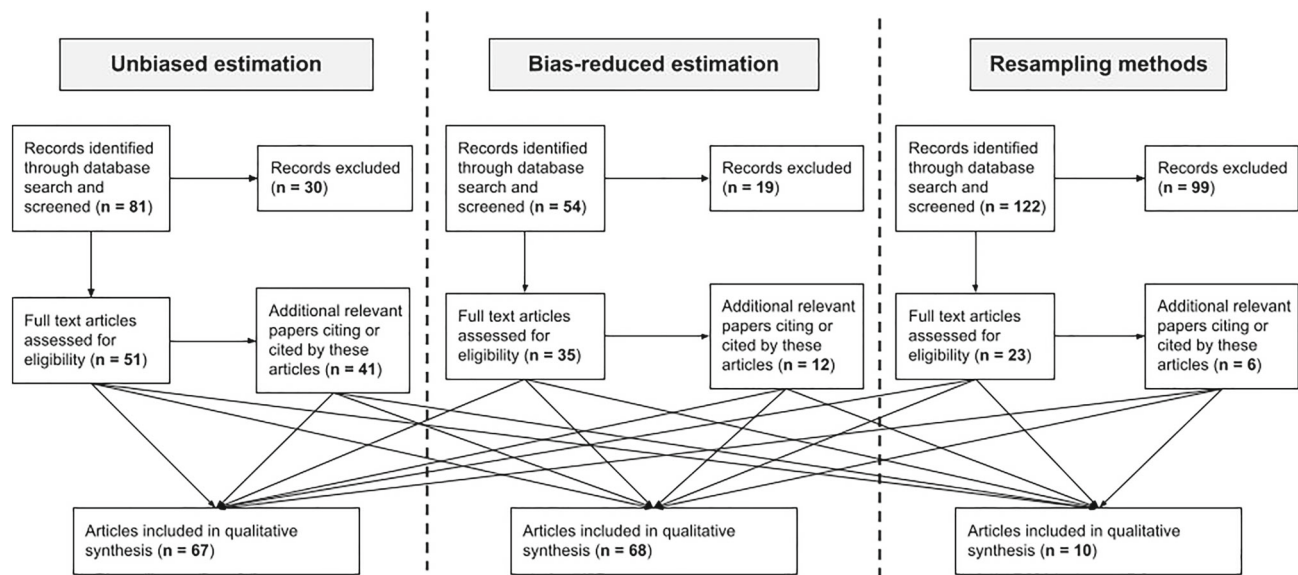
FIGURE 1 Diagram showing the flow of information through the different phases of the systematic review. The Bayesian and empirical Bayes methods are included in the "bias-reduced" category

We excluded papers not involving adaptive clinical trial designs, as well as those that did not propose or evaluate unbiased or bias-reduced estimators. Two authors (DSR and PP) separately reviewed the papers, extracted data and checked each other's work for accuracy.

Our search strategy retrieved a total of 257 papers, of which 148 were excluded immediately as they were not relevant based on the title and abstract. We then looked for additional relevant papers citing or cited by these remaining 109, which added 59 papers. After completing a full text review of these papers, a total of 145 were deemed relevant, and information about the trial contexts, advantages, limitations, code availability and case studies was extracted for qualitative synthesis (Figure 1). Full results giving a summary of each paper extracted for qualitative analysis are found in the supporting materials.

4 | COMPARISONS OF ESTIMATORS, TRIAL EXAMPLES AND SOFTWARE

In this section, we summarize the results of our systematic review in Table 3 by classifying them according to the broad class of adaptive design used. For each design class, we give references to the relevant literature for the different types of estimators that have been proposed. Where the literature focuses on designs that allow for multiple adaptive features within a single trial, we have used our judgment to place them into a class of adaptive designs, and refer the reader to Sections 5 and 6 for further clarifying details. In Table 3, we also give a summary of general pros and cons of the different estimators and the comparisons between them as given in the literature. Finally, we point out examples of the use of different estimators on trial data, as well as where software/code is available. Note that this list of software/code is by no means exhaustive, but only includes those that were specifically mentioned in the reviewed methodological literature or those of which the authors were already aware.

The systematic review identified mean-unbiased estimators, median unbiased estimators, resampling-based estimators and bias-reduced estimators. The bias-reduced estimators include those that estimate bias and subtract it from the MLE and those that use shrinkage methods. Detailed descriptions and explanations of the different types of point estimators referenced in Table 3 are given in the following two sections, with Section 5 focusing on unbiased estimation (both mean-unbiased and median-unbiased) and Section 6 focusing on bias-reduced estimation (including Bayesian approaches and resampling-based methods).

Looking at the results as a whole, we can see that the literature on unbiased and bias-adjusted estimation for adaptive designs has grown rapidly in recent years, reflecting the increased use of adaptive designs in practice. However, only a few of these methodological papers present real-life trial examples (such as reanalyzes of trials that have already been

TABLE 3 Summary of unbiased and bias-reduced estimators, trial examples, and software/code for different classes of adaptive designs

Design	Method(s)	Pros and cons	Trial examples and software/code
Group sequential	*Mean-unbiased estimation* Chang et al,[32] Kim,[33] Emerson and Fleming,[34] Emerson and Kittelson,[35] Liu and Hall,[36] Jung and Kim,[37] Liu et al,[38,39] Porcher and Desseaux,[40] Zhao et al[41] *Secondary parameters*: Gorfine,[42] Liu and Hall,[43] Liu et al,[44] Kunz and Kieser[45] *Diagnostic tests*: Shu et al,[46] Pepe et al[47] *Median-unbiased estimation* Kim,[33,48] Emerson and Fleming,[34] Todd et al,[49] Troendle and Yu,[50] Hall and Liu,[51] Koyama and Chen,[52] Wittes,[53] Porcher and Desseaux,[40] Shimura et al[54] *Secondary parameters*: Hall and Yakir[55] *Adaptive group sequential trials*: Wassmer,[56] Brannath et al,[57] Gao et al,[58,59] Levin et al,[60] Mehta et al,[61] Nelson et al[62]	The UMVUE has zero bias, but tends to have a higher MSE than the naive or bias-adjusted estimators Computation can be complex and extensive if the number of looks is relatively large (≥ 4), but otherwise can be simpler than for bias-adjusted estimators UMVUE can be conditionally biased MUE reduces bias compared to the naive estimator, but can have an increased MSE. Bias-adjusted estimators can have a lower MSE as well Calculation of the MUE can be complicated, and results can depend on the ordering of the sample space MUEs can be derived for adaptive group sequential designs, unlike for other estimation methods	Beta-Blocker Heart Attack Trial, see Gorfine[42] Phase II trial in patients with adenocarcinoma, see Kunz and Kieser[45] GI06-101 trial in hepatobiliary cancer, see Zhao et al[41] *Software/code*: OptGS R package Trials for acute bronchitis reported in Wassmer et al[63] Multicenter Automatic Defibrillator Implantation Trial (MADIT), see Hall and Liu,[51] Hall and Yakir[55] Non-small cell lung cancer trial, see Wassmer[56] Randomized Aldactone Evaluation Study (RALES), see Wittes[53] Trial reported in Troger et al[64] Clinical Evaluation of Pertuzumab and Trastuzumab (CLEOPATRA) trial, see Shimura et al[54] *Software/code*: ADDPLAN[59] *Note*: Also implemented in software such as SAS, East, as well as R packages such as rpact, RCTdesign, AGSDest, and OptGS. See https://panda.shef.ac.uk/techniques/group-sequential-design-gsd/categories/27

TABLE 3 (Continued)

Design	Method(s)	Pros and cons	Trial examples and software/code
Resampling	*Parametric*: Pinheiro and DeMets,[65] Wang and Leung,[66] Leung and Wang,[67] Cheng and Shen,[68] Magnusson and Turnbull[17] *Nonparametric*: Leblanc and Crowley[69]	Essentially the same procedure can be used under different stopping rules and different study designs Bootstrap algorithms can be computationally intensive Bias is substantially reduced, with reasonable MSE Nonparametric approaches are robust to model misspecification	Trial for nasopharyngeal cancer, see Leblanc and Crowley[69]
Bias-reduced	Whitehead,[70] Chang et al,[32] Tan and Xiong,[71] Todd et al,[49] Li and DeMets,[72] Fan et al,[73] Liu et al,[74] Guo and Liu,[75] Porcher and Desseaux,[40] Shimura et al,[76] Li[77] *Adaptive group sequential trials*: Levin et al[60] *Conditional*: Troendle and Yu,[50] Coburger and Wassmer[78] *Secondary parameters*: Whitehead,[79] Liu et al,[44,74] Yu et al[80] *Health economic outcomes*: Flight[81]	The MSE of the bias-adjusted MLE is typically lower than that of the UMVUE, particularly for small sample sizes Shrinkage-type estimators (which also have a Bayesian interpretation) can reduce both the conditional bias and MSE further	Three different phase II studies, see Tan and Xiong[71] Trial of immunosuppression for bone marrow transplantation, see Whitehead[79] Two cardiovascular trials (MERIT-HF and COPERNICUS), see Fan et al[73] Trial in familial adenomatous polyposis, see Liu et al[74] Phase II trial in endometrial cancer, see Shimura et al[76] GUSTO trial for myocardial infarctions, see Marschner and Schou[82] *Software/code*: RCTdesign R package, OptGS R package[60]
Bayesian	Hughes and Pocock,[83] Pocock and Hughes[84]	Useful for producing shrinkage of unexpectedly large/imprecise observed treatment effects that arise in trials that stop early Bias reduction depends on the specification of the prior distribution.	None

TABLE 3 (Continued)

Design	Method(s)	Pros and cons	Trial examples and software/code
Sample size re-estimation	*Mean-unbiased estimation* *Unconditional:* Liu et al,[85,86] Kunzmann and Kieser[87] *Conditional (UMVCUE):* Kunzmann and Kieser,[87] Broberg and Miller[88]	Estimators have zero bias, either unconditionally or conditionally. However, the MSE tends to be greater than for the naive estimator and bias-reduced estimators Does not guarantee compatibility with the test decision (see Kunzmann and Kieser[87]) Estimators can have an explicit representation making computation easy	Schizophrenia trial, see Broberg and Miller[88]
	Median-unbiased estimation Lawrence and Hung,[89] Liu et al,[85] Wang et al,[90] Liu et al,[86] Kunzmann and Kieser,[87] Nhacolo and Brannath[91] *Conditional perspective:* Broberg and Miller[88] *Flexible sample size adaptations:* Bauer et al,[92] Liu and Chi,[93] Brannath et al,[94] Lawrence and Hung,[89] Proschan et al,[95] Brannath et al[96]	MUE tends to have small mean bias, and can also have smaller MSE than the naive estimator Does not guarantee compatibility with the test decision MUE can be calculated for flexible adaptation rules that are not completely prespecified in advance, unlike for other estimation methods	Coronary artery disease trial, see Wang et al[90] Trial on reperfusion therapy for acute myocardial infarction, see Brannath et al[96] *Software/code:* R packages such as rpact and adpss
	Bias-reduced Denne,[97] Coburger and Wassmer,[98] Cheng and Shen,[99] Shen and Cheng,[100] Liu et al,[85] Tremmel,[101] Broberg and Miller[88]	Proposed bias-reduced estimates are nearly unbiased with practical sample sizes, with similar variance to the naive estimator Numerical problems can occur when calculating adjusted estimators, and observations close to the critical boundaries can lead to unreasonably extreme adjusted estimators	Type II Coronary Intervention Study, see Denne[97] Colon cancer trial, see Shen and Cheng[100] Trial in chronic lymphocytic leukemia, see Tremmel[101] Schizophrenia trial, see Broberg and Miller[88]
	Bayesian Kunzmann and Kieser,[87] Grayling and Mander[102]	Guarantees compatibility with the test decision Reduces MSE of MLE except for very small or very large values of the success probability Reduces absolute bias compared with MLE except for small values of the success probability, where there can be a substantial positive bias Can reduce MSE substantially compared to the UMVUE for certain response rates	None

TABLE 3 (Continued)

Design	Method(s)	Pros and cons	Trial examples and software/code
Multi-arm multi-stage designs (with treatment selection)	*Mean-unbiased estimation* *Two-stage designs*: Cohen and Sackrowitz,[103] Tappin,[104] Bowden and Glimm,[105,106] Pepe et al,[47] Koopmeiners et al,[107] Robertson et al,[108,109] Robertson and Glimm[110] *Multi-stage designs*: Bowden and Glimm,[106] Stallard and Kimani[111] *Seamless phase II/III trials*: Kimani et al,[112] Robertson et al[109]	UMVCUEs are conditionally unbiased. Compared to the MLE, the conditional MSE tends to be lower, but unconditionally the MSE can substantially increase UMVCUEs in the literature tend to have a closed-form expression, allowing for easy computation In some settings, the UMVCUE can have comparable MSE to bias-adjusted estimators	Trial for the treatment of anxiety disorder, see Kimani et al[112] and Robertson et al[109] INHANCE study, see Robertson and Glimm[110] ADVENT trial, see Stallard and Kimani[111] PROVE trial[113] implements approach of Stallard and Kimani[111] *Software/code*: Bowden and Glimm[106]
	Resampling Pickard and Chang,[114] Whitehead et al[115]	Provides a reasonable balance between bias and MSE across several scenarios Approach can be applied to endpoints coming from a variety of distributions (including normal and binomial) Approaches are robust to model misspecification	*Software/code*: Whitehead et al[115]
	Bias-reduced *Two-stage designs*: Coad,[116] Shen,[117] Stallard et al,[118] Pepe et al,[47] Luo et al,[119,120] Bebu et al,[121,122] Koopmeiners et al,[107] Brückner et al[123] *Multi-stage designs*: Coad,[116] Stallard and Todd,[124] Bowden and Glimm[106] *Seamless phase II/III trials*: Kimani et al[112]	Bias-adjusted MLE can have relatively low MSE and acceptably small bias in some scenarios Shrinkage methods can be the most effective in reducing the MSE Bias-reduced estimators can run into computational/convergence problems Estimators can overcorrect for bias	Phase II study in colorectal cancer, see Luo et al[119] *Software/code*: Luo et al[119] FOCUS trial in advanced colorectal cancer, see Brückner et al[123] Phase III trial in Alzheimer's, see Stallard and Todd[124] *Software/code*: Bowden and Glimm[106] *Software/code*: Kimani et al[112]
	Bayesian *Two-stage designs*: Carreras and Brannath,[125] Bowden et al,[126] Brückner et al[123] *Multi-stage designs*: Bunouf and Lecoutre[127]	Shrinkage estimators can perform favorably compared with the MLE in terms of bias and MSE Approaches can run into practical and theoretical complications, necessitating further modifications such as estimating a between-arm heterogeneity parameter	FOCUS trial in advanced colorectal cancer, see Brückner et al[123]

TABLE 3 (Continued)

Design	Method(s)	Pros and cons	Trial examples and software/code
Response-adaptive randomization	*Mean-unbiased estimation* Bowden and Trippa[128]	Mean-unbiased estimators can have a large MSE and can be very computationally intensive to calculate	Glioblastoma trial with multiple treatments
	Bias-reduced Coad,[129] Morgan,[130] Marschner[131]	The conditional MLE can be very effective at eliminating the conditional bias that is present in the unconditional MLE, but this comes at the cost of a loss of efficiency except for more extreme designs The penalized MLE exhibits very little conditional bias and is not subject to substantial efficiency loss (compared to the unconditional MLE) when the realized design is close to its average	None
Adaptive enrichment designs	*Mean-unbiased estimation* *Two-stage designs*: Kimani et al,[132-134] Kunzmann et al,[135] Di Stefano et al[136]	UMVCUE is unbiased, but tends to have a higher MSE than the MLE UMVCUE can also have a larger MSE than shrinkage/bias-adjusted estimators, but compensates by the corresponding eradication of bias UMVCUEs are easily computable, as they have a closed form expression	MILLY phase II study in asthma, see Kunzmann et al[135] *Software/code*: Kimani et al[134]
	Bias-reduced *Two-stage designs*: Kunzmann et al,[135] Kimani et al,[133] Di Stefano et al[136]	Can have lower MSE than UMVCUE that is comparable to the MSE of the MLE, but has residual bias	MILLY phase II study in asthma, see Kunzmann et al[135]
	Resampling Magnusson and Turnbull,[17] Kunzmann et al,[135] Simon and Simon[137]	Bootstrap estimator can have a higher bias than the MLE in the two-stage setting of Kunzmann et al,[135] but Simon and Simon[137] found it was very effective at correcting for bias in their multimarker setting Procedures can be computationally intensive	MILLY phase II study in asthma, see Kunzmann et al[135]
	Bayesian *Two-stage designs*: Kunzmann et al,[135] Kimani et al,[133] Di Stefano et al[136]	Estimators exhibit a higher bias than the MLE in many situations, with varying MSE properties	MILLY phase II study in asthma, see Kunzmann et al[135]

Abbreviations: MLE, maximum likelihood estimator; MSE, mean squared error; MUE, median unbiased estimator; UMVCUE, uniformly minimum variance conditionally unbiased estimator; UMVUE, uniformly minimum variance unbiased estimator.

completed), and even fewer provide software/code. In terms of the different classes of adaptive designs, group sequential designs have received the most attention by far, and some of the estimation methods are implemented in widely available statistical software such as R or SAS. This reflects the relatively long history of group sequential designs and their more widespread use in practice. In contrast, we found very little literature on estimation for trials using RAR. With the increasing use of such trials (see, eg, the ISPY-2[138] and BATTLE[139] trials), this may be an important gap in the literature to fill.

Another potential gap is that some of the methods are restricted to two-stage designs (especially for adaptive enrichment trials), and so a natural extension would be to develop these methods for multi-stage trials. As well, most of the methodology has focused on binary and normally distributed outcomes, with comparatively few proposals tailored for trials with time to event outcomes. It is also unclear to what extent existing methodology can apply to longitudinal outcomes (which may be particularly common for interim analyses) or to cluster randomized trials.

In terms of different types of estimators, generally speaking there is a bias-variance tradeoff (see the start of Section 6) unbiased estimators may pay the "price" for complete unbiasedness by having a high variance. Conversely, bias-reduced estimators may have a lower variance than unbiased estimators, but may "overcorrect" and thereby introduce bias in the opposite direction. Another general observation is that mean-unbiased estimators in the literature require pre-specified boundaries and decision rules, and so are not applicable to flexible adaptive designs with arbitrary changes. Finally, resampling-based methods have received comparatively little attention as opposed to other types of bias-reduced estimation. For example, as far as we could tell, there has not been work looking explicitly at their use for sample size re-estimation designs. Given that resampling-based methods can be applied even to complex trial designs (see, eg, Whitehead et al[115]), their use could potentially provide one solution to the problem of estimation for trials which combine different types of adaptations together and hence will not fit neatly into the classes of adaptive designs used above.

5 | UNBIASED ESTIMATION

5.1 | Mean-unbiased estimation

In order to achieve exact mean-unbiasedness, typically it is necessary to find an estimator whose distribution is independent of the trial adaptations. One key setting where this is possible is in classical group sequential trials. Given that there is a prespecified number of observations at the time of the first interim analysis, then the sample mean (which corresponds to the MLE) at the end of the first stage is (unconditionally) unbiased as no adaptations to the trial have occurred. However, this estimator is also clearly inefficient for estimating the overall treatment effect, as it does not use any information that may arise from later stages of the trial.

Another setting where unbiased estimation is possible is multi-stage trials with ranking and selection from a subset of candidates, such as candidate treatments or patient subgroups. The sample mean (MLE) for the selected candidate(s) calculated using only the final stage data will be conditionally unbiased, conditional on the ranking and selection that has taken place in the previous stages. However, again this estimator is clearly inefficient as it ignores all the information about the selected candidate(s) from the previous stages of the trial. Moreover, unlike the first stage estimator described in the previous paragraph, it cannot be used to estimate the effect of a treatment or subpopulation that was dropped at an interim analysis.

5.1.1 | Minimum variance unbiased estimation

One way to obtain a more efficient estimator that uses more of the information from the trial, while still maintaining exact mean-unbiasedness, is to apply the Rao-Blackwell theorem. This theorem implies that if U is an unbiased estimator of the unknown parameter of interest θ, and T is a sufficient statistic, then the estimator $\hat{U} = E(U|T)$ is unbiased and $\text{var}(\hat{U}) \leq \text{var}(U)$. In certain cases, this "Rao-Blackwellization" technique allows the derivation of the unbiased estimator that has the smallest possible variance, which is known as the uniformly minimum variance unbiased estimator (UMVUE). The Lehmann-Scheffé theorem[140] states that if $\mathcal{T}$ is a sufficient and complete statistic, and U is an unbiased estimator, then the Rao-Blackwellized estimator $\hat{U} = E(U|\mathcal{T})$ is the unique UMVUE.

The derivation of UMVUEs (from an unconditional perspective) has been a focus in the literature on group sequential trials, with work by Chang et al[32] and Jung and Kim[37] for binary response data (see also Porcher and Desseaux[40] and Zhao et al[41] for Simon's two-stage designs), and Kim,[33] Emerson and Fleming,[34] and Emerson and Kittelson[35] for normally distributed endpoints. Liu and Hall[36] proved that the sufficient statistic for the latter setting is in fact not complete, but that the Rao-Blackwellized estimator is still UMVUE among all "truncation-adaptable" unbiased estimators—that is, where inference following stopping early does not require knowledge of the future analyses. For simplicity, in what follows we do not make this technical distinction when describing UMVUEs. Similar, more general theoretical results can be found in Liu et al[38] for group sequential tests for distributions in a one-parameter exponential family and Liu et al[39] for multivariate normal group sequential tests. See also Liu et al[141] and Liu and Pledger[142] for further theoretical results for general two-stage adaptive designs.

Methodology has also been developed to calculate the UMVUE for secondary parameters or endpoints in group sequential trials. Liu and Hall[43] derived the UMVUE in the context of correlated Brownian motions, while Gorfine[42] considered normally distributed endpoints where the secondary parameter is the mean in a subgroup of subjects. Similarly, Liu et al[44] derived the UMVUE for a secondary probability, such as the rate of toxicity, in group sequential trials with binary endpoints. Kunz and Kieser[45] also derived UMVUEs for secondary endpoints in two-stage trials with binary endpoints. Further settings include estimating the sensitivity and specificity of a diagnostic test using a group sequential design, where UMVUEs have been derived by Shu et al[46] and Pepe et al.[47]

Moving away from classical group sequential designs, Liu et al[86] derived the UMVUE for a two-stage adaptive design (with sample size adjustment based on conditional power). Kunzmann and Kieser[87] derived the UMVUE general two-stage adaptive designs with binary endpoints, where the second stage sample size is determined by a function of the response rate observed in stage one. Meanwhile, Liu et al[85] proposed Rao-Blackwellized unbiased estimators for both primary and secondary endpoints in the context of a two-stage adaptive trial with sample size re-estimation. Bowden and Trippa[128] showed how to calculate a Rao-Blackwellized unbiased estimator for trials using RAR with binary endpoints.

In some trial contexts, it may be more appropriate to find unbiased estimators *conditional* on the adaptations that have taken place. For example, in multi-stage trials with treatment selection, typically there would be greater interest in estimating the properties of the better-performing treatments that are selected at the end of the trial, rather than those treatments that are dropped for futility. In addition, it may not be possible to find a complete statistic for the parameter of interest without additionally conditioning on the selection rule used (see, eg, Cohen and Sackrowitz[103]). For both these reasons, in settings such as multi-stage trials with ranking and selection, there has been a focus on deriving uniformly minimum variance *conditionally* unbiased estimators (UMVCUEs, also known as CUMVUEs), where the conditioning is on the triggered adaptations.

One of the first papers to take this conditional perspective and calculate the UMVCUE was Cohen and Sackrowitz,[103] in the context of a two-stage drop-the-loser trial with treatment selection and normally distributed endpoints (see Tappin[104] for the setting with binary endpoints), where only the best-performing treatment is taken forward to the final stage. This was subsequently extended by Bowden and Glimm[105] to allow the calculation of the UMVCUE for treatments that were not the best-performing, Bowden and Glimm[106] for multi-stage drop-the-loser trials (this was only a Rao-Blackwellized estimator) and Robertson and Glimm[110] for unknown variances. Meanwhile, Koopmeiners et al[107] derived the UMVCUE for a two-stage trial evaluating continuous biomarkers. A similar line of work focused on deriving UMVCUEs for binary data, with Pepe et al[47] looking at two-stage designs testing the sensitivity of a dichotomous biomarker (this UMVCUE can also be applied for Simon's two stage designs, see Porcher and Desseaux,[40] which was extended by Robertson et al[108]).

For seamless phase II/III trials with a normally distributed endpoint, Kimani et al[112] showed how to calculate a Rao-Blackwellized conditionally unbiased estimator, which was subsequently extended by Robertson et al.[109] More recently, Stallard and Kimani[111] derived the UMVCUE for MAMS trials (again with normally distributed endpoints) with treatment selection and early stopping for futility, conditional on any prespecified selection or stopping rule.

A recent setting where UMVCUEs have been derived is two-stage adaptive enrichment designs, where biomarkers are used to select a patient population to investigate in the second stage. Kimani et al[132] and Kunzmann et al[135] derived the UMVCUE for such designs with normally distributed endpoints, which was extended by Kimani et al[133] to allow the biomarker cut-off to be determined by the first stage data, and Kimani et al[134] and Di Stefano et al[136] for time-to-event endpoints. Finally, in the context of two-stage adaptive designs where the second-stage sample size depends on the result of the first stage through a prespecified function, Kunzmann

and Kieser[87] and Broberg and Miller[88] considered UMVCUEs for normally distributed and binary endpoints, respectively.

5.2 | Median-unbiased estimation

While exact mean-unbiasedness is a common criterion to aim for, it is not always feasible or desirable to use mean-unbiased estimators. The main potential disadvantage is an increase in the MSE when compared with the usual MLE, due to the bias-variance trade-off. As well, the calculation of Rao-Blackwellized estimators may become difficult or infeasible for more flexible trial designs, where adaptations are not fully prespecified in advance. For both these reasons, median-unbiased estimators (MUEs) have been proposed, which can have a lower MSE than mean-unbiased estimators in some settings, and can be derived for a wider class of adaptive trial designs.

Consider testing a hypothesis about some parameter of interest θ. One common approach to calculating MUEs is as follows:

1. Define an ordering of the trial design space with respect to evidence against H_0. For example, in group sequential trials the predominant option is stage-wise ordering: this depends on the boundary crossed, the stage at which stopping occurs and the value of the standardized test statistic (in decreasing order of priority).
2. Using this ordering, define a P-value function $P(\theta)$ which gives the probability that, at the stage the trial stopped, more extreme evidence against H_0 could have been observed.
3. At the point the trial stops, find the MUE $\widehat{\theta}_{\text{MU}}$, which satisfies $P(\widehat{\theta}_{\text{MU}}) = 0.5$.

The last step is essentially finding a 50% confidence bound for θ, or equivalently the midpoint of a symmetric $100(1 - \alpha)\%$ confidence interval (constructed using the P-value function, which takes into account the previous adaptations or stopping rules).

A number of papers have derived MUEs for group sequential trials, including Kim,[33,48] Emerson and Fleming,[34] Todd[49] and Troendle and Yu.[50] Hall and Liu[51] provided MUEs that account for overrunning, while Hall and Yakir[55] derived MUEs for secondary parameters. Note also early work by Woodroofe[143] showed how to calculate a MUE but in the context of sequential probability ratio tests. Meanwhile, MUEs for Simon's two-stage design were derived by Koyama and Chen.[52]

For adaptive group sequential trials Wassmer[56] and Nelson et al[62] showed how to calculate MUEs for normally distributed endpoints, while Brannath et al[57] derived a MUE for survival endpoints. These results were subsequently extended by Gao et al,[59] who presented general theory for point and interval estimation (see also Mehta et al[61]). Gao et al[58] also derived a MUE for an adaptive group sequential design that tests for noninferiority followed by testing for superiority.

Moving away from group sequential trials, Wang et al[90] and Liu et al[86] focused on two-stage adaptive designs with sample size adjustments based on conditional power considerations, and showed how to calculate MUEs in these settings. Lawrence and Hung[89] proposed a MUE for the setting where the maximum information of a trial is adjusted on the basis of unblinded data (see also Liu et al[85] for an improved version of this estimator). More generally, MUEs for two-stage trials with an adaptive sample size based on a prespecified function of the interim results are given by Kunzmann and Kieser[87] and Nhacolo and Brannath.[91] Meanwhile, Li et al[144] derived MUEs in the context of "2-in-1" adaptive designs, which allow for a seamless expansion of a phase II trial into a phase III trial or an expansion of biomarker positive populations, where the endpoints for the interim and final analyses can be different.

Like for mean-unbiased estimation, the conditional perspective has also been used for median-unbiased estimation. The previously mentioned work by Hall and Yakir[55] for group sequential trials is an early example of this conditional perspective, while the approach of Koopmeiners et al[107] can also be used to calculate a conditional MUE for two-stage group sequential designs. In the specific context of two-stage group sequential designs with survival endpoints, Shimura et al[54] showed how to calculate conditional MUEs, for the cases where the study stops or does not stop early for either efficacy or futility. Meanwhile, Broberg and Miller[88] considered the general setting of a two-stage sample size adjustable design based on interim data, and derived the MUE conditional on continuing to the second stage.

MUEs can be derived for flexible adaptive designs, which allow arbitrary changes to the trial at each stage based on the interim results and/or external information. Bauer et al[92] first described the construction of MUEs for two-stage designs based on combination tests and conditional error functions; see also Liu and Chi,[93] Proschan et al,[95] and Brannath et al.[96]

Finally, Brannath et al[94] gave a general method to calculate MUEs when using recursive combination tests for flexible multi-stage adaptive trials.

6 | BIAS-REDUCED ESTIMATION

While using an unbiased estimator whenever there is one available for a particular adaptive design may seem like a straightforward choice, this may not always be the best option, due to the inherent bias-variance tradeoff: reducing the bias of an estimator can come at the cost of increasing its variance (thus decreasing its precision) substantially. For example, estimation procedures that involve conditioning on a statistic that is informative about the parameter of interest (eg, the stopping time) address the bias, but discard the information that is contained in that statistic and hence lead to a loss of efficiency. The bias-variance tradeoff can lead to unbiased treatment effect estimates being accompanied by large standard errors and wide confidence intervals, which may be of little use practically. Therefore, reducing the bias but not completely eliminating it sometimes proves to be a better solution than aiming for complete unbiasedness.

6.1 | Analytical and iterative procedures

Cox recognized as early as 1952[145] that in sequential test procedures the option to stop early introduces bias to the MLE. They proposed a modified MLE which can be used for repeated significance tests[146] but not more general group sequential tests. Whitehead[70] focused on the sequential probability ratio test[147] and the triangular test,[148] and derived an analytic expression to quantify the bias and proposed a bias-adjusted estimator $\tilde{\theta}$ obtainable by subtracting the bias from the MLE:

$$\tilde{\theta} = \widehat{\theta} - \text{bias}(\tilde{\theta}).$$

Evaluating the bias at $\tilde{\theta}$ leads to an iterative procedure, and this method can be used for various types of endpoints including continuous, binary, and time-to-event. Guo and Liu[75] argued that a simpler single-iteration version could be used, where the estimated bias of the MLE is subtracted rather than the bias at the unknown $\tilde{\theta}$:

$$\tilde{\theta}_s = \widehat{\theta} - \text{bias}(\widehat{\theta}).$$

They suggested that this approach could be extended from single-arm phase II trials to more general estimation problems involving designs with early stopping. Note that these proposals aim to reduce the unconditional bias of the MLE, see below for the conditional perspective.

Chang et al[32] and applied Whitehead's idea to group sequential designs with binary response data, while Tan and Xiong[71] proposed using the bias-adjusted MLE to estimate response rates in trials using fully sequential and group sequential conditional probability ratio tests. Li and DeMets[72] derived an exact analytical expression for the bias of the MLE for a group sequentially monitored Brownian motion process with a linear drift, and used it to construct a bias-adjusted estimator based on the arguments of Whitehead.[70] Todd[49] studied the bias-adjusted MLE further in simulations of group sequential designs with a normally distributed endpoint using the O'Brien-Fleming or triangular test. Denne[97] postulated that the same bias-adjusted MLE could also be used to estimate the treatment effect following a two-stage sample size re-estimation design on the basis of conditional power, which was subsequently investigated by Liu et al[85] (who also considered a bias-adjusted MLE for secondary endpoints). Levin et al[60] found that the bias-adjusted MLE could be applied to a wider class of adaptive group sequential designs with predefined rules.

Whitehead[79] extended the idea of the bias-adjusted MLE to the estimation of secondary endpoints in sequential trials using two different approaches: conditional on the primary endpoint (in which case the MLE is unbiased) and unconditional. Liu et al[44] further extended Whitehead's approach to estimate secondary probabilities after termination of a sequential phase II trial with response rate as the primary endpoint. The bias-adjusted MLE has also been proposed for a group sequential strategy to monitor toxicity incidents in clinical trials by Yu et al,[80] for which they noted the bias of the usual MLE was "considerably large." Meanwhile, Flight[81] showed how to use the bias-adjusted MLE for health economic outcomes following a group sequential design.

For trials with treatment selection, early work by Coad[116] in the setting with two treatments and either a two or three-stage trial (where only one treatment is selected to be carried forward to the final stage), showed how to derive

a bias-adjusted MLE. In the context of adaptive enrichment designs with time-to-event endpoints, Di Stefano et al[136] derived both single and multiple iteration versions of a bias-adjusted MLE. Finally, for response-adaptive trials, Coad[129] and Morgan[130] showed how to approximate the bias of the MLE, which then allows the construction of a bias-adjusted estimator following the idea of Whitehead.[70]

6.1.1 | Conditional perspective

Troendle and Yu[50] developed methods for estimating treatment differences in group sequential designs with normally distributed endpoints that condition on the stopping time S of the trial, thereby reducing the conditional bias, that is, the discrepancy between the conditional expectation of the estimator and the parameter value:

$$\text{Conditional bias}(\theta) = E(\widehat{\theta}|S) - \theta.$$

They used a similar iterative approach as Whitehead[70] to compute the estimator. Coburger and Wassmer[98] took this one step further, by conditioning both on stopping time as well as the sample sizes of the stages up to that point; whereas Troendle and Yu had assumed equal sample sizes for all stages. They also considered adaptive group sequential designs, where the unconditional bias is no longer computable, but the conditional bias can be approximated. Tremmel[101] extended these considerations to stratified binomial and time-to-event endpoints. Coburger and Wassmer[78] suggested using a similar conditional bias-adjusted estimator not only when a trial stops but also during the course of an adaptive sample size re-estimation design, to prevent an overestimation of the sample sizes for subsequent stages.

Fan et al[73] showed that the conditional bias-reduced estimator for group sequential designs as proposed by Troendle and Yu[50] is equivalent to a conditional MLE as well as to a conditional moment estimator. They also proposed two modified estimators which are hybrids of the conditional and unconditional MLE that have smaller variance and MSE. The relationship between the conditional MLE and the conditional bias-reduced estimator was also pointed out by Liu et al,[74] who additionally derived the conditional MLE for secondary endpoints in a group sequential design and adaptive group sequential design. More general theoretical results for sequential trials focusing on the conditional MLE can be found in Molenberghs et al[149] and Milanzi et al.[150,151]

Marschner[131] recently extended this notion of conditional MLEs to more general adaptive designs, as part of a unifying formulation of adaptive designs and a general approach to their analysis. More specifically, 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 (conditional on the realized design). Marschner also proposed a penalized MLE, which weights the design likelihood and can vary between the two extremes of the unconditional and conditional MLEs. Examples of the conditional and penalized MLEs were given in the context of a RAR design.

Shimura et al[76] modified Troendle and Yu's conditional bias-reduced estimator for two-stage group sequential trials with normal endpoints by combining it with the shrinkage idea of Thompson[152]: in the calculation of the bias adjustment term, they replaced the MLE with a weighted average of the MLE at the interim analysis and a "prior" estimate of the effect size, $\widehat{\theta}_* = c\widehat{\theta} + (1-c)\theta_0$, with θ_0 the effect size assumed in the initial sample size calculation and shrinkage weights c determined as a function of the MLE as proposed by Thompson.[152] Although the authors presented their idea in a frequentist framework, it can also be interpreted as Bayesian, with θ_0 the prior and $\widehat{\theta}_*$ the posterior expectation. Meanwhile, Li[77] proposed using a weighted sum of the MLE and sample proportion for two-stage group sequential designs with binary responses, where the weight is based on estimates of the MSEs of these estimators. Pepe et al[47] and Koopmeiners et al[107] considered conditional bias-reduced estimation in the context of group sequential diagnostic biomarker studies with a single interim analysis for futility, conditional on not stopping early.

Cheng and Shen[99] derived an easily computable bias-adjusted estimator of the treatment effect in a "self-designing" trial (where the sample size is sequentially determined to achieve a target power) with a normally distributed endpoint.[153] They further proposed a modified estimator, involving a shrinkage factor, to reduce bias when block sizes are small. For "self-designing" trials with censored time-to-event endpoints and group sequential sample size re-estimation,[154] Shen and Cheng[100] developed bias-adjusted estimators of the hazard ratio. Broberg and Miller[88] considered two-stage sample size adjustable designs with normally distributed endpoints and derived a conditional bias-adjusted MLE. Meanwhile, in the context of "2-in-1" adaptive trial designs, Li et al[144] derived a conditionally bias-adjusted estimator.

Moving on to trial designs with treatment selection, when multiple treatments are compared to a control and the most promising treatment is selected at the first interim analysis (with or without further early stopping opportunities at

subsequent interim analyses), Stallard and Todd[124] developed a bias-adjusted estimator for the treatment effect. Kimani et al[112] adapted this estimator for the difference between the selected and control treatments in seamless phase II/III designs with normally distributed endpoints. Unlike Stallard and Todd's estimator,[124] which aimed to be approximately unbiased conditional on the selected treatment, this new estimator aimed to be approximately unbiased conditional on both the selected treatment and continuation to the second stage. In the context of selecting the treatment dose with the highest response rate (using a normal approximation), Shen[117] also derived a stepwise bias overcorrection. When there are two treatments being compared with a control, Stallard et al[118] proposed a family of approximately conditionally unbiased designs.

Brückner et al[123] extended Stallard and Todd's bias-reduced estimator to two-stage multi-arm designs involving a common control and treatment selection at the interim analysis with time-to-event endpoints. Despite deriving analytical expressions for the selection bias of the MLE in the selected and stopped treatment arms, respectively, they experienced computational difficulties and convergence problems. They only defined the resulting bias-reduced estimator for the interim analysis, and proposed using a two-stage estimator for the final analysis, which is a weighted average of the bias-reduced estimator at the interim analysis and the MLE.

Luo et al[119] used the method of conditional moments to derive a bias-adjusted estimator of the response rate following a two-stage drop-the-loser design with a binary endpoint. Adopting a stochastic process framework (as opposed to the conventional random variable viewpoint) for clinical trials with adaptive elements, Luo et al[120] proposed approximating treatment effects based on a general estimating equation to match the first conditional moment. This generic approach can be used for a variety of adaptive designs with endpoints following any probability distribution. They illustrated it for two-stage treatment selection designs comparing multiple treatments against a control.

Kunzmann et al[135] applied the conditional moment estimator of Luo et al[119] to two-stage adaptive enrichment designs with a normally distributed endpoint and binary biomarker with a prespecified cut-off. Additionally, they studied two hybrid estimators, one between the UMVCUE and the conditional moment estimator, and the other one combining the UMVCUE and MLE. Kimani et al[133] studied single- and multi-iteration bias-adjusted estimators for the treatment effect in the selected subpopulation of a two-stage adaptive threshold enrichment design allowing the cut-off of a predictive biomarker based on first stage data.

Finally, Bebu et al[121] devised a bias-corrected conditional MLE for the mean of the selected treatment in two-stage designs and normally distributed endpoints. They focused on the case of two experimental treatments, but this was extended by Bebu et al[122] to designs involving the selection of more than one treatment arm, selection rules based on both efficacy and safety, including a control arm, adjusting for covariates, and binomial endpoints. Bowden and Glimm[106] further extended the method to multi-stage drop-the-loser designs which must always proceed to the final stage (ie, no early stopping rules).

6.2 | Bayesian and empirical Bayes approaches

Several Bayesian approaches to bias-reduced estimation in adaptive designs have been developed, typically for trials using frequentist frameworks. These differ from fully Bayesian adaptive designs in that they only use the accumulating data, rather than the posterior, to make decisions about adaptations (eg, stopping or selecting arms) and rely primarily on classical frequentist inference such as hypothesis testing, but often require additional assumptions or information, such as the specification of a prior.

Hughes and Pocock[83] and Pocock and Hughes[84] proposed a Bayesian shrinkage method for estimating treatment effects in trials stopped at an interim analysis, whereby a prior quantifying the plausibility of different treatment effect sizes is specified at the outset of the trial and combined with the trial data using Bayes' rule, producing a posterior estimate of the treatment effect that is shrunk toward the median of the prior distribution. They acknowledged that this will be sensitive to the choice of prior and warned against using an overoptimistic prior as this would lead to little or no shrinkage, saying that their method "works best in the hands of prior specifiers who are realists by nature."[83]

Hwang[155] proposed using Lindley's[156] shrinkage estimator for estimating the mean of the best treatment in single-stage multi-arm designs with 4 or more study treatments and a normally distributed endpoint. This was extended by Carreras and Brannath[125] to multi-arm two-stage drop-the-losers designs. Their two-stage version is a weighted mean of Lindley's estimator for the first stage and the MLE of the selected treatment for the second stage. It shrinks the MLE toward the mean across all arms, but only works with a minimum of 4 study treatment arms. For designs with 2 or 3 study treatments, they replaced Lindley's estimator with the best linear unbiased predictor of a random effects

model for the first stage estimator. This method also allows for bias-reduced estimation of the second best treatment mean.

Of note, Lindley's estimator can be viewed as an empirical Bayes estimator of the posterior mean of the posterior distribution of a specific treatment mean given its MLE. This fact was exploited by Bowden et al,[126] who adapted Carreras and Brannath's method and replaced their two-stage estimator with a single shrinkage equation for both stages. This, however, adds both practical and theoretical complications, necessitating further modifications such as estimating a between-arm heterogeneity parameter. The authors suggest this approach could be applied to other MAMS designs.

Along the lines of Hwang's approach,[155] Brückner et al[123] derived two shrinkage estimators for two-stage multi-arm designs with a common control, treatment selection at the interim analysis, and time-to-event endpoints. They both shrink the MLE of the selected treatment toward the overall log-hazard ratio. The authors also defined a two-stage estimator similar to that of Carreras and Brannath,[125] as well as a bias-corrected Kaplan-Meier estimator for the selected treatment.

Bunouf and Lecoutre[127] derived two Bayesian estimators for multi-stage designs with binary endpoints where after every stage a decision is made whether to stop or continue recruiting: one estimator is the posterior mean based on a "design-corrected" prior, the other the posterior mode based on a similarly corrected uniform prior, and for the latter they also provided an easy-to-use approximation for two-stage designs. Meanwhile, Kunzmann et al[135] developed two empirical Bayes estimators for two-stage adaptive enrichment designs with a single prespecified subgroup and prespecified decision rule based on a binary biomarker and a normally distributed endpoint with known variance. Kimani et al[133] extended this work to adaptive threshold enrichment designs where the biomarker cut-off is determined based on first stage data.

Finally, Kunzmann and Kieser[87] developed an estimator for two-stage single-arm trials with a binary endpoint with sample size adjustment and early stopping for futility or efficacy. This estimator minimizes the expected MSE subject to compatibility with the test decision and monotonicity conditions, and can be viewed as a constrained posterior mean estimate. Similarly, in the same trial context Grayling and Mander[102] proposed an optimal estimator that minimizes a weighted sum of the expected MSE and bias.

6.3 | Resampling-based methods

The methods for unbiased and biased-reduced estimation considered in the previous subsections require explicit formulas based on the population response model as well as the trial design and adaptations. For example, UMVUEs and UMVCUEs can often be given as closed-form expressions, while for MUEs a *P*-value function is often specified. Meanwhile, bias-reduced estimators tend to have explicit expressions for an estimate of the bias, leading to equations that can be solved numerically.

An alternative class of methods is based on *resampling* procedures, where the trial data is resampled or generated via a parametric bootstrap a large number of times, and the resulting trial replicates can then be used to give empirical estimates of the bias. One advantage of these methods is that essentially the same procedure can be used under a variety of different stopping rules and trial designs, including complex designs as detailed below.

In the group sequential setting, Wang and Leung[66] showed how to use a parametric bootstrap to calculate a bias-corrected estimate for the mean of normally distributed endpoints with either known or unknown variance. This methodology was generalized by Leung and Wang,[67] who proposed a generic stochastic approximation approach based on a parametric bootstrap, which can be used for non-independent and identically distributed (iid) data and a large class of trial designs, including sequential ones. Cheng and Shen[68] proposed a variant of the Wang and Leung[66] procedure for complex group sequential monitoring rules that assess the predicted power and expected loss at each interim analysis. Magnusson and Turnbull[17] proposed a double bootstrap procedure for group sequential enrichment designs incorporating subgroup selection. Meanwhile, Pinheiro and DeMets[65] described a simulation approach to estimate the bias of the MLE and then proposed constructing a bias-reduced estimator for the treatment difference, although this is equivalent to Whitehead's bias-adjusted MLE.[70]

Kunzmann et al[135] carried out further work in the context of adaptive enrichment trials. In the two-stage trial setting with a binary biomarker and normally distributed endpoints, they considered a number of alternatives to the MLE, including a parametric bootstrap procedure. More generally, Simon and Simon[137] considered multi-stage enrichment designs that develop model-based predictive classifiers based on multiple markers, and showed how to use a parametric bootstrap

method for bias correction for generic response distributions. Finally, in trials with treatment selection, Pickard and Chang[114] considered parametric bootstrap corrections for a two-stage drop-the-losers design, for normally and binomially distributed endpoints.

Thus far, all the methods presented above are fully parametric, that is, assuming that the trial outcome data comes from a probability distribution with a fixed set of parameters. An alternative approach is to use *nonparametric* bootstrap procedures to correct for bias. An early example of this approach was given by Leblanc and Crowley,[69] who proposed a nonparametric bootstrap in the context of group sequential designs with censored survival data using log rank testing. Subsequently, Rosenkranz[157] described using nonparametric bootstrap to correct the bias of treatment effect estimates following selection (eg, selecting the treatment with the maximum effect). More recently, Whitehead et al[115] proposed an estimation approach based on the method of Rao-Blackwellization (ie, targeting unbiased estimation), but used a nonparametric bootstrap procedure to recreate replicate observations in a complicated group sequential trial setting comparing multiple treatments with a control.

7 | DISCUSSION

Motivated by recent FDA guidance on estimation after adaptive designs, this article provides a review of the potential methodological solutions to biased estimation that exist in the literature. We found that there is a growing body of work proposing and evaluating a range of unbiased and bias-adjusted estimators for a wide variety of adaptive trial designs. Our ambition is that this article, combined with the annotated bibliography given in the Supplementary Information, provides an easily accessible and comprehensive resource for researchers in this rapidly evolving area. However, we note that statistical software and code to easily calculate adjusted estimators is relatively sparse, which is an obstacle to the uptake of methods in practice (see Grayling and Wheeler[158] and part II of this article series). It also remains the case that for more complex or novel adaptive designs, adjusted estimators may not exist.

From a methodological perspective, there remain important open questions regarding estimation after adaptive trials, as already mentioned at the end of Section 6. More generally, an estimator for a parameter of interest should ideally have all of the following properties:

1. Adequately reflects the adaptive design used.
2. No or small bias.
3. Low MSE (reflecting a favorable bias-variance trade-off).
4. Is easily computable.
5. A procedure for calculating an associated confidence interval that:
   a. Has the correct coverage.
   b. Is consistent/compatible with the hypothesis test decisions (including early stopping).

Constructing an overarching framework for estimation after adaptive designs that has all these properties would be very useful, although challenging. Some initial steps in this direction have been taken by Kunzmann and Kieser[87,159] in the context of adaptive two-stage single-arm trials with a binary endpoint. Even if such a framework is not feasible more generally, the issue of constructing confidence intervals remains an important question that has so far received less attention in the literature than point estimation.

The property that an estimator should "adequately reflect the adaptive design used" also has implications when considering the Bayesian analysis of adaptive designs, as pointed out by an anonymous reviewer. The likelihood function will typically be insensitive to the adaptive design and sampling scheme (see, eg, Section 4.1 of Marschner[131]). Thus, while the frequentist interpretation will take account of the adaptive design, the pure Bayesian approach following the likelihood principle will not. In practice this would mean that a pure Bayesian analysis would ignore the adaptive nature of the design and the issues discussed in this article would disappear. However, this (arguably) is intuitively unappealing and provides a rationale for evaluating Bayesian-motivated approaches from a frequentist perspective.

In part II of this article series, we explore the practical considerations surrounding the use of unbiased and bias-reduced estimators for adaptive designs. There, we review the use of such estimators in current practice and illustrate their application to a real trial design. We also provide a set of guidelines for best practice, considering their use in adaptive trials from the design stage through to the final reporting of results.

DATA AVAILABILITY STATEMENT

All of the data that support the findings of this study are available within the paper and supplementary information.

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. 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
2. 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
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. *Trials*. 2020;21(1):528. doi:10.1186/s13063-020-04334-x
4. 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
5. Quinlan J, Gaydos B, Maca J, Krams M. Barriers and opportunities for implementation of adaptive designs in pharmaceutical product development. *Clin Trials*. 2010;7(2):167-173. doi:10.1177/1740774510361542
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 Clinical Trials.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. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50(4):1029. doi:10.2307/2533441
12. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286-1290. doi:10.1111/j.0006-341X.1999.01286.x
13. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biom J*. 1999;41(6):698-696.
14. 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
15. 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
16. Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med*. 2003;22(14):2239-2256. doi:10.1002/sim.1430
17. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Stat Med*. 2013;32(16):2695-2714. doi:10.1002/sim.5738
18. Chiu YD, Koenig F, Posch M, Jaki T. Design and estimation in clinical trials with subpopulation selection. *Stat Med*. 2018;37(29):4335-4352. doi:10.1002/sim.7925
19. Robertson DS, Wason JMS. Familywise error control in multi-armed response-adaptive trials. *Biometrics*. 2019;75(3):885-894. doi:10.1111/biom.13042

20. Robertson DS, Lee KM, Lopez-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. *Stat Sci*. 2022. Accessed April 11, 2021. http://arxiv.org/abs/2005.00564

21. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Stat Med*. 2000;19(7):901-911. doi:10.1002/(sici)1097-0258(20000415)19:7<901::aid-sim405>3.0.co;2-l

22. Magirr D, Jaki T, Koenig F, Posch M. Sample size reassessment and hypothesis testing in adaptive survival trials. *PLoS One*. 2016;11(2):e0146465. doi:10.1371/journal.pone.0146465

23. U. S. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics: guidance for industry; November 2019. https://www.fda.gov/media/78495/download

24. Robertson DS, Choodari-Oskooei B, Dimairo M, et al. Point estimation for adaptive trial designs II: practical considerations and guidance. *Accepted in Stat Med*. 2022.

25. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Berlin Heidelberg: Springer; 2016.

26. 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

27. Freidlin B, Korn EL, Abrams JS. Bias, operational bias, and generalizability in phase II/III trials. *J Clin Oncol*. 2018;36(19):1902-1904. doi:10.1200/JCO.2017.77.0479

28. Murad MH, Chu H, Lin L, Wang Z. The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evid-Based Med*. 2018;23(3):84-86. doi:10.1136/bmjebm-2018-110891

29. Uschner D, Rosenberger WF. Bias control in randomized controlled clinical trials. In: Piantadosi S, Meinert CL, eds. *Principles and Practice of Clinical Trials*. Cham: Springer International Publishing; 2020:1-20. doi:10.1007/978-3-319-52677-5&uscore;219-1

30. Barton BA. Controlling bias in randomized clinical trials. In: Piantadosi S, Meinert CL, eds. *Principles and Practice of Clinical Trials*. Cham: Springer International Publishing; 2020:1-17. doi:10.1007/978-3-319-52677-5&uscore;214-1

31. 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

32. Chang MN, Wieand HS, Chang VT. The bias of the sample proportion following a group sequential phase II clinical trial. *Stat Med*. 1989;8(5):563-570. doi:10.1002/sim.4780080505

33. Kim K. Point estimation following group sequential tests. *Biometrics*. 1989;45(2):613-617. doi:10.2307/2531502

34. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika*. 1990;77(4):875-892. doi:10.1093/biomet/77.4.875

35. Emerson SS, Kittelson JM. A computationally simpler algorithm for the UMVUE of a normal mean following a group sequential trial. *Biometrics*. 1997;53(1):365-369.

36. Liu A, Hall WJ. Unbiased estimation following a group sequential test. *Biometrika*. 1999;86(1):71-78. doi:10.1093/biomet/86.1.71

37. Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials: estimation of binomial probability. *Stat Med*. 2004;23(6):881-896. doi:10.1002/sim.1653

38. Liu A, Hall WJ, Yu KF, Wu C. Estimation following a group sequential test for distributions in the one-parameter exponential family. *Stat Sin*. 2006;16:165-181.

39. Liu A, Wu C, Yu KF, Yuan W. Completeness and unbiased estimation of mean vector in the multivariate group sequential case. *J Multivar Anal*. 2007;98(3):505-516. doi:10.1016/j.jmva.2006.01.001

40. Porcher R, Desseaux K. What inference for two-stage phase II trials? *BMC Med Res Methodol*. 2012;12(1):117. doi:10.1186/1471-2288-12-117

41. Zhao J, Yu M, Feng XP. Statistical inference for extended or shortened phase II studies based on Simon's two-stage designs. *BMC Med Res Methodol*. 2015;15(1):48. doi:10.1186/s12874-015-0039-5

42. Gorfine M. Estimation of a secondary parameter in a group sequential clinical trial. *Biometrics*. 2001;57(2):589-597. doi:10.1111/j.0006-341X.2001.00589.x

43. Liu A, Hall WJ. Unbiased estimation of secondary parameters following a sequential test. *Biometrika*. 2001;88(3):895-900. doi:10.1093/biomet/88.3.895

44. Liu A, Wu C, Yu KF, Gehan E. Supplementary analysis of probabilities at the termination of a group sequential phase II trial. *Stat Med*. 2005;24(7):1009-1027. doi:10.1002/sim.1990

45. Kunz CU, Kieser M. Estimation of secondary endpoints in two-stage phase II oncology trials. *Stat Med*. 2012;31(30):4352-4368. doi:10.1002/sim.5585

46. Shu Y, Liu A, Li Z. Point and interval estimation of accuracies of a binary medical diagnostic test following group sequential testing. *Philos Trans R Soc Math Phys Eng Sci*. 2008;366(1874):2335-2345. doi:10.1098/rsta.2008.0041

47. Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility: early termination of biomarker studies. *Stat Med*. 2009;28(5):762-779. doi:10.1002/sim.3506

48. Kim K. Improved approximation for estimation following closed sequential tests. *Biometrika*. 1988;75(1):121-128. doi:10.1093/biomet/75.1.121

49. Todd S. Point and interval estimation following a sequential clinical trial. *Biometrika*. 1996;83(2):453-461. doi:10.1093/biomet/83.2.453

50. 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

51. Hall WJ, Liu A. Sequential tests and estimators after overrunning based on maximum-likelihood ordering. *Biometrika*. 2002;89(3):699-708. doi:10.1093/biomet/89.3.699

52. Koyama T, Chen H. Proper inference from Simon's two-stage designs. *Stat Med*. 2008;27(16):3145-3154. doi:10.1002/sim.3123

53. Wittes J. Stopping a trial early—and then what? *Clin Trials*. 2012;9(6):714-720. doi:10.1177/1740774512454600

54. 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

55. Hall WJ, Yakir B. Inference about a secondary process following a sequential trial. *Biometrika*. 2003;90(3):597-611. doi:10.1093/biomet/90.3.597

56. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biom J*. 2006;48(4):714-729. doi:10.1002/bimj.200510190

57. Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. *Biometrics*. 2009;65(2):539-546. doi:10.1111/j.1541-0420.2008.01101.x

58. Gao P, Liu L, Mehta C. Adaptive designs for noninferiority trials. *Biom J*. 2013;55(3):310-321. doi:10.1002/bimj.201200034

59. Gao P, Liu L, Mehta C. Exact inference for adaptive group sequential designs. *Stat Med*. 2013;32(23):3991-4005. doi:10.1002/sim.5847

60. Levin GP, Emerson SC, Emerson SS. An evaluation of inferential procedures for adaptive clinical trial designs with pre-specified rules for modifying the sample size. *Biometrics*. 2014;70(3):556-567. doi:10.1111/biom.12168

61. Mehta C, Liu L, Ghosh P, Gao P. Exact inference for adaptive group sequential designs. In: Liu R, Tsong Y, eds. *Pharmaceutical Statistics*. Springer Proceedings in Mathematics & Statistics. Vol 218. Cham: Springer International Publishing; 2019:131-139. doi:10.1007/978-3-319-67386-8&uscore;10

62. Nelson BS, Liu L, Mehta C. A simulation-based comparison of estimation methods for adaptive and classical group sequential clinical trials. *Pharm Stat*. 2022;21(3):599-611. doi:10.1002/pst.2188

63. Wassmer G, Eisebitt R, Coburger S. Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Inf J*. 2001;35(4):1131-1146. doi:10.1177/009286150103500410

64. Tröger W, Galun D, Reif M, Schumann A, Stanković N, Milićević M. *Viscum album* [L.] extract therapy in patients with locally advanced or metastatic pancreatic cancer: a randomised clinical trial on overall survival. *Eur J Cancer*. 2013;49(18):3788-3797. doi:10.1016/j.ejca.2013.06.043

65. Pinheiro JC, DeMets DL. Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika*. 1997;84(4):831-845. doi:10.1093/biomet/84.4.831

66. Wang YG, Leung DHY. Bias reduction via resampling for estimation following sequential tests. *Seq Anal*. 1997;16:249-267. doi:10.1080/07474949708836386

67. Leung DH, Wang Y. Bias reduction using stochastic approximation. *Aust N Z J Stat*. 1998;40(1):43-52. doi:10.1111/1467-842X.00005

68. Cheng Y, Shen Y. An efficient sequential design of clinical trials. *J Stat Plan Inference*. 2013;143(2):283-295. doi:10.1016/j.jspi.2012.07.015

69. LeBlanc M, Crowley J. Using the bootstrap for estimation in group sequential designs: an application to a clinical trial for nasopharyngeal cancer. *Stat Med*. 1999;18(19):2635-2644. doi:10.1002/(sici)1097-0258(19991015)18:19<2635::aid-sim200>3.0.co;2-7

70. 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

71. Tan M, Xiong X. Continuous and group sequential conditional probability ratio tests for phase II clinical trials. *Stat Med*. 1996;15(19):2037-2051. doi:10.1002/(SICI)1097-0258(19961015)15:19<2037::AID-SIM339>3.0.CO;2-Z

72. Li Z, DeMets DL. On the bias of estimation of a Brownian motion drift following group sequential tests. *Stat Sin*. 1999;9:923-937.

73. 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

74. Liu A, Troendle JF, Yu KF, Yuan VW. Conditional maximum likelihood estimation following a group sequential test. *Biom J*. 2004;46(6):760-768. doi:10.1002/bimj.200410076

75. Guo HY, Liu A. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. *J Biopharm Stat*. 2005;15(5):773-781. doi:10.1081/BIP-200067771

76. Shimura M, Maruo K, Gosho M. Conditional estimation using prior information in 2-stage group sequential designs assuming asymptotic normality when the trial terminated early. *Pharm Stat*. 2018;17(5):400-413. doi:10.1002/pst.1859

77. Li Q. An MSE-reduced estimator for the response proportion in a two-stage clinical trial. *Pharm Stat*. 2011;10(3):277-279. doi:10.1002/pst.414

78. Coburger S, Wassmer G. Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biom J*. 2003;45(7):812-825. doi:10.1002/bimj.200390051

79. Whitehead J. Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics*. 1986;42(3):461-471.

80. Yu J, Hutson AD, Siddiqui AH, Kedron MA. Group sequential control of overall toxicity incidents in clinical trials—non-Bayesian and Bayesian approaches. *Stat Methods Med Res*. 2016;25(1):64-80. doi:10.1177/0962280212440535

81. Flight L. The Use of Health Economics in the Design and Analysis of Adaptive Clinical Trials [PhD thesis]. Sheffield: University of Sheffield. University of Sheffield; 2020. http://etheses.whiterose.ac.uk/27924/

82. 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

83. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med*. 1988;7(12):1231-1242. doi:10.1002/sim.4780071204

84. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials*. 1989;10(Suppl 4):209S-221S. doi:10.1016/0197-2456(89)90059-7

85. Liu A, Wu C, Yu KF. Point and interval estimation of primary and secondary parameters in a two-stage adaptive clinical trial. *J Biopharm Stat*. 2008;18(2):211-226. doi:10.1080/10543400701697125

86. Liu Q, Li G, Anderson KM, Lim P. On efficient two-stage adaptive designs for clinical trials with sample size adjustment. *J Biopharm Stat*. 2012;22(4):617-640. doi:10.1080/10543406.2012.678226

87. Kunzmann K, Kieser M. Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint. *Stat Med*. 2017;36(6):971-984. doi:10.1002/sim.7200

88. Broberg P, Miller F. Conditional estimation in two-stage adaptive designs: conditional estimation in two-stage adaptive designs. *Biometrics*. 2017;73(3):895-904. doi:10.1111/biom.12642

89. Lawrence J, Hung HMJ. Estimation and confidence intervals after adjusting the maximum information. *Biom J*. 2003;45(2):143-152. doi:10.1002/bimj.200390001

90. Wang Y, Li G, Shih WJ. Estimation and confidence intervals for two-stage sample-size-flexible design with LSW likelihood approach. *Stat Biosci*. 2010;2(2):180-190. doi:10.1007/s12561-010-9023-0

91. Nhacolo A, Brannath W. Interval and point estimation in adaptive phase II trials with binary endpoint. *Stat Methods Med Res*. 2019;28(9):2635-2648. doi:10.1177/0962280218781411

92. Bauer P, Brannath W, Posch M. Flexible two-stage designs: an overview. *Methods Inf Med*. 2001;40(2):117-121.

93. Liu Q, Chi GYH. On sample size and inference for two-stage adaptive designs. *Biometrics*. 2001;57(1):172-177. doi:10.1111/j.0006-341X.2001.00172.x

94. Brannath W, Posch M, Bauer P. Recursive combination tests. *J Am Stat Assoc*. 2002;97(457):236-244. doi:10.1198/016214502753479374

95. Proschan MA, Liu Q, Hunsberger S. Practical midcourse sample size modification in clinical trials. *Control Clin Trials*. 2003;24(1):4-15. doi:10.1016/S0197-2456(02)00240-4

96. Brannath W, König F, Bauer P. Estimation in flexible two stage designs. *Stat Med*. 2006;25(19):3366-3381. doi:10.1002/sim.2258

97. Denne JS. Estimation following extension of a study on the basis of conditional power. *J Biopharm Stat*. 2000;10(2):131-144. doi:10.1081/BIP-100101018

98. Coburger S, Wassmer G. Conditional point estimation in adaptive group sequential test designs. *Biom J*. 2001;43(7):821-833.

99. Cheng Y, Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics*. 2004;60(4):910-918. doi:10.1111/j.0006-341X.2004.00246.x

100. Shen Y, Cheng Y. Adaptive design: estimation and inference with censored data in a semiparametric model. *Biostatistics*. 2006;8(2):306-322. doi:10.1093/biostatistics/kxl011

101. Tremmel LT. Statistical inference after an adaptive group sequential design: a case study. *Ther Innov Regul Sci*. 2010;44(5):589-598. doi:10.1177/009286151004400506

102. Grayling MJ, Mander AP. Optimised point estimators for multi-stage single-arm phase II oncology trials. *J Biopharm Stat*. 2022;23:1-15. doi:10.1080/10543406.2022.2041656

103. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Stat Probab Lett*. 1989;8(3):273-278. doi:10.1016/0167-7152(89)90133-8

104. Tappin L. Unbiased estimation of the parameter of a selected binomial population. *Commun Stat Theory Methods*. 1992;21(4):1067-1083. doi:10.1080/03610929208830831

105. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biom J*. 2008;50(4):515-527. doi:10.1002/bimj.200810442

106. Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. *Biom J*. 2014;56(2):332-349. doi:10.1002/bimj.201200245

107. 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

108. Robertson DS, Prevost AT, Bowden J. Correcting for bias in the selection and validation of informative diagnostic tests. *Stat Med*. 2015;34(8):1417-1437. doi:10.1002/sim.6413

109. Robertson DS, Prevost AT, Bowden J. Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis-driven selection rules. *Stat Med*. 2016;35(22):3907-3922. doi:10.1002/sim.6974

110. Robertson DS, Glimm E. Conditionally unbiased estimation in the normal setting with unknown variances. *Commun Stat Theory Methods*. 2019;48(3):616-627. doi:10.1080/03610926.2017.1417429

111. Stallard N, Kimani PK. Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. *Biometrika*. 2018;105(2):495-501. doi:10.1093/biomet/asy004

112. Kimani PK, Todd S, Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. *Stat Med*. 2013;32(17):2893-2910. doi:10.1002/sim.5757

113. 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

114. Pickard MD, Chang M. A flexible method using a parametric bootstrap for reducing bias in adaptive designs with treatment selection. *Stat Biopharm Res*. 2014;6(2):163-174. doi:10.1080/19466315.2014.897251

115. Whitehead J, Desai Y, Jaki T. Estimation of treatment effects following a sequential trial of multiple treatments. *Stat Med*. 2020;39(11):1593-1609. doi:10.1002/sim.8497

116. Coad DS. Sequential estimation for two-stage and three-stage clinical trials. *J Stat Plan Inference*. 1994;42(3):343-351. doi:10.1016/0378-3758(94)00033-6

117. Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. *Stat Med*. 2001;20(13):1913-1929. doi:10.1002/sim.842

118. Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. *J Stat Plan Inference*. 2008;138(6):1629-1638. doi:10.1016/j.jspi.2007.05.045

119. Luo X, Wu SS, Xiong J. Parameter estimation following an adaptive treatment selection trial design. *Biom J*. 2010;52(6):823-835. doi:10.1002/bimj.200900134

120. Luo X, Li M, Shih WJ, Ouyang P. Estimation of treatment effect following a clinical trial with adaptive design. *J Biopharm Stat*. 2012;22(4):700-718. doi:10.1080/10543406.2012.676534

121. Bebu I, Luta G, Dragalin V. Likelihood inference for a two-stage design with treatment selection. *Biom J*. 2010;52(6):811-822. doi:10.1002/bimj.200900170

122. Bebu I, Dragalin V, Luta G. Confidence intervals for confirmatory adaptive two-stage designs with treatment selection: confidence intervals for adaptive two-stage designs. *Biom J*. 2013;55(3):294-309. doi:10.1002/bimj.201200053

123. 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

124. Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. *J Stat Plan Inference*. 2005;135(2):402-419. doi:10.1016/j.jspi.2004.05.006

125. 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

126. Bowden J, Brannath W, Glimm E. Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. *Stat Med*. 2014;33(3):388-400. doi:10.1002/sim.5920

127. Bunouf P, Lecoutre B. On Bayesian estimators in multistage binomial designs. *J Stat Plan Inference*. 2008;138(12):3915-3926. doi:10.1016/j.jspi.2008.02.014

128. Bowden J, Trippa L. Unbiased estimation for response adaptive clinical trials. *Stat Methods Med Res*. 2017;26(5):2376-2388. doi:10.1177/0962280215597716

129. Coad DS. Estimation following sequential tests involving data-dependent treatment allocation. *Stat Sin*. 1994;4(2):693-700.

130. Morgan CC. Estimation following group-sequential response-adaptive clinical trials. *Control Clin Trials*. 2003;24(5):523-543. doi:10.1016/S0197-2456(03)00062-X

131. Marschner IC. A general framework for the analysis of adaptive experiments. *Stat Sci*. 2021;36(3):465-492. doi:10.1214/20-STS803

132. Kimani PK, Todd S, Stallard N. Estimation after subpopulation selection in adaptive seamless trials. *Stat Med*. 2015;34(18):2581-2601. doi:10.1002/sim.6506

133. Kimani PK, Todd S, Renfro LA, Stallard N. Point estimation following two-stage adaptive threshold enrichment clinical trials: estimators for adaptive threshold enrichment clinical trials. *Stat Med*. 2018;37(22):3179-3196. doi:10.1002/sim.7831

134. Kimani PK, Todd S, Renfro LA, et al. Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. *Stat Med*. 2020;39(19):2568-2586. doi:10.1002/sim.8557

135. Kunzmann K, Benner L, Kieser M. Point estimation in adaptive enrichment designs. *Stat Med*. 2017;36(25):3935-3947. doi:10.1002/sim.7412

136. Di Stefano F, Pannaux M, Correges A, Galtier S, Robert V, Saint-Hilary G. A comparison of estimation methods adjusting for selection bias in adaptive enrichment designs with time-to-event endpoints. *Stat Med*. 2022;41(10):1767-1779. doi:10.1002/sim.9327

137. Simon N, Simon R. Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics*. 2018;19(1):27-41. doi:10.1093/biostatistics/kxw054

138. 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

139. 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

140. Casella G, Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Thomson Learning; 2002.

141. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *J Am Stat Assoc*. 2002;97(460):1034-1041. doi:10.1198/016214502388618852

142. Liu Q, Pledger G. On design and inference for two-stage adaptive clinical trials with dependent data. *J Stat Plan Inference*. 2006;136(6):1962-1984. doi:10.1016/j.jspi.2005.08.015

143. Woodroofe M. Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika*. 1992;79(2):347-353. doi:10.1093/biomet/79.2.347

144. Li W, Bai X, Deng Q, Liu F, Chen C. Estimation of treatment effect in 2-in-1 adaptive design and some of its extensions. *Stat Med*. 2021;40(11):2556-2577. doi:10.1002/sim.8917

145. Cox DR. A note on the sequential estimation of means. *Math Proc Camb Philos Soc*. 1952;48(3):447-450. doi:10.1017/S0305004100027857

146. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc Ser Gen*. 1969;132(2):235-244. doi:10.2307/2343787

147. Wald A. *Sequential Analysis*. Hoboken, NJ: John Wiley; 1947.

148. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester: Horwood; 1983.

149. Molenberghs G, Kenward MG, Aerts M, et al. On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Stat Methods Med Res*. 2014;23(1):11-41. doi:10.1177/0962280212445801

150. Milanzi E, Molenberghs G. Properties of estimators in exponential family settings with observation-based stopping rules. *J Biom Biostat*. 2015;7(1):272. doi:10.4172/2155-6180.1000272

151. Milanzi E, Molenberghs G, Alonso A, et al. Estimation after a group sequential trial. *Stat Biosci*. 2015;7(2):187-205. doi:10.1007/s12561-014-9112-6

152. Thompson JR. Some shrinkage techniques for estimating the mean. *J Am Stat Assoc*. 1968;63(321):113-122. doi:10.1080/01621459.1968.11009226

153. Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics*. 1999;55(1):190-197. doi:10.1111/j.0006-341x.1999.00190.x

154. Shen Y, Cai J. Sample size reestimation for clinical trials with censored survival data. *J Am Stat Assoc*. 2003;98(462):418-426. doi:10.1198/016214503000206

155. Hwang JT. Empirical Bayes estimation for the means of the selected populations. *Sankhya Ind J Stat*. 1993;55(2):285-304.

156. Lindley DV. Discussion of professor Stein's paper "confidence sets for the mean of a multivariate normal distribution". *J R Stat Soc Ser B*. 1962;24:265-296.

157. Rosenkranz GK. Bootstrap corrections of treatment effect estimates following selection. *Comput Stat Data Anal*. 2014;69:220-227. doi:10.1016/j.csda.2013.08.010

158. 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

159. Kunzmann K, Kieser M. Test-compatible confidence intervals for adaptive two-stage single-arm designs with binary endpoint. *Biom J*. 2018;60(1):196-206. doi:10.1002/bimj.201700018

SUPPORTING INFORMATION

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