Statistics in Medicine

WILEY

Statistics in Medicine

REVIEW ARTICLE OPEN ACCESS

Confidence Intervals for Adaptive Trial Designs I: A Methodological Review

David S. Robertson[1] | Thomas Burnett[2] | Babak Choodari-Oskooei[3] | Munya Dimairo[4] | Michael Grayling[5] | Philip Pallmann[6] | Thomas Jaki[1,7]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK | [2]University of Bath, Bath, UK | [3]MRC Clinical Trials Unit at UCL, London, UK | [4]Sheffield Centre for Health and Related Research (SCHARR), University of Sheffield, Sheffield, UK | [5]Statistics and Decision Sciences, Janssen R&D, High Wycombe, UK | [6]Centre for Trials Research, Cardiff University, Cardiff, UK | [7]University of Regensburg, Regensburg, Germany

Correspondence: David S. Robertson (david.robertson@mrc-bsu.cam.ac.uk)

Received: 22 October 2024 | **Revised:** 30 April 2025 | **Accepted:** 6 June 2025

Funding: This work was supported by UK Medical Research Council (MC_UU_00002/14, MC_UU_0040_03, MC_UU_00004_09, MC_UU_12023_29), Health and Care Research Wales, Cancer Research UK, and National Institute for Health and Care Research.

Keywords: adaptive design | bootstrap | coverage | estimation | flexible design | group sequential | interim analyses | repeated analyses

ABSTRACT

Regulatory guidance notes the need for caution in the interpretation of confidence intervals (CIs) constructed during and after an adaptive clinical trial. Conventional CIs of the treatment effects are prone to undercoverage (as well as other undesirable properties) in many adaptive designs (ADs) because they do not take into account the potential and realized trial adaptations. This paper is the first in a two-part series that explores CIs for adaptive trials. It provides a comprehensive review of the methods to construct CIs for ADs, while the second paper illustrates how to implement these in practice and proposes a set of guidelines for trial statisticians. We describe several classes of techniques for constructing CIs for adaptive clinical trials before providing a systematic literature review of available methods, classified by the type of AD. As part of this, we assess, through a proposed traffic light system, which of several desirable features of CIs (such as achieving nominal coverage and consistency with the hypothesis test decision) each of these methods holds.

1 | Introduction

An adaptive design (AD) is a clinical trial design that allows for prospectively planned modifications to one or more aspects of the trial based on accumulating data from participants in the trial [1–3]. These planned modifications vary widely in their intent and scope, but carry a high-level commonality of increasing flexibility and improving efficiency, while maintaining trial integrity and validity. The most common types of AD include those that can select a patient (sub)population (adaptive enrichment designs), modify the randomization ratio to, for example, favor better performing arms (response-adaptive randomization designs), revise the recruitment target based on an updated power calculation (sample size reestimation designs), select promising treatment(s) out of several experimental options (multi-arm multi-stage (MAMS) designs) and terminate the trial early for efficacy or futility (group sequential designs). For a more detailed overview, see, for example, Bretz et al. [4], Pallmann et al. [1], Burnett et al. [5], and the PANDA online resource [6].

Whilst there is now a very large body of literature relating to ADs, the majority has focused on the key question of hypothesis testing, that is, how to enable the inclusion of various types of trial adaptations while maintaining control of decision error rates

(e.g., type I and type II errors). Monitoring the accumulating data within a trial requires adjustment to control the overall type I error rate, for example to account for multiplicity from repeatedly applying statistical testing over time during an AD. A key feature underpinning the methodology for type I error rate control (e.g., in group sequential designs) is the 'independent increment structure', which means that the test statistic calculated using the data gathered in a new stage of a trial (i.e., between two consecutive interim analyses) is statistically independent of the information gathered in all previous stages. This allows a mathematically tractable expression for the joint distribution of the test statistics over time [7] and hence the calculation of stopping boundaries.

The related but distinct issue of the computation of inferential quantities, such as treatment effect estimates and confidence intervals (CIs), has received comparatively less explicit attention. Consequently, to stimulate the field and simultaneously offer practical assistance to researchers, a recent two-part series of articles [8, 9] sought to: (a) review available methods to remove or reduce bias in point estimates following an adaptive trial, (b) explore how to choose and implement an estimator, and (c) guide on how to report estimated effects. Both of these articles on point estimation acknowledged the importance of methods for obtaining CIs following a trial utilizing an AD, but left an in-depth discussion of this topic as out-of-scope. Indeed, whilst point estimates (summary measures of treatment effects) are often the primary focus of a study's final analysis as a core attribute of an estimand [10], capturing uncertainty around such estimates correctly is also essential to aid interpretation. It is CIs that capture this uncertainty by offering an interval that is expected to typically contain the unknown parameter of interest. When choosing the CI to calculate, an important consideration in practice is the desirable properties of a CI procedure. Principally, this relates to the CI having the desired coverage probability (i.e., the long-run probability that the CI contains the true unknown treatment effect of interest). However, it may also include numerous other considerations, including that the CI will indeed be an interval (i.e., it is not disjoint), that narrower CIs are preferred as they are more informative (reflecting less uncertainty), that the CI will always contain an associated point estimate, that it will always be consistent with the decision rule (i.e., with an associated hypothesis test), and that it is computationally feasible to implement.

Put simply, the problem therein for ADs is then that use of standard CI methodology (i.e., CIs constructed using methods that do not account for the fact an AD has been used) for parameters of interest may not result in these desirable properties. Indeed, recent regulatory guidance highlights the important role CIs play, along with the pitfalls of utilizing CI methods developed for conventional fixed sample designs for ADs. Specifically, the U.S. Food and Drug Administration (FDA) notes that "confidence intervals for the primary and secondary endpoints may not have correct coverage probabilities for the true treatment effects" and thus "confidence intervals should be presented with appropriate cautions regarding their interpretation" [11]. It also notes the need to pre-specify methods used to compute CIs at the end of an adaptive trial as also reflected in the Adaptive designs CONSORT Extension (ACE) guidance [3, 12]. The European Medicines Agency (EMA) guidance takes an arguably stronger viewpoint, by stating "methods to … provide confidence intervals with pre-specified coverage probability are required" if an AD is going to be deployed in a regulated setting [13].

Thus the availability of methodology for CI construction, specific to ADs, is of critical importance. In this paper, we refer to such CIs as 'adjusted' CIs; that is, accounting for the adaptive nature of the design when computing CIs. For certain ADs, comparisons of available adjusted CIs have been made—see, for example, comparisons for Simon two-stage trials [14], phase II/III trials [15], and (adaptive) group sequential trials [16, 17]. However, a wider and more up-to-date overview of available methodology is needed to facilitate the use of these methods and to identify key open questions for future research. It is this overview we seek to provide here.

This article proceeds as follows. First, in Section 2 we elaborate on the issue of using 'standard' CI methodology for trials using an AD, discussing the potential problems this can cause, before describing several broad classes of available techniques for computing adjusted CIs suited to ADs in Section 3. Following this, in Section 4 we provide a literature review of available methods for constructing CIs for ADs, including a traffic light system that categorizes which of several desirable features each method holds. We then conclude in Section 5 with a discussion of the current landscape of methods for computing CIs after an adaptive trial.

2 | Potential Problems With Using Standard CIs for ADs

In a traditional (fixed) design for a clinical trial with a single primary outcome, the sample space of that outcome is one-dimensional and so the construction of CIs (just like *p*-values) at the end of the trial is relatively straightforward (at least for commonly encountered continuous outcomes; for discrete outcomes there is additional complexity). This is often achieved by 'inverting' the hypothesis test to obtain desired CI bounds around the maximum likelihood estimate (MLE), as demonstrated more explicitly in Section 3. These 'standard' CIs often produce desired coverage under correct distributional assumptions of the parameter of interest and are consistent with the test decision.

On the contrary, in ADs, several factors may render the use of standard CIs inappropriate and complicate the methods for deriving appropriate CIs for ADs at the end of the trial. First, the possible outcomes at the end of the trial depend on the timing of interim analyses, observed results, and adaptation rule considered. As such, the outcome sample space is no longer one-dimensional but multi-dimensional, as it now includes the stopping stage of the trial (for example).

Second, standard CIs are often constructed based on distributional assumptions that are no longer met in an AD. For example, a selection, enrichment, or stopping rule in an AD may mean that the distribution of standard estimators of the parameter of interest is no longer normal, but the standard CI assumes an underlying normal distribution. This can then lead to a disconnect between the test decision (which does account for the adaptive features of the design such as truncation of the test statistic distribution due to selection) and the standard CI. As a result, a

test decision and the decision derived from the standard CI (i.e., whether it includes the null effect) may differ, which leads to substantial problems in interpretation and communication of the trial results; see the real data example for an adaptive enrichment design in Wassmer and Dragalin [18].

Third, as reflected by Robertson et al. [8, 9], the MLE after an AD is potentially biased in the sense that it will tend to systematically deviate from the true value; and standard CIs tend to be centered around this biased estimate. Finally, the use of standard CIs that do not account for the adaptive nature of the design tends to produce incorrect coverage, often lower than the desired nominal coverage, although higher coverage may also occur. Just like the statistical bias of standard point estimates, the level of incorrect coverage of standard intervals can be impacted by many factors including the magnitude of the underlying treatment effect, trial adaptation considered, decision rules (e.g., stopping boundaries), or the probability of triggering adaptations (e.g., stopping or selection). This incorrect coverage of standard CIs may lead to challenges around the overall interpretation of the study results, as well as their use in secondary research such as evidence synthesis (e.g., systematic reviews and meta-analyses) and even health economic evaluations.

3 | Methodology for Constructing CIs

In this Section, we describe general methodology for constructing CIs and illustrate how these methods can be applied to ADs.

3.1 | Fundamental Concepts and Definitions

We first introduce some fundamental concepts and definitions for CIs in general. Suppose we have a random sample $\boldsymbol{X}$ from a probability distribution with parameter θ, which is the single parameter of interest in the trial (we defer the case for multiple parameters of interest until Section 4.1). A CI for θ with confidence level (or confidence coefficient) $1-\alpha$ is a random interval $(L(X), U(X))$ that has the following claimed property: $P(L(X) < \theta < U(X)) = 1-\alpha$ for all θ. Note that sometimes this is replaced by $P(L(X) < \theta < U(X)) \geq 1-\alpha$ in the literature, particularly for discrete outcomes where it may not be possible to achieve equality for all values of θ.

The coverage probability (often shortened to just 'coverage') of a CI is given by $P(L(X) < \theta < U(X))$. The confidence level in the definition above is the 'nominal' coverage probability. If all assumptions used in deriving a CI are met, this nominal coverage probability will equal the (actual/true) coverage probability (known as 'exact' coverage). However, if these assumptions are not met, such as in the context of many ADs, then the actual coverage may be greater than the nominal coverage probability (known as overcoverage, and the CI is termed a conservative CI) or less than the nominal coverage probability (known as undercoverage, and the CI is termed an anti-conservative CI). While exact coverage is ideal and is targeted, overcoverage is generally more acceptable than undercoverage. Arguably, when a CI has undercoverage it may be more accurately described simply as an 'interval' since it does not have the desired confidence level.

Another important concept for CIs is the distinction between *one-sided* versus *two-sided* CIs, which correspond (at least in theory) with one-sided versus two-sided hypothesis testing. To fix ideas, consider the case where θ takes values on the real line, that is, $\theta \in (-\infty, +\infty)$. A two-sided CI corresponds with a two-sided hypothesis test of $H_0 : \theta = \theta_0$ that is the alternative $H_1 : \theta \neq \theta_0$, and would be of the form (l, u) where l and u both take finite values. In contrast, a one-sided CI corresponds to a one-sided hypothesis test of $H_0 : \theta = \theta_0$ versus the alternative $H_1 : \theta > \theta_0$ (or $H_1 : \theta < \theta_0$), and would be of the form (l, ∞) (or $(-\infty, u)$). In practice, for one-sided hypothesis tests it is common to replace a $1-\alpha$ level one-sided CI (l, ∞) with a $1-2\alpha$ level two-sided CI (l', u) with $l' = l$ and $u < \infty$, or a $1-\alpha$ level one-sided CI $(-\infty, u)$ with a $1-2\alpha$ level two-sided CI (l, u') with $u' = u$ and $l > -\infty$.

3.2 | Inverting a Hypothesis Test Statistic

A widely used technique for constructing CIs is by exploiting the duality between hypothesis tests and CIs. Suppose we are testing a null hypothesis H_0 for a parameter of interest θ with a corresponding level-α hypothesis test procedure. If H_0 is true, a $100(1-\alpha)\%$ CI corresponds to the values of θ for which H_0 is not rejected at level α. This implies [19] that rejecting H_0 whenever such a $100(1-\alpha)\%$ CI does not include the null effect is equivalent to rejecting H_0 whenever the level-α test procedure yields a p-value of less than α. This guarantees that the CI is *consistent* with the hypothesis testing decision (see Section 3.4).

Of note, whilst the $100(1-\alpha)\%$ CI is the set of values of θ for which H_0 would not be rejected, the acceptance region of a level-α test is the set of values of the test statistic for which H_0 would not be rejected. An inversion of the test procedure is achieved by 'mapping' these two sets of values onto each other, thereby using the acceptance region of an existing hypothesis test to identify a set of values for the parameter of interest (i.e., a confidence set) which is consistent with not rejecting H_0. In practice, one would typically 'invert' the equation for the test statistic corresponding to the boundary values, thereby identifying the CI bounds. Where this cannot be done analytically, numerical approaches can be employed to find the CI bounds.

A key example in the context of ADs are repeated CIs (RCIs) used for group sequential designs (as well as MAMS designs). We defer the motivation and general definition of RCIs to Section 4.2.1, and only describe their construction for group sequential designs in terms of inverting a hypothesis test statistic as described in Jennison and Turnbull [20]. Consider a two-sided group sequential test of the hypothesis $H_0 : \theta = \theta_0$ with type I error probability α. This has the form:

$$\text{Reject } H_0 \text{ at stage } k \text{ if } \mid Z_k\left(\theta_0\right) \mid \geq c_k(\alpha), k = 1, \ldots, K$$

where $Z_1, \ldots, Z_K$ are the cumulative standardized test statistics and $c_1(\alpha), \ldots, c_K(\alpha)$ are the group sequential critical values. The RCI at stage k, denoted I_k, is defined by inverting this group sequential test, that is, by defining $I_k = \left\{\theta_0 : |Z_k(\theta_0)| < c_k(\alpha)\right\}$.

Another class of CIs for group sequential designs that are also based on inverting a hypothesis test statistic are 'final' CIs, which can be calculated at the stage the trial stops according to the

pre-specified stopping rules (see also Section 4.2.1). However, rather than working with the group sequential test as described above, such CIs rely on a chosen ordering of the sample space to determine which values of the pair (k_T, Z_T) are more extreme evidence against the null hypothesis, where k_T denotes the stage the trial stops at and Z_T denotes the observed cumulative standardized test statistic. Given the chosen ordering, one can derive an acceptance region and corresponding hypothesis test of $H_0 : \theta = \theta_0$ with type I error probability α. These hypothesis tests can then be inverted to give a CI for θ.

A number of different types of ordering have been proposed in the literature, including the following [20]:

- *Stagewise (or analysis time) ordering*: outcomes are ordered by when the trial stops for efficacy, with earlier stopping for efficacy considered more extreme evidence than later stopping for efficacy, even if the final effect size is smaller.
- *MLE (or sample mean) ordering*: outcomes are ordered by the value of the MLE, which is equivalent to ordering outcomes by the sample mean (or sample mean difference) for many outcome types.
- *Likelihood ratio ordering*: outcomes are ordered by the value of the (standardized) test statistics, which is the ordering induced by the likelihood ratio test statistic [21].
- *Score test* ordering: outcomes are ordered by the value of the score test statistic.

More details of these orderings and an illustrative example are given in Appendix A.1 for the interested reader. Depending on the ordering chosen, the resulting CI can have marked differences in terms of the desirable properties described in Section 4.3. For example, some orderings will give a CI that may not necessarily agree with the original group sequential test. For further discussion on these issues, see Jennison and Turnbull [20, 22]. There, they recommend the use of CIs based on stagewise ordering, because this always produces true intervals (instead of possibly the union of disjoint intervals), is consistent with the decision of the group sequential tests, and is the only method available when the information levels at the interim analyses are unpredictable.

3.3 | Constructing a Pivotal Quantity

Alternatively to inverting the distribution one may instead construct CIs through using a pivotal quantity [19]. Suppose one has some data $\boldsymbol{X} = (X_1, \ldots, X_n)$ and an (unknown) parameter of interest, say θ (note this may be a vector of parameters but for convenience we will consider only a single parameter). Then a pivotal quantity is defined as a function of the data X and the parameter θ, denote this by $g(X, \theta)$, where the distribution of $g(X, \theta)$ does not depend on θ.

Assuming the distribution of $g(X, \theta)$ is known one may find L and U such that

$$P(L < g(X, \theta) < U) = 1 - \alpha.$$

Thus with some manipulation one may find corresponding $\widehat{\theta}_L$ and $\widehat{\theta}_U$ such that

$$P\left(\widehat{\theta}_L < \theta < \widehat{\theta}_U\right) = 1 - \alpha,$$

and thus a $100(1-\alpha)\%$ CI for θ is given by

$$\left[\widehat{\theta}_L, \widehat{\theta}_U\right].$$

To solidify this concept let us consider an example using the normal distribution. Suppose we have independent normally distributed data $\boldsymbol{X} = (X_1, \ldots, X_n)$ with an unknown mean θ and known variance σ^2, that is

$$X_i \sim N(\theta, \sigma^2) \text{ for all } i = 1, \ldots, n$$

To construct a $100(1-\alpha)\%$ CI for θ we shall use a pivotal quantity. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then the pivotal quantity is given by

$$g(X, \theta) = \frac{\overline{X} - \theta}{\sigma/\sqrt{n}}$$

where conveniently we thus know that $g(X, \theta) \sim N(0, 1)$. Defining $\Phi^{-1}(\ldots)$ to be the inverse cumulative distribution function (CDF) of the standard normal distribution we may then find a CI for the pivotal quantity by choosing $L = -\Phi^{-1}(1 - \alpha/2)$ and $U = \Phi^{-1}(1 - \alpha/2)$ satisfying the condition that

$$P(L < g(X, \theta) < U) = 1 - \alpha.$$

With a little work we can thus find the CI,

$$\begin{aligned} 1 - \alpha &= P(L < g(X, \theta) < U) \\ &= P\left(\overline{X} - \Phi^{-1}(1 - \alpha/2)\, \sigma/\sqrt{n} < \theta < \overline{X} + \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}\right) \end{aligned}$$

Thus a $100(1-\alpha)\%$ CI for θ is given by

$$\left[\overline{X} - \Phi^{-1}(1 - \alpha/2)\, \sigma/\sqrt{n}, \overline{X} + \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}\right].$$

It is useful to note the consequence of these well-known results in the context of the normal distribution. It is regularly the case that for the ADs that we consider for this work we wish to provide inference for the sample mean of our data in which case for sufficiently large n, via the central limit theorem, a normal approximation will often suffice. The extension of these methods to ADs depends on the situation in which they are applied but this core concept of finding a pivot corresponding to each estimate for which we require a CI may be applied. This has been applied in several works, see Section 4.4.

As an example of the application of finding an (approximate) pivot for ADs, Woodroofe [23] proposed the following in the context of group sequential designs with parameter of interest θ. Let Z_T and I_T denote the cumulative standardized test statistic and (Fisher) information level, respectively, at the stage the trial stops (denoted T). If the sample sizes were fixed then the statistic $Z'_T(\theta) = \frac{Z_T - \theta I_T}{\sqrt{I_T}}$ would (asymptotically) follow a standard normal distribution. However, due to the potential for early stopping, this is not the case. Instead, the following modification of the statistic $Z'_T(\theta)$ gives a new statistic that more closely follows a standard normal distribution: $Z^*_T(\theta) = \frac{Z'_T(\theta) - \mu(\theta)}{\sigma(\theta)}$ where $\mu(\theta)$ is the mean of $Z'_T(\theta)$ and $\sigma(\theta)$ is the standard deviation of $Z'_T(\theta)$.

3.4 | Bootstrap/Resampling Approaches

The approaches discussed so far in this Section require knowledge of a suitable distribution for the estimate, either directly or asymptotically. Bootstrap and other resampling based methods bypass the need for this by constructing approximate CIs through resampling from the data. To construct a CI for some estimate the bootstrap method will sample from the data (with replacement) repeatedly each giving an estimate of the parameter of interest. These estimates are used to approximate the distribution of the parameter from which we estimate the CI.

To give a little more insight let us consider our previous example from Section 3.3. We have data $\boldsymbol{X} = (X_1, \ldots, X_n)$ and wish to make inference on the mean θ. We may estimate this by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

supposing this has a known distribution (or through use of the other methods in this Section) we might then construct an appropriate $100(1-\alpha)\%$ CI.

Alternatively without making any distributional assumptions we may construct this CI using bootstrap resampling [19]. We draw M bootstrap samples of size n from these data by sampling the data uniformly at random with replacement, denoting these samples by $\boldsymbol{X}^{(i)} = \left(X_1^{(i)}, \ldots, X_n^{(i)}\right)$ for $i = 1, \ldots, M$. From each sample we construct the corresponding estimate

$$\hat{\theta}^{(i)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(i)}.$$

For large n the distribution of these $\widehat{\theta}^{(i)}$ converges to the unknown distribution of $\widehat{\theta}$. Without loss of generality we re-index these bootstrap estimates such that they are in order with

$$\widehat{\theta}^{(1)} \leq \widehat{\theta}^{(2)} \leq \ldots \leq \widehat{\theta}^{(M)}.$$

We then have that

$$P\left(\widehat{\theta}^{(M(\alpha/2))} < \theta < \widehat{\theta}^{(M(1-\alpha/2))}\right) \approx 1-\alpha$$

and thus an approximate $100(1-\alpha)\%$ CI for θ is given by

$$\left[\widehat{\theta}^{(M(\alpha/2))}, \widehat{\theta}^{(M(1-\alpha/2))}\right].$$

This is the so-called 'percentile interval' which is the simplest, but other (potentially better) choices of bootstrap CI exist [24].

This concept can be directly extended to the setting of ADs by using a bootstrap procedure as above while accounting for the adaptive nature of the design. For any given estimator, one may resample from the corresponding data used in constructing the estimate and use this in the construction of an appropriate CI. Such methods can be found in the methodological summary in Section 4.4.

As an example of the use of bootstrap for ADs, consider the context of response-adaptive randomization for a multi-armed trial with K arms and binary endpoints, where the allocation probabilities are adapted based on the accumulated patient response data. Rosenberger and Hu [25] describe how to use a bootstrap procedure in this context:

1. Obtain the observed data from the trial; that is, the vector of observed success proportions $\widehat{P} = \left(\widehat{p}_1, \ldots, \widehat{p}_K\right)$ and sample sizes on each arm $= \left(n_1, \ldots, n_K\right)$.
2. Simulate the response-adaptive allocation rule M times, using $\widehat{P}$ as the assumed true response probabilities.
3. Compute bootstrap estimates of the vector of response probabilities $\hat{P}_1^*, \ldots, \hat{P}_M^*$ and sample sizes $N_1^*, \ldots, N_M^*$ from the simulations.
4. For each $i = 1, \ldots, K$ order $\hat{p}_i^{*1}, \ldots, \hat{p}_i^{*M}$ to form the ordered sequence $\hat{p}_i^{*(1)}, \ldots, \hat{p}_i^{*(M)}$
5. The simplest $100(1-\alpha)\%$ CI for p_i is then $\left(\hat{p}_i^{*(M\alpha/2)}, \hat{p}_i^{*(M(1-\alpha)/2)}\right)$.

3.5 | Hybrid Approaches

Finally, it is also possible to combine the approaches described above to construct CIs for ADs. For example, Chuang and Lai [26, 27] proposed a hybrid technique consisting of elements of both CIs based on ordering the sample space (see the end of Section 3.2) and bootstrap methods following (group) sequential trials. Specifically, they suggested replacing normal quantiles (used in the calculations of the upper and lower confidence bounds) with quantiles from a bootstrap distribution. A different kind of hybrid strategy was discussed by Kimani et al. [15] who compared different CI methods for seamless phase II/III trials in simulations and noted that the properties of CIs could be optimized by combining lower and upper bounds based on different methods, for example, a lower bound based on a method known to have high power (to maximize the chance of rejecting the null hypothesis) with an upper bound based on a method which has exact or close-to-nominal coverage to provide an accurate upper limit for the size of the treatment effect. Such a compound CI would inevitably be asymmetric in most cases.

4 | Methodological Review of Adjusted CIs for ADs

4.1 | Search Strategy and Paper Selection

We conducted a database search of Scopus on 7 April 2025 of all published papers (not including preprints) up to that date. We used a "title, abstract, keywords" search, with the following predefined search term: ((("confidence interval" OR "confidence region" OR "confidence limit" OR "confidence band") AND ("adaptive design" OR "adaptive trial" OR "adaptive clinical trial" OR "group sequential" OR "sequential trial" OR "sequential clinical trial" OR "drop the loser" OR "response adaptive" OR "sample size re-estimation" OR "seamless phase" OR "multi arm multi stage" OR "adaptive enrichment" OR "master protocol" OR "platform trial")).

Our search strategy retrieved a total of 366 papers, of which 218 were excluded as they were not relevant based on the title and abstract. We then looked for additional relevant papers citing or cited by the remaining 147, which added 29 papers. We conducted a full text review of these 176 papers. Information about the trial contexts, methodology used, advantages, limitations, code availability and case studies were extracted for qualitative synthesis. A PRISMA flow chart of this process can be found in Appendix A.2 (Figure A1). Full results giving a summary of each paper can be found in the Data S1.

Before presenting a summary of the results of the literature review in Section 4.4, in Section 4.2 we first briefly describe some key methodological concepts found in the literature review, followed by a discussion of desirable criteria for CIs in Section 4.3.

4.2 | Key Methodological Concepts Found in the Literature Review

4.2.1 | Repeated versus Final CIs

In an AD, CIs are used for end-of-study interpretation of results, that is, at the final analysis of the trial data (hence the term 'final' CI). However, CIs can also be used for monitoring the trial while it is ongoing and providing quantification of uncertainty around treatment effect estimates at *interim* analyses/looks at the accumulating data. One key example of this is when presenting data to an independent data monitoring committee such as the Data and Safety Monitoring Board (DSMB) for deciding whether or not to stop a trial early for efficacy or lack-of-benefit/futility.

This second motivation leads to the concept of *repeated* CIs (RCIs), as already briefly introduced in Section 3.2, which are a sequence of interval estimates for θ that can be calculated at *any* interim look at the trial data. More formally, given a trial with a maximum of K stages (or equivalently $K-1$ interim analyses) the RCIs for a parameter θ are defined [20] as a sequence of intervals $I_k, k=1, \ldots, K$, for which simultaneous coverage probability is maintained at level $1-\alpha$, so that the following equation holds:

$$Pr_\theta\big(\theta \in I_k \text{ for all } k = 1, \ldots, K\big) = 1-\alpha \text{ for all } \theta.$$

Note that in fact if τ is any random stopping time for the trial (i.e., the stage the trial stops at, taking values in $\{1, \ldots, K\}$) then the above equation implies that $Pr_\theta\big(\theta \in I_\tau\big) = 1-\alpha$ for all θ. Hence, RCIs can be used while still maintaining the $1-\alpha$ confidence limit *regardless of how the decision to stop the study was reached*.

Consequently, final CIs (as described at the end of Section 3.2) can only be calculated in a valid way at the stage a trial stops according to a pre-specified stopping rule, whereas RCIs are not tied to a stopping rule and can be calculated at any interim analysis. However, this means that RCIs will typically be wider than final CIs, particularly for earlier interim analyses [28].

Note that in the group sequential literature, sometimes the term 'exact' CI has been used for final CIs as a reference to their exact coverage probability (as opposed to potentially conservative coverage for RCIs). However, 'exact' CIs more generally can refer to being based on the exact distribution of the trial outcomes, for example, for binary data being based on a binomial distribution rather than a normal approximation to the binomial distribution. In our literature review, when we refer to 'exact' CIs we are referring to this latter, more general definition.

4.2.2 | Individual CIs versus Simultaneous Confidence Sets

Thus far, we have focused on the case where there is a single hypothesis and corresponding parameter θ of interest. However, many types of ADs consider multiple hypotheses and corresponding parameters simultaneously. MAMS designs are a key example, with multiple treatment arms being compared against a common control arm. To fix ideas, suppose we have null hypotheses $H_{01}, \ldots, H_{0J}$ with corresponding parameters of interest $\theta_1, \ldots, \theta_J$. Given a random data sample X, suppose we calculate individual two-sided CIs $\big(L_i(X), U_i(X)\big)$ for each parameter independently. These will achieve the correct *individual* (also known as marginal or univariate) coverage, that is, $P\big(\, L_i(X) < \theta_i < U_i(X) \,\big) = 1-\alpha$ for $i = 1, \ldots, J$. However, the *overall* or *simultaneous* coverage probability of the J CIs is not necessarily controlled, that is, $P\big(\, L_1(X) < \theta_1 < U_1(X), \ldots, L_J(X) < \theta_J < U_J(X) \,\big) \neq 1-\alpha$.

To achieve control of the simultaneous coverage probability requires specification of a multivariate *confidence set* or *confidence region* C(X), with the property that $P(\boldsymbol{\theta} \in C(X)) = 1-\alpha$, where $\theta = \big(\theta_1, \ldots, \theta_J\big)$. In general, such confidence sets are not necessarily a cross product of CIs. In order to recover CIs for each parameter, one can enlarge the confidence region to fit within a multidimensional rectangle. This comes with the disadvantage that the resulting CIs may be inconsistent with the test decision (see Section 4.3), as noted by Posch et al. [29].

In a trial that tests multiple hypotheses, when considering constructing CIs for each parameter of interest it is therefore important to decide whether correct individual coverage or simultaneous coverage is desired. This is closely linked to the concept of adjusting for multiplicity, that is, whether one wants to control the marginal type I error rate (known as the pairwise type I error rate in the context of comparing with a common control arm) for each hypothesis at level α or controlling the overall familywise error rate (FWER), which is the probability of making at least one type I error. If there is interest in controlling the FWER, then the CIs should reflect this and hence simultaneous confidence sets/intervals should be considered. Conversely, if no adjustment for multiplicity is required (e.g., when the hypotheses tested are independent) then the usual individual CIs would be sufficient.

4.2.3 | Conditional versus Unconditional CIs

A final key distinction of CIs for ADs is whether they have the correct coverage *conditionally* or *unconditionally*. Intuitively, unconditional coverage refers to the coverage averaged across all possible realizations of an adaptive trial. In contrast, conditional coverage refers to the coverage averaged over a particular subset of trial realizations. For example, we might be interested in the

coverage of the CI conditional on a trial continuing to the final stage, or a particular treatment arm being selected at the final analysis. More precisely, and returning to the setting with a single parameter of interest θ, the conditional coverage of a CI $(L(X), U(X))$ is defined as $P(L(X) < \theta < U(X) \mid S)$, where S is a particular (random) event of interest, such as the stopping stage of a group sequential trial.

A detailed discussion of the merits of conditional versus unconditional inference, including CIs, is beyond the scope of this paper. We refer the interested reader to Strickland and Casella [30], Fan and DeMets [31], Marschner and Schou [32], and Marschner et al. [33] for useful discussion about conditional CIs in the context of group sequential trials. More general discussion and a proposed framework to view about conditional versus unconditional inference for ADs can be found in Marschner [34].

4.3 | Desirable Criteria for CIs

In addition to coverage (see Section 3.1), there have been other proposed desirable criteria for CIs in the literature. The main criteria include:

- Correct coverage (arguably essential),
- Width (all other things being equal, a smaller width is desirable),
- Consistency/compatibility with the hypothesis test (see below),
- Contains the point estimate of interest,
- (Approximate) symmetry around the point estimate of interest,
- Is informative (see below),
- Is in fact an interval (i.e., not a union of disjoint intervals, or the empty set),
- Is computationally feasible/simple to implement.

A CI is *consistent/compatible* with the hypothesis testing decision if it excludes the parameter value(s) that are rejected by the hypothesis test, and conversely includes the parameter value(s) that are *not* rejected by the hypothesis test. If a CI is *not* consistent/compatible with the hypothesis testing decision then this can lead to problems with study interpretation and the communication of results. In theory at least, it is possible to 'invert' the hypothesis test used (see Section 3.2) to obtain a CI that is always consistent with the hypothesis test decision, but not all CIs are constructed in this way as seen in Section 3.

A CI is *informative* if it restricts the possible parameter space. For example, if θ corresponds to the success probability of a binomial distribution then a two-sided CI needs to be strictly contained in [0,1]. More formally, consider a parameter θ that takes values in the set (a,b), where a and b may be infinite. Clearly, it is desirable that for a two-sided CI $(L(X), U(X))$, we have $L(X) > a$ and $U(X) < b$. Similarly, for an upper one-sided CI, it is desirable that $U(X) < b$, while for a lower one-sided CI, it is desirable that $L(X) > a$.

Note there is a stricter definition of a CI being informative for CIs that are compatible with a corresponding hypothesis test. Aside from the criteria above, a CI in this case is informative if it additionally provides more information than the hypothesis testing decision. For example, consider testing the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative $H_1 : \theta > \theta_0$. If H_0 is rejected, then a CI is only informative if $L(X) > \theta_0$.

4.4 | Summary of Literature Review

As part of our summary of the literature review, we use a 'traffic light' system for the different classes of methods for constructing CIs, following the suggested desirable criteria for CIs given above (apart from symmetry since this is closely related to containing the point estimate of interest). For simplicity, in the definitions that follow we consider the setting with a single parameter of interest θ and corresponding method for constructing a two-sided CI denoted $(L(X), U(X))$ given a random data sample X with target coverage (i.e., claimed confidence level) of $1 - \alpha$.

4.4.1 | Coverage

Green: CI has (actual) coverage equal to $1 - \alpha$ for all values of θ, that is, $P(L(X) \leq \theta \leq U(X)) = 1 - \alpha$ for all θ. We add labels 'A' for Analytical and 'S' for Simulation if this property has been shown analytically or only by simulation, respectively.

Amber: CI has (actual) coverage greater than or equal to $1 - \alpha$ for all values of θ, that is, $P(L(X) \leq \theta \leq U(X)) \geq 1 - \alpha$ for all θ. We add labels 'A' for Analytical and 'S' for Simulation as above.

Red: CI has (actual) coverage less than $1 - \alpha$ for at least one value of θ, that is, $P(L(X) \leq \theta \leq U(X)) < 1 - \alpha$ for some θ.

4.4.2 | Interval

Green: For all realizations of X, the method will result in a single CI $(L(X), U(X))$ for all $\alpha \in (0,1)$.

Red: For some realizations of X, the method does *not* result in a single CI $(L(X), U(X))$ for some $\alpha \in (0,1)$.

4.4.3 | Consistent

Green: The CI is always consistent as it is constructed by inverting the hypothesis test used by the AD.

Red: The CI can be inconsistent, that is, there are explicit examples of where the CI and hypothesis test decision for the AD are conflicting.

4.4.4 | Informative

Green: For all realizations of X, the procedure will result in an informative CI for all $\alpha \in (0,1)$.

Red: For some realizations of X, the procedure does *not* result in an informative CI for some $\alpha \in (0,1)$.

4.4.5 | Contains MLE

In Section 4.3, one of the desirable criteria for CIs is that it "contains the point estimate of interest". A detailed discussion of what the point estimate of interest could be for an AD is out of scope of this paper, and we refer the reader to Robertson et al. [8, 9]. For the purposes of the literature review, we simply use the usual end-of-trial MLE as our point estimator of interest, both because this is the most commonly reported estimator and because this has been proposed as a criterion for assessing CIs [20]. Note that in theory, a CI could contain an unbiased or bias-reduced point estimate but not the MLE.

Green: For all realizations of X, the procedure will result in a CI that contains the MLE for all $\alpha \in (0,1)$.

Red: For some realizations of X, the procedure does *not* result in a CI that contains the MLE for some $\alpha \in (0,1)$.

4.4.6 | Computation

We note that for calculating a single CI (i.e., for a single trial realization), there are no computational concerns for almost all methods given modern computational power. However, when it comes to assessing the performance of CIs through simulation studies (e.g., at the planning stage of the trial), computation can still be a (major) limitation.

Green: CI can be calculated using analytical formulae that can easily be evaluated using standard statistical software or code (such as R).

Amber: Calculation of the CI involves the use of numerical optimization and/or computer simulation.

Armed with this traffic light system, we summarize the results of our literature review in Table 1 by classifying the CI methods used for each broad class of AD. For each combination of design class and CI method, we provide a summary of any key features of the CI method reported in the literature, list (key) references in the literature (categorized by outcome type, for example) and also show how the CI method performs as assessed by the traffic light system. Note that sometimes it is unclear whether a property holds for a given CI method, because it has not been adequately explored in the literature so far, which we denote as a '?' in Table 1.

Looking at the summary of the literature review as a whole, the number of papers proposing methods for constructing adjusted CIs for ADs has grown quite rapidly in the past 15 or so years. In terms of the properties described by the traffic light system, in terms of coverage it is generally clear (at least by simulation) which methods result in CIs that have correct coverage as well as over- or undercoverage. We caution however that even methods that are 'green' above rely on the assumptions used to derive the CIs holding exactly, and that methods that are 'red' will have differing levels of undercoverage.

All of the main CI methods above were 'green' in terms of being an interval and being informative, but specific subcases (not covered by the traffic light system) can be 'red', as an example of methods possibly returning the empty set (and hence also being uninformative) see Fan and DeMets [31] and Hartung and Knapp [107] in the context of conditional CIs for group sequential designs. In the context of RCIs for group sequential designs, Brookmeyer & Crowley [133] show how in rare cases the RCI may not be an interval if the information levels depend on the parameter of interest θ, with further discussion on this point in Jennison and Turnbull [20].

In contrast, consistency is harder to assess, with this criterion being unclear for all classes of ADs apart from group sequential designs. This also applies to the criterion of the MLE being included in the CI, reflecting the gap sometimes seen in the literature on point estimation and the literature on CIs for ADs. Finally, all methods (except for the adjusted asymptotic CI for response-adaptive randomization) were 'yellow' in terms of computation, although this encompasses a wider range of computational complexity.

By far the majority of the proposed methodology for CIs has focused on group sequential designs, which is understandable given their long history and widespread use. Other classes of ADs have received comparatively little attention when it comes to adjusted CIs, which is reflected in the unclear properties for some of the CI methods. Finally, most of the methodology has focused on (at least asymptotically) normally-distributed outcomes or binary outcomes, with comparatively few proposals tailored for trials with time-to-event outcomes.

5 | Discussion

In our literature review of the methods for constructing CIs for ADs, we found that there is a growing body of work proposing and evaluating a range of CIs for a variety of ADs. Our hope is that this paper, combined with the annotated bibliography given in the Data S1, provides an easily-accessible and comprehensive resource for trialists and methodologists working on ADs. However, statistical software and code to calculate adjusted CIs unfortunately remains relatively rare (see the annotated bibliography), which is an obstacle to the uptake of methods in practice (see Grayling and Wheeler [134]). In addition, for more complex or novel ADs, adjusted CIs may not currently exist in the literature.

From a methodological perspective, while CIs for group sequential designs are very well-developed, this is much less the case for other classes of ADs. In particular, for response-adaptive randomization and adaptive enrichment designs, only a handful of papers proposing adjusted CIs exist. Another feature is that methods specific for ADs with longitudinal endpoints (including time-to-event endpoints) have received comparatively little attention. The use of such endpoints is challenging even from a hypothesis testing viewpoint, as the independent increments assumption may no longer hold [135]. From an estimation viewpoint, as pointed out by an anonymous reviewer, another complication is that if there is an interaction between follow-up time and treatment effects, the estimand can then be a function of the

TABLE 1 | Summary of properties of different methods of constructing CIs, categorised by the broad class of adaptive design.

Design	Method(s)
Group sequential	***Final confidence intervals*** *Tsiatis et al. (1984)* [35], *Jennison and Turnbull (1999)* [20] **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** A • Colours above assume the use of stage-wise ordering of the sample space and that the canonical joint distribution of the group sequential test statistics holds exactly (e.g., for normally distributed data with a known variance – see Jennison and Turnbull, 1999 [20]) • For an example of where the CI does not contain the MLE, see Tsiatis et al. (1984) [35] • Can fail to be an interval if other orderings of the sample space are used, see e.g., Rosner and Tsiatis (1988) [16], Emerson and Fleming (1990) [36] • Coverage can be conservative when a normal approximation used for binary outcomes, and there can be inconsistency with the test decision; see Lloyd (2021) [37] for a discussion of these features • For non-normal outcomes, a 'hybrid' strategy can be used which combines the CI approach with bootstrap resampling, see Chuang and Lai (1998, 2000) [27, 28], Lai and Li (2006) [38] • Conditional perspective considered in Strickland and Casella (2003) [31], Ohman (1996) [39], Fan and DeMets (2006) [32], Koopmeiners et al. (2012) [40], Marschner and Schou (2019) [33], Marschner et al. (2022) [34] **See also**: *Siegmund (1978)* [41], *Kim and DeMets (1987)* [42], *Rosner and Tsiatis (1988)* [16], *Chang (1989)* [21], *Facey and Whitehead (1990)* [43], *Emerson and Fleming (1990)* [36], *Wittes (2012)* [44], *Hampson et al. (2017)* [45], *Hanscom et al. (2022)* [46] **Binary outcomes**: *Jennison and Turnbull (1983)* [47], *Chang and O'Brien (1986)* [48], *Duffy and Santner (1987)* [49], *Emerson (1995)* [50], *Chang (2004)* [51], *Jung and Kim (2004)* [52], *Dallas (2008)* [53], *Porcher and Desseaux (2012)* [14], *Kirk and Fay (2014)* [54], *Yu et al. (2016)* [55], *Shan (2018)* [56], *Lloyd (2021, 2022)* [37, 57], *Cao and Jung (2024)* [58] **Delayed responses/Overrunning**: *Hall and Liu (2002)* [59], *Hampson and Jennison (2013)* [60], *Zeng et al. (2015)* [61], *Shan (2018)* [62], *Zhao et al. (2015)* [63] **Repeated measures**: *Lee et al. (2002)* [64] ***Repeated confidence intervals*** *Jennison and Turnbull (1984* [65], *1989* [66], *1999* [20]*)* **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** A • Colours above assume that the information levels do not depend on the unknown parameter of interest and that the canonical joint distribution of the group sequential test statistics holds exactly (e.g., for normally distributed data with a known variance – see Jennison and Turnbull, 1999 [20]) • If information levels do depend on the unknown parameter of interest, it may not be an interval – see e.g., Jennison and Turnbull (1999) [20] • RCIs are not tied to any specific stopping rule, so consistency depends on what group sequential test is used to derive the RCI, see e.g., Jennison and Turnbull (1989 [66], with discussion). If the test matches the design, then consistency is guaranteed. • Contains the MLE by construction

(Continues)

TABLE 1 | (Continued)

Design	Method(s)
	See also *Jennison and Turnbull (1990* [29], *1991* [67]*), Davis and Hardy (1992)* [68], *Fleming and DeMets (1993)* [69], *Cook (1994)* [70], *Lee (1995)* [71], *Hu and Lagakos (1999a, b)* [72, 73], *Posch et al. (2008)* [74], *Zhao et al. (2009)* [75], *Zhang et al. (2016)* [76], *Nowak et al. (2022)* [77], *Nelson et al. (2022)* [17] **Time-to-event/Survival outcomes**: *Jennison and Turnbull (1985)* [78], *Williams (1996)* [79], *Bernado and Ibrahim (2000)* [80] **Binary outcomes**: *Lin et al. (1991)* [81], *Coe and Tamhane (1993)* [82] **Repeated measures**: *Wei et al. (1990)* [83], *Jiang (1999)* [84] **Equivalence tests**: *Jennison and Turnbull (1993)* [85]
	Adjusted asymptotic confidence intervals *Woodroofe (1992)* [23], *Todd et al. (1996)* [86] **Coverage**: S; **Interval**: ; **Consistency**: ?; **Informative**: ; **Contains MLE**: ?; **Computation**:  • Colours above are for normally distributed outcomes (with known variance) • Slight undercoverage can occur for certain parameter values, but conservative coverage is also observed in simulations • CI is centred around the median unbiased estimator (and not the MLE) **See also:** **Binary outcomes:** *Todd and Whitehead (1997)* [87], *Coad and Govindarajulu (2000)* [88] **Time-to-event/Survival outcomes:** *Coad and Woodroofe (1996* [89], *1997* [90]*)* **Secondary parameters:** *Whitehead et al. (2000)* [91]
	Bootstrap/resampling procedures *Snapinn (1994)* [92], *Chuang and Lai (1998* [27], *2000* [28]*)* **Coverage**: S; **Interval**: ; **Consistency**: ?; **Informative**: ; **Contains MLE**: ?; **Computation**:  • Can be applied to group sequential trials regardless of stopping boundaries or patient outcome distribution • Undercoverage can occur and can be substantial • Computation involves repeated trial simulations, but for calculating a single confidence interval is not time-consuming • Conditional perspective considered in Pepe et al. (2009) [93], Shimura et al. (2017) [94]

(Continues)

TABLE 1 | (Continued)

Design	Method(s)
	Adaptive group sequential/group sequential with sample size re-estimation **Repeated confidence intervals:** *Lehmacher and Wassmer (1999)* [95], *Wassmer et al. (2001)* [96], *Wassmer (2003)* [97], *Hartung and Knapp (2006)* [98], *Mehta et al. (2007)* [99] **Final confidence intervals:** *Wassmer (2006)* [100], *Brannath et al. (2009)* [101], *Wang et al. (2010)* [102], *Gao et al. (2013)* [103], *Gao and Mehta (2013)* [104], *Mehta et al. (2019)* [105], *Gao and Li (2024)* [106] **See also:** *Hartung and Knapp (2010* [107], *2011* [108]*)*, and the review article by *Nelson et al. (2022)* [17] • The above references are for adaptive group sequential designs (i.e., group sequential designs additionally encompassing sample size re-estimation), but given the variety of different designs this (sub)class includes, we do not give a traffic light categorisation
Multi-arm multi-stage designs (with treatment selection)	***Repeated confidence intervals*** *Zhao et al. (2009)* [75], *Jaki and Magirr (2013)* [109] **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** A • Colours above assume that the canonical joint distribution of the group sequential test statistics holds (at least approximately) • Assumes all promising treatments (for a range of definitions of 'promising') are taken forward after each interim analysis **See also:** **Single stage multi-arm designs:** *Liu (1995)* [110]
	Drop-the-loser designs *Sampson and Sill (2005)* [111], *Stallard and Todd (2005)* [112], *Sill and Sampson (2009)* [113], *Wu et al. (2010)* [114], *Neal et al. (2011)* [115], *Magirr et al. (2013)* [116], *Bowden and Glimm (2014)* [117], *Kimani et al. (2014)* [15], *Carreras et al. (2015)* [118], *Brückner et al. (2017)* [119], *Whitehead et al. (2020)* [120], *Gao and Li (2024)* [106] • The above references are specifically for drop-the-loser designs, but encompass a variety of different methods for constructing CIs, hence we do not give a traffic light categorisation • Some methods e.g., the asymptotic approach of Bowden and Glimm (2014) [117] can have undercoverage **See also:** **Methods using a normal approximation:** *Shun et al. (2007)* [121], *Bowden and Glimm (2008)* [122] **Review article for seamless phase II/III trials:** *Kimani et al. (2014)* [15]

(Continues)

TABLE 1 | (Continued)

Design	Method(s)
Response-adaptive randomisation (RAR)	***Bootstrap/resampling procedures*** *Rosenberger and Hu (1999)* [25], *Bandyopadhyay and Biswas (2003)* [123], *Baldi Antognini et al. (2022)* [124], *Lane (2022)* [125] **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** S | ? | ? • Can be applied to multi-arm trials as well as different outcome types • Slight undercoverage can occur for some parameter values, but conservative coverage is commonly observed in simulation studies • Computation involves repeated trial simulations, but for calculating a single confidence interval it is not time-consuming • Both conditional and unconditional bootstrap procedures have been proposed, see *Lane (2022)* [125]
	Final confidence intervals *Wei et al. (1990)* [126] **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** S | ? | ? • Colours above are for a two-arm trial with binary outcomes • The specific RAR procedure considered in the simulations is the randomised play-the-winner (RPW) rule • Undercoverage can occur for large treatment differences, otherwise conservative coverage is observed • Computational burden depends on the RAR procedure used
	Adjusted asymptotic confidence intervals *Tolusso and Wang (2011)* [127] **Coverage** | **Interval** | **Consistency** | **Informative** | **Contains MLE** | **Computation** S | ? | ? • Colours above are for a two-arm trial with binary outcomes • The specific RAR procedure considered in the simulations is the randomised RPW rule • Undercoverage can occur for large treatment differences, otherwise conservative coverage is observed • Simple closed form expression given for the RPW rule, but in general computational burden depends on the RAR procedure used
Adaptive enrichment designs	*Brannath et al. (2009)* [128], *Rosenblum (2013)* [129], *Wu et al. (2014)* [130], *Wassmer and Dragalin (2015)* [18], *Kimani et al. (2020)* [131], *Ishii et al. (2025)* [132] • The above references encompass a variety of different methods for constructing CIs, hence we do not give a traffic light categorisation • These methods in general can be very computationally intensive

timing of the interim analyses. This can be avoided by defining the primary estimand in such a way to account for the possible interaction, for example, by using the restricted mean survival time (RMST) which does not make a proportional hazards assumption [136].

More generally, while coverage is always reported in simulation studies for CIs, the other desirable properties and related performance measures described in Section 4.3 are much less often reported (as also reflected in the traffic light system results in Table 1). Hence, we encourage methodologists working on new CI methods to consider reporting a wider variety of performance measures. In the future, it would also be helpful to have methodological proposals around how to appropriately combine different metrics/performance measures of interest. Related to this, some proposed desirable properties of CIs are more contentious than others. For example, the property of containing the MLE (and the related symmetry property) may not necessarily be a good one to assess, as conceivably an adjusted CI might contain an appropriately bias-adjusted estimate but not the MLE.

We note that our focus in this paper has exclusively been on frequentist CIs. We have not discussed the construction of other types of intervals, for example, prediction intervals and tolerance intervals (see e.g., Vardeman [137] and Krishnamoorthy & Mathew [138]). Nor have we addressed fixed-width CI construction (i.e., where the design of the trial itself is chosen to achieve a CI with a desired coverage and of a certain width, through a minimal sample size), all of which have had very limited discussion in the context of ADs.

With the growing popularity of Bayesian methods for ADs (and clinical trials more generally) there is growing interest in using the Bayesian paradigm for inference about the treatment effect. However, CIs are inherently a frequentist concept involving the repeated resampling or realizations of the adaptive trial in question. In contrast, from a Bayesian perspective the uncertainty around the parameter of interest, θ, is quantified in terms of a credible interval, based on the posterior probability density of θ itself. Nonetheless, the frequentist properties of credible intervals [137] (such as coverage) could in theory be investigated, although this was out of scope of our literature review and our paper more generally. A complicating factor is that the properties of credible intervals would additionally depend on the choice of prior distribution. An interesting approach that combines aspects of both Bayesian and frequentist methods is the confidence *distribution* approach as described by Marschner [139]. These confidence distributions provide a posterior-like probability distribution that does not require the specification of priors, and is compatible for frequentist inference.

In part II of this paper series [140], we explore the practical considerations surrounding the use of CIs for ADs. There, we illustrate their application to a two-stage group sequential trial design. We also provide a set of guidelines for best practice, considering the use of CIs in ADs from the design stage through to the final reporting of results.

Acknowledgments

T Jaki and DS Robertson received funding from the UK Medical Research Council (MC_UU_00002/14 and MC_UU_0040_03). B Choodari-Oskooei was supported by the MRC grant (MC_UU_00004_09 and MC_UU_12023_29). The Centre for Trials Research receives infrastructure funding from Health and Care Research Wales and Cancer Research UK. The Sheffield Clinical Trials Research Unit (CTRU) within SCHARR received funding from the National Institute for Health and Care Research (NIHR).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All of the data that support the findings of this study are available within the paper and Data S1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

1. P. Pallmann, A. W. Bedding, B. Choodari-Oskooei, et al., "Adaptive Designs in Clinical Trials: Why Use Them, and How to Run and Report Them," *BMC Medicine* 16, no. 1 (2018): 29, https://doi.org/10.1186/s12916-018-1017-7.

2. M. Dimairo, E. Coates, P. Pallmann, et al., "Development Process of a Consensus-Driven CONSORT Extension for Randomised Trials Using an Adaptive Design," *BMC Medicine* 16, no. 1 (2018): 210, https://doi.org/10.1186/s12916-018-1196-2.

3. M. Dimairo, P. Pallmann, J. Wason, 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* 369 (2020): m115, https://doi.org/10.1136/bmj.m115.

4. F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch, "Adaptive Designs for Confirmatory Clinical Trials," *Statistics in Medicine* 28, no. 8 (2009): 1181–1217, https://doi.org/10.1002/sim.3538.

5. T. Burnett, P. Mozgunov, P. Pallmann, S. S. Villar, G. M. Wheeler, and T. Jaki, "Adding Flexibility to Clinical Trial Designs: An Example-Based Guide to the Practical Use of Adaptive Designs," *BMC Medicine* 18, no. 1 (2020): 352, https://doi.org/10.1186/s12916-020-01808-2.

6. M. Dimairo, P. Pallmann, T. Jaki, et al., "PANDA: A Practical Adaptive and Novel Designs and Analysis Toolkit. University of Sheffield." accessed October 21, 2024. https://panda.shef.ac.uk/.

7. P. Armitage, C. K. McPherson, and B. C. Rowe, "Repeated Significance Tests on Accumulating Data," *Journal of the Royal Statistical Society. Series A, General* 132, no. 2 (1969): 235, https://doi.org/10.2307/2343787.

8. D. S. Robertson, B. Choodari-Oskooei, M. Dimairo, L. Flight, P. Pallmann, and T. Jaki, "Point Estimation for Adaptive Trial Designs I: A Methodological Review," *Statistics in Medicine* 42, no. 2 (2023): 122–145, https://doi.org/10.1002/sim.9605.

9. D. S. Robertson, B. Choodari-Oskooei, M. Dimairo, L. Flight, P. Pallmann, and T. Jaki, "Point Estimation for Adaptive Trial Designs II: Practical Considerations and Guidance," *Statistics in Medicine* 42, no. 14 (2023): 2496–2520, https://doi.org/10.1002/sim.9734.

10. B. C. Kahan, J. Hindley, M. Edwards, S. Cro, and T. P. Morris, "The Estimands Framework: A Primer on the ICH E9(R1) Addendum," *BMJ* 384 (2024): e076316, https://doi.org/10.1136/bmj-2023-076316.

11. Center for Drug Evaluation and Research, "Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry." 2020, accessed October 21, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

12. M. Dimairo, P. Pallmann, J. Wason, 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* 21, no. 1 (2020): 528, https://doi.org/10.1186/s13063-020-04334-x.

13. European Medicines Agency, "Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned With An Adaptive Design." 2007, accessed October 21, 2024. https://www.emwa.org/Documents/Freelancer/adaptivestudydesign/EMA%20adaptive%20design_Oct%202007.pdf.

14. R. Porcher and K. Desseaux, "What Inference for Two-Stage Phase II Trials?," *BMC Medical Research Methodology* 12, no. 1 (2012): 117, https://doi.org/10.1186/1471-2288-12-117.

15. P. K. Kimani, S. Todd, and N. Stallard, "A Comparison of Methods for Constructing Confidence Intervals After Phase II/III Clinical Trials," *Biometrical Journal* 56, no. 1 (2014): 107–128, https://doi.org/10.1002/bimj.201300036.

16. G. L. Rosner and A. A. Tsiatis, "Exact Confidence Intervals Following a Group Sequential Trial: A Comparison of Methods," *Biometrika* 75, no. 4 (1988): 723–729, https://doi.org/10.1093/biomet/75.4.723.

17. B. S. Nelson, L. Liu, and C. Mehta, "A Simulation-Based Comparison of Estimation Methods for Adaptive and Classical Group Sequential Clinical Trials," *Pharmaceutical Statistics* 21, no. 3 (2022): 599–611, https://doi.org/10.1002/pst.2188.

18. G. Wassmer and V. Dragalin, "Designing Issues in Confirmatory Adaptive Population Enrichment Trials," *Journal of Biopharmaceutical Statistics* 25, no. 4 (2015): 651–669, https://doi.org/10.1080/10543406.2014.920869.

19. G. Casella and R. Berger, *Statistical Inference* (CRC Press, 2024).

20. C. Jennison and B. W. Turnbull, *Group Sequential Methods With Applications to Clinical Trials* (CRC Press, 1999).

21. M. N. Chang, "Confidence Intervals for a Normal Mean Following a Group Sequential Test," *Biometrics* 45, no. 1 (1989): 247–254, https://doi.org/10.2307/2532050.

22. M. A. Proschan, K. K. G. Lan, and J. T. Wittes, "Inference Following a Group-Sequential Trial," in *Statistical Monitoring of Clinical Trials: A Unified Approach* (Springer, 2006), 113–136, https://doi.org/10.1007/978-0-387-44970-8_7.

23. M. Woodroofe, "Estimation After Sequential Testing: A Simple Approach for a Truncated Sequential Probability Ratio Test," *Biometrika* 79, no. 2 (1992): 347–353, https://doi.org/10.2307/2336845.

24. T. J. DiCiccio and B. Efron, "Bootstrap Confidence Intervals," *Statistical Science* 11, no. 3 (1996): 189–228, https://doi.org/10.1214/ss/1032280214.

25. W. F. Rosenberger and F. Hu, "Bootstrap Methods for Adaptive Designs," *Statistics in Medicine* 18, no. 14 (1999): 1757–1767, https://doi.org/10.1002/(SICI)1097-0258(19990730)18:14<1757::AID-SIM212>3.0.CO;2-R.

26. C. S. Chuang and T. L. Lai, "Resampling Methods for Confidence Intervals in Group Sequential Trials," *Biometrika* 85, no. 2 (1998): 317–332, https://doi.org/10.1093/biomet/85.2.317.

27. C. S. Chuang and T. L. Lai, "Hybrid Resampling Methods for Confidence Intervals," *Statistica Sinica* 10 (2000): 1–50.

28. C. Jennison and B. W. Turnbull, "Statistical Approaches to Interim Monitoring of Medical Trials: A Review and Commentary," *Statistical Science* 5, no. 3 (1990): 299–317, https://doi.org/10.1214/ss/1177012099.

29. M. Posch, F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer, "Testing and Estimation in Flexible Group Sequential Designs With Adaptive Treatment Selection," *Statistics in Medicine* 24, no. 24 (2005): 3697–3714, https://doi.org/10.1002/sim.2389.

30. P. A. Ohman Strickland and G. Casella, "Conditional Inference Following Group Sequential Testing," *Biometrical Journal* 45, no. 5 (2003): 515–526, https://doi.org/10.1002/bimj.200390029.

31. X. Fan and D. L. Demets, "Conditional and Unconditional Confidence Intervals Following a Group Sequential Test," *Journal of Biopharmaceutical Statistics* 16, no. 1 (2006): 107–122, https://doi.org/10.1080/10543400500406595.

32. I. C. Marschner and I. M. Schou, "Underestimation of Treatment Effects in Sequentially Monitored Clinical Trials That Did Not Stop Early for Benefit," *Statistical Methods in Medical Research* 28, no. 10–11 (2019): 3027–3041, https://doi.org/10.1177/0962280218795320.

33. I. C. Marschner, M. Schou, and A. J. Martin, "Estimation of the Treatment Effect Following a Clinical Trial That Stopped Early for Benefit," *Statistical Methods in Medical Research* 31, no. 12 (2022): 2456–2469, https://doi.org/10.1177/09622802221122445.

34. I. C. Marschner, "A General Framework for the Analysis of Adaptive Experiments," *Statistical Science* 36, no. 3 (2021): 465–492, https://doi.org/10.1214/20-STS803.

35. A. A. Tsiatis, G. L. Rosner, and C. R. Mehta, "Exact Confidence Intervals Following a Group Sequential Test," *Biometrics* 40, no. 3 (1984): 797–803, https://doi.org/10.2307/2530924.

36. S. S. Emerson and T. R. Fleming, "Parameter Estimation Following Group Sequential Hypothesis Testing," *Biometrika* 77, no. 4 (1990): 875–892, https://doi.org/10.1093/biomet/77.4.875.

37. C. J. Lloyd, "Exact Confidence Limits After a Group Sequential Single Arm Binary Trial," *Statistics in Medicine* 40, no. 10 (2021): 2389–2399, https://doi.org/10.1002/sim.8909.

38. T. L. Lai and W. Li, "Confidence Intervals in Group Sequential Trials With Random Group Sizes and Applications to Survival Analysis," *Biometrika* 93, no. 3 (2006): 641–654, https://doi.org/10.1093/biomet/93.3.641.

39. P. A. Ohman and G. Casella, "Conditional Inference Following Group Sequential Testing. Cornell University." 1996, https://ecommons.cornell.edu/server/api/core/bitstreams/e049d6c7-f4b8-4f64-8a40-2e33b0c253b6/content.

40. J. S. Koopmeiners, Z. Feng, and M. S. Pepe, "Conditional Estimation After a Two-Stage Diagnostic Biomarker Study That Allows Early Termination for Futility," *Statistics in Medicine* 31, no. 5 (2012): 420–435, https://doi.org/10.1002/sim.4430.

41. D. Siegmund, "Estimation Following Sequential Tests," *Biometrika* 65, no. 2 (1978): 341–349, https://doi.org/10.2307/2335213.

42. K. Kim and D. L. DeMets, "Confidence Intervals Following Group Sequential Tests in Clinical Trials," *Biometrics* 43, no. 4 (1987): 857–864, https://doi.org/10.2307/2531539.

43. K. M. Facey and J. Whitehead, "An Improved Approximation for Calculation of Confidence Intervals After a Sequential Clinical Trial," *Statistics in Medicine* 9, no. 11 (1990): 1277–1285, https://doi.org/10.1002/sim.4780091107.

44. J. Wittes, "Stopping a Trial Early - and Then What?," *Clinical Trials* 9, no. 6 (2012): 714–720, https://doi.org/10.1177/1740774512454600.

45. L. V. Hampson, R. Fisch, L. M. Van, and T. Jaki, "Asymmetric Inner Wedge Group Sequential Tests With Applications to Verifying Whether Effective Drug Concentrations Are Similar in Adults and Children,"

Statistics in Medicine 36, no. 3 (2017): 426–441, https://doi.org/10.1002/sim.7154.

46. B. S. Hanscom, D. J. Donnell, T. R. Fleming, et al., "Evaluating Group-Sequential Non-Inferiority Clinical Trials Following Interim Stopping: The HIV Prevention Trials Network 083 Trial," *Clinical Trials* 19, no. 6 (2022): 605–612, https://doi.org/10.1177/17407745221118371.

47. C. Jennison and B. W. Turnbull, "Confidence Intervals for a Binomial Parameter Following a Multistage Test With Application to MIL-STD 105D and Medical Trials," *Technometrics* 25, no. 1 (1983): 49–58, https://doi.org/10.2307/1267726.

48. M. N. Chang and P. C. O'Brien, "Confidence Intervals Following Group Sequential Tests," *Controlled Clinical Trials* 7, no. 1 (1986): 18–26, https://doi.org/10.1016/0197-2456(86)90004-8.

49. D. E. Duffy and T. J. Santner, "Confidence Intervals for a Binomial Parameter Based on Multistage Tests," *Biometrics* 43, no. 1 (1987): 81–93, https://doi.org/10.2307/2531951.

50. S. S. Emerson, "Stopping a Clinical Trial Very Early Based on Unplanned Interim Analyses: A Group Sequential Approach," *Biometrics* 51, no. 3 (1995): 1152–1162, https://doi.org/10.2307/2533015.

51. M. N. Chang, "Improved Confidence Intervals for a Binomial Parameter Following a Group Sequential Phase II Clinical Trial," *Statistics in Medicine* 23, no. 18 (2004): 2817–2826, https://doi.org/10.1002/sim.1878.

52. S. H. Jung and K. M. Kim, "On the Estimation of the Binomial Probability in Multistage Clinical Trials," *Statistics in Medicine* 23, no. 6 (2004): 881–896, https://doi.org/10.1002/sim.1653.

53. M. J. Dallas, "Accounting for Interim Safety Monitoring of an Adverse Event Upon Termination of a Clinical Trial," *Journal of Biopharmaceutical Statistics* 18, no. 4 (2008): 631–638, https://doi.org/10.1080/10543400802071311.

54. J. L. Kirk and M. P. Fay, "An Introduction to Practical Sequential Inferences via Single-Arm Binary Response Studies Using the Binseqtest R Package," *American Statistician* 68, no. 4 (2014): 230–242, https://doi.org/10.1080/00031305.2014.951126.

55. J. Yu, A. D. Hutson, A. H. Siddiqui, and M. A. Kedron, "Group Sequential Control of Overall Toxicity Incidents in Clinical Trials – Non-Bayesian and Bayesian Approaches," *Statistical Methods in Medical Research* 25, no. 1 (2016): 64–80, https://doi.org/10.1177/0962280212440535.

56. G. Shan, "Exact Confidence Limits for the Probability of Response in Two-Stage Designs," *Statistics* 52, no. 5 (2018): 1086–1095, https://doi.org/10.1080/02331888.2018.1469023.

57. C. J. Lloyd, "Exact Confidence Limits Compatible With the Result of a Sequential Trial," *Journal of Statistical Planning and Inference* 217 (2022): 171–176, https://doi.org/10.1016/j.jspi.2021.07.014.

58. S. Cao and S. H. Jung, "Confidence Intervals for Odds Ratio From Multistage Randomized Phase II Trials," *Statistics in Medicine* 43, no. 12 (2024): 2359–2367, https://doi.org/10.1002/sim.10073.

59. W. J. Hall and A. Liu, "Sequential Tests and Estimators After Overrunning Based on Maximum-Likelihood Ordering," *Biometrika* 89, no. 3 (2002): 699–708, https://doi.org/10.1093/biomet/89.3.699.

60. L. V. Hampson and C. Jennison, "Group Sequential Tests for Delayed Responses (With Discussion)," *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 75, no. 1 (2013): 3–54, https://doi.org/10.1111/j.1467-9868.2012.01030.x.

61. D. Zeng, F. Gao, K. Hu, C. Jia, and J. G. Ibrahim, "Hypothesis Testing for Two-Stage Designs With Over or Under Enrollment," *Statistics in Medicine* 34, no. 16 (2015): 2417–2426, https://doi.org/10.1002/sim.6490.

62. G. Shan, "Exact Confidence Limits for the Response Rate in Two-Stage Designs With Over- Or Under-Enrollment in the Second Stage," *Statistical Methods in Medical Research* 27, no. 4 (2018): 1045–1055, https://doi.org/10.1177/0962280216650918.

63. J. Zhao, M. Yu, and X. P. Feng, "Statistical Inference for Extended or Shortened Phase II Studies Based on Simon's Two-Stage Designs," *BMC Medical Research Methodology* 15, no. 1 (2015): 48, https://doi.org/10.1186/s12874-015-0039-5.

64. J. W. Lee, S. J. Jo, D. L. DeMets, and K. Kim, "Confidence Intervals Following Group Sequential Tests in Clinical Trials With Multivariate Observations," *Journal of Statistical Computation and Simulation* 72, no. 3 (2002): 247–259, https://doi.org/10.1080/00949650212386.

65. C. Jennison and B. W. Turnbull, "Repeated Confidence Intervals for Group Sequential Clinical Trials," *Controlled Clinical Trials* 5, no. 1 (1984): 33–45, https://doi.org/10.1016/0197-2456(84)90148-X.

66. C. Jennison and B. W. Turnbull, "Interim Analyses: The Repeated Confidence Interval Approach," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 51, no. 3 (1989): 305–361.

67. C. Jennison and B. W. Turnbull, "Exact Calculations for Sequential t, X2 and F Tests," *Biometrika* 78, no. 1 (1991): 133–141, https://doi.org/10.1093/biomet/78.1.133.

68. B. R. Davis and R. J. Hardy, "Repeated Confidence Intervals and Prediction Intervals Using Stochastic Curtailment," *Communications in Statistics - Theory and Methods* 21, no. 2 (1992): 351–368, https://doi.org/10.1080/03610929208830783.

69. T. R. Fleming and D. L. DeMets, "Monitoring of Clinical Trials: Issues and Recommendations," *Controlled Clinical Trials* 14, no. 3 (1993): 183–197, https://doi.org/10.1016/0197-2456(93)90002-U.

70. R. J. Cook, "Interim Monitoring of Bivariate Responses Using Repeated Confidence Intervals," *Controlled Clinical Trials* 15, no. 3 (1994): 187–200, https://doi.org/10.1016/0197-2456(94)90056-6.

71. S. J. Lee, "Group Sequential Monitoring of Clinical Trials With Multivariate Outcomes," *Drug Information Journal* 29, no. 1_suppl (1995): 1563S–1582S, https://doi.org/10.1177/00928615950290S106.

72. X. J. Hu and S. W. Lagakos, "Group Sequential Analyses for the Mean Function of a Repeated Measure Process," *Statistics in Medicine* 18, no. 17–18 (1999): 2287–2299, https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2287::AID-SIM255>3.0.CO;2-D.

73. X. Hu and S. Lagakos, "Interim Analyses Using Repeated Confidence Bands," *Biometrika* 86, no. 3 (1999): 517–529, https://doi.org/10.1093/biomet/86.3.517.

74. M. Posch, G. Wassmer, and W. Brannath, "A Note on Repeated *p*-values for Group Sequential Designs," *Biometrika* 95, no. 1 (2008): 253–256, https://doi.org/10.1093/biomet/asm080.

75. L. Zhao, X. J. Hu, and S. W. Lagakos, "Statistical Monitoring of Clinical Trials With Multivariate Response and/or Multiple Arms: A Flexible Approach," *Biostatistics* 10, no. 2 (2009): 310–323, https://doi.org/10.1093/biostatistics/kxn037.

76. Q. Zhang, D. Lai, and B. R. Davis, "Repeated Confidence Intervals and Prediction Intervals Using Stochastic Curtailment Under Fractional Brownian Motion," *Communications in Statistics - Theory and Methods* 45, no. 14 (2016): 4295–4306, https://doi.org/10.1080/03610926.2014.919400.

77. C. P. Nowak, T. Mütze, and F. Konietschke, "Group Sequential Methods for the Mann-Whitney Parameter," *Statistical Methods in Medical Research* 31, no. 10 (2022): 2004–2020, https://doi.org/10.1177/09622802221107103.

78. C. Jennison and B. W. Turnbull, "Repeated Confidence Intervals for the Median Survival Time," *Biometrika* 72, no. 3 (1985): 619–625, https://doi.org/10.1093/biomet/72.3.619.

79. P. L. Williams, "Sequential Monitoring of Clinical Trials With Multiple Survival Endpoints," *Statistics in Medicine* 15, no. 21 (1996): 2341–2357, https://doi.org/10.1002/(SICI)1097-0258(19961115)15:21<2341::AID-SIM453>3.0.CO;2-N.

80. M. V. Patricia Bernardo and J. G. Ibrahim, "Group Sequential Designs for Cure Rate Models With Early Stopping in Favour of the Null Hypothesis," *Statistics in Medicine* 19, no. 22 (2000): 3023–3035, https://doi.org/10.1002/1097-0258(20001130)19:22<3023::AID-SIM638>3.0.CO;2-X.

81. D. Y. Lin, L. J. Wei, and D. L. DeMets, "Exact Statistical Inference for Group Sequential Trials," *Biometrics* 47, no. 4 (1991): 1399–1408, https://doi.org/10.2307/2532394.

82. P. R. Coe and A. C. Tamhane, "Exact Repeated Confidence Intervals for Bernoulli Parameters in a Group Sequential Clinical Trial," *Controlled Clinical Trials* 14, no. 1 (1993): 19–29, https://doi.org/10.1016/0197-2456(93)90047-H.

83. L. J. Wei, J. Q. Su, and J. M. Lachin, "Interim Analyses With Repeated Measurements in a Sequential Clinical Trial," *Biometrika* 77, no. 2 (1990): 359–364, https://doi.org/10.1093/biomet/77.2.359.

84. W. Jiang, "Group Sequential Procedures for Repeated Events Data With Frailty," *Journal of Biopharmaceutical Statistics* 9, no. 3 (1999): 379–399, https://doi.org/10.1081/BIP-100101183.

85. C. Jennison and B. W. Turnbull, "Sequential Equivalence Testing and Repeated Confidence Intervals, With Applications to Normal and Binary Responses," *Biometrics* 49, no. 1 (1993): 31–43, https://doi.org/10.2307/2532600.

86. S. Todd, "Point and Interval Estimation Following a Sequential Clinical Trial," *Biometrika* 83, no. 2 (1996): 453–461, https://doi.org/10.1093/biomet/83.2.453.

87. S. Todd and J. Whitehead, "Confidence Interval Calculation for a Sequential Clinical Trial of Binary Responses," *Biometrika* 84, no. 3 (1997): 737–743, https://doi.org/10.1093/biomet/84.3.737.

88. D. S. Coad and Z. Govindarajulu, "Corrected Confidence Intervals Following a Sequential Adaptive Clinical Trial With Binary Responses," *Journal of Statistical Planning and Inference* 91, no. 1 (2000): 53–64, https://doi.org/10.1016/S0378-3758(00)00129-4.

89. D. S. Coad and M. B. Woodroofe, "Corrected Confidence Intervals After Sequential Testing With Applications to Survival Analysis," *Biometrika* 83, no. 4 (1996): 763–777, https://doi.org/10.1093/biomet/83.4.763.

90. D. S. Coad and M. B. Woodroofe, "Approximate Confidence Intervals After a Sequential Clinical Trial Comparing Two Exponential Survival Curves With Censoring," *Journal of Statistical Planning and Inference* 63, no. 1 (1997): 79–96, https://doi.org/10.1016/S0378-3758(96)00204-2.

91. J. Whitehead, S. Todd, and W. J. Hall, "Confidence Intervals for Secondary Parameters Following a Sequential Test," *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 62, no. 4 (2000): 731–745, https://doi.org/10.1111/1467-9868.00260.

92. S. M. Snapinn, "Use of the Randomization Distribution in the Analysis of Sequential Clinical Trials," *Communications in Statistics - Theory and Methods* 23, no. 2 (1994): 485–498, https://doi.org/10.1080/03610929408831268.

93. M. S. Pepe, Z. Feng, G. Longton, and J. Koopmeiners, "Conditional Estimation of Sensitivity and Specificity From a Phase 2 Biomarker Study Allowing Early Termination for Futility," *Statistics in Medicine* 28, no. 5 (2009): 762–779, https://doi.org/10.1002/sim.3506.

94. M. Shimura, M. Gosho, and A. Hirakawa, "Comparison of Conditional Bias-Adjusted Estimators for Interim Analysis in Clinical Trials With Survival Data," *Statistics in Medicine* 36, no. 13 (2017): 2067–2080, https://doi.org/10.1002/sim.7258.

95. W. Lehmacher and G. Wassmer, "Adaptive Sample Size Calculations in Group Sequential Trials," *Biometrics* 55, no. 4 (1999): 1286–1290, https://doi.org/10.1111/j.0006-341X.1999.01286.x.

96. G. Wassmer, R. Eisebitt, and S. Coburger, "Flexible Interim Analyses in Clinical Trials Using Multistage Adaptive Test Designs," *Drug Information Journal* 35, no. 4 (2001): 1131–1146, https://doi.org/10.1177/009286150103500410.

97. G. Wassmer, "Data-Driven Analysis Strategies for Proportion Studies in Adaptive Group Sequential Test Designs," *Journal of Biopharmaceutical Statistics* 13, no. 4 (2003): 585–603, https://doi.org/10.1081/BIP-120024196.

98. J. Hartung and G. Knapp, "Repeated Confidence Intervals in Self-Designing Clinical Trials and Switching Between Noninferiority and Superiority," *Biometrical Journal* 48, no. 4 (2006): 697–709, https://doi.org/10.1002/bimj.200510213.

99. C. R. Mehta, P. Bauer, M. Posch, and W. Brannath, "Repeated Confidence Intervals for Adaptive Group Sequential Trials," *Statistics in Medicine* 26, no. 30 (2007): 5422–5433, https://doi.org/10.1002/sim.3062.

100. G. Wassmer, "Planning and Analyzing Adaptive Group Sequential Survival Trials," *Biometrical Journal* 48, no. 4 (2006): 714–729, https://doi.org/10.1002/bimj.200510190.

101. W. Brannath, C. R. Mehta, and M. Posch, "Exact Confidence Bounds Following Adaptive Group Sequential Tests," *Biometrics* 65, no. 2 (2009): 539–546, https://doi.org/10.1111/j.1541-0420.2008.01101.x.

102. Y. Wang, G. Li, and W. J. Shih, "Estimation and Confidence Intervals for Two-Stage Sample-Size-Flexible Design With LSW Likelihood Approach," *Statistics in Biosciences* 2, no. 2 (2010): 180–190, https://doi.org/10.1007/s12561-010-9023-0.

103. P. Gao, L. Liu, and C. Mehta, "Adaptive Designs for Noninferiority Trials," *Biometrical Journal* 55, no. 3 (2013): 310–321, https://doi.org/10.1002/bimj.201200034.

104. P. Gao, L. Liu, and C. Mehta, "Exact Inference for Adaptive Group Sequential Designs," *Statistics in Medicine* 32, no. 23 (2013): 3991–4005, https://doi.org/10.1002/sim.5847.

105. C. Mehta, L. Liu, P. Ghosh, and P. Gao, "Exact Inference for Adaptive Group Sequential Designs," in *Pharmaceutical Statistics*, ed. R. Liu and Y. Tsong (Springer International Publishing, 2019), 131–139, https://doi.org/10.1007/978-3-319-67386-8_10.

106. P. Gao and Y. Li, "Adaptive Multiple Comparison Sequential Design (AMCSD) for Clinical Trials," *Journal of Biopharmaceutical Statistics* 34, no. 3 (2024): 424–440, https://doi.org/10.1080/10543406.2023.2233590.

107. J. Hartung and G. Knapp, "Adaptive Group Sequential Confidence Intervals for the Ratio of Normal Means," *Sankhya B* 72, no. 1 (2010): 76–95, https://doi.org/10.1007/s13571-010-0005-5.

108. J. Hartung and G. Knapp, "Statistical Inference in Adaptive Group Sequential Trials With the Standardized Mean Difference as Effect Size," *Sequential Analysis* 30, no. 1 (2011): 94–113, https://doi.org/10.1080/07474946.2011.539926.

109. T. Jaki and D. Magirr, "Considerations on Covariates and Endpoints in Multi-Arm Multi-Stage Clinical Trials Selecting all Promising Treatments," *Statistics in Medicine* 32, no. 7 (2013): 1150–1163, https://doi.org/10.1002/sim.5669.

110. W. Liu, "A Group Sequential Procedure for all-Pairwise Comparisons of k Treatments Based on the Range Statistic," *Biometrics* 51, no. 3 (1995): 946–955, https://doi.org/10.2307/2532995.

111. A. R. Sampson and M. W. Sill, "Drop-The-Losers Design: Normal Case," *Biometrical Journal* 47, no. 3 (2005): 257–268; discussion 269–281, https://doi.org/10.1002/bimj.200410119.

112. N. Stallard and S. Todd, "Point Estimates and Confidence Regions for Sequential Trials Involving Selection," *Journal of Statistical Planning and Inference* 135, no. 2 (2005): 402–419, https://doi.org/10.1016/j.jspi.2004.05.006.

113. M. W. Sill and A. R. Sampson, "Drop-The-Losers Design: Binomial Case," *Computational Statistics and Data Analysis* 53, no. 3 (2009): 586–595, https://doi.org/10.1016/j.csda.2008.07.031.

114. S. S. Wu, W. Wang, and M. C. K. Yang, "Interval Estimation for Drop-The-Losers Designs," *Biometrika* 97, no. 2 (2010): 405–418, https://doi.org/10.1093/biomet/asq003.

115. D. Neal, G. Casella, M. C. K. Yang, and S. S. Wu, "Interval Estimation in Two-Stage, Drop-The-Losers Clinical Trials With Flexible Treatment Selection," *Statistics in Medicine* 30, no. 23 (2011): 2804–2814, https://doi.org/10.1002/sim.4308.

116. D. Magirr, T. Jaki, M. Posch, and F. Klinglmueller, "Simultaneous Confidence Intervals That Are Compatible With Closed Testing in Adaptive Designs," *Biometrika* 100, no. 4 (2013): 985–996, https://doi.org/10.1093/biomet/ast035.

117. J. Bowden and E. Glimm, "Conditionally Unbiased and Near Unbiased Estimation of the Selected Treatment Mean for Multistage Drop-The-Losers Trials," *Biometrical Journal* 56, no. 2 (2014): 332–349, https://doi.org/10.1002/bimj.201200245.

118. M. Carreras, G. Gutjahr, and W. Brannath, "Adaptive Seamless Designs With Interim Treatment Selection: A Case Study in Oncology," *Statistics in Medicine* 34, no. 8 (2015): 1317–1333, https://doi.org/10.1002/sim.6407.

119. M. Brückner, A. Titman, and T. Jaki, "Estimation in Multi-Arm Two-Stage Trials With Treatment Selection and Time-To-Event Endpoint," *Statistics in Medicine* 36, no. 20 (2017): 3137–3153, https://doi.org/10.1002/sim.7367.

120. J. Whitehead, Y. Desai, and T. Jaki, "Estimation of Treatment Effects Following a Sequential Trial of Multiple Treatments," *Statistics in Medicine* 39, no. 11 (2020): 1593–1609, https://doi.org/10.1002/sim.8497.

121. Z. Shun, K. K. G. Lan, and Y. Soo, "Interim Treatment Selection Using the Normal Approximation Approach in Clinical Trials," *Statistics in Medicine* 27, no. 4 (2008): 597–618, https://doi.org/10.1002/sim.2990.

122. J. Bowden and E. Glimm, "Unbiased Estimation of Selected Treatment Means in Two-Stage Trials," *Biometrical Journal* 50, no. 4 (2008): 515–527, https://doi.org/10.1002/bimj.200810442.

123. U. Bandyopadhyay and A. Biswas, "Interval Estimation in Adaptive Allocations Using Point Estimates," *Calcutta Statistical Association Bulletin* 54, no. 1–2 (2003): 31–44, https://doi.org/10.1177/0008068320030103.

124. A. Baldi Antognini, M. Novelli, and M. Zagoraiou, "A New Inferential Approach for Response-Adaptive Clinical Trials: The Variance-Stabilized Bootstrap," *Test* 31, no. 1 (2022): 235–254, https://doi.org/10.1007/s11749-021-00777-9.

125. A. Lane, "Conditional Information and Inference in Response-Adaptive Allocation Designs," *Statistics in Medicine* 41, no. 2 (2022): 390–406, https://doi.org/10.1002/sim.9243.

126. L. J. Wei, R. T. Smythe, D. Y. Lin, and T. S. Park, "Statistical Inference With Data-Dependent Treatment Allocation Rules," *Journal of the American Statistical Association* 85, no. 409 (1990): 156–162, https://doi.org/10.1080/01621459.1990.10475319.

127. D. Tolusso and X. Wang, "Interval Estimation for Response Adaptive Clinical Trials," *Computational Statistics and Data Analysis* 55, no. 1 (2011): 725–730, https://doi.org/10.1016/j.csda.2010.06.016.

128. W. Brannath, E. Zuber, M. Branson, et al., "Confirmatory Adaptive Designs With Bayesian Decision Tools for a Targeted Therapy in Oncology," *Statistics in Medicine* 28, no. 10 (2009): 1445–1463, https://doi.org/10.1002/sim.3559.

129. M. Rosenblum, "Confidence Intervals for the Selected Population in Randomized Trials That Adapt the Population Enrolled," *Biometrical Journal* 55, no. 3 (2013): 322–340, https://doi.org/10.1002/bimj.201200080.

130. S. S. Wu, Y. H. Tu, and Y. He, "Testing for Efficacy in Adaptive Clinical Trials With Enrichment," *Statistics in Medicine* 33, no. 16 (2014): 2736–2745, https://doi.org/10.1002/sim.6127.

131. P. K. Kimani, S. Todd, L. A. Renfro, et al., "Point and Interval Estimation in Two-Stage Adaptive Designs With Time to Event Data and Biomarker-Driven Subpopulation Selection," *Statistics in Medicine* 39, no. 19 (2020): 2568–2586, https://doi.org/10.1002/sim.8557.

132. R. Ishii, K. Takahashi, K. Maruo, and M. Gosho, "Statistical Inference for a Two-Stage Adaptive Seamless Design Using Different Binary Endpoints," *Statistics in Medicine* 44, no. 6 (2025): e70003, https://doi.org/10.1002/sim.70003.

133. R. Brookmeyer and J. Crowley, "A Confidence Interval for the Median Survival Time," *Biometrics* 38, no. 1 (1982): 29–41, https://doi.org/10.2307/2530286.

134. M. J. Grayling and G. M. Wheeler, "A Review of Available Software for Adaptive Clinical Trial Design," *Clinical Trials* 17, no. 3 (2020): 323–331, https://doi.org/10.1177/1740774520906398.

135. K. Kim and A. A. Tsiatis, "Independent Increments in Group Sequential Tests: A Review," *SORT: Statistics and Operations Research Transactions* 44 (2020): 223–264, https://doi.org/10.2436/20.8080.02.101.

136. P. Royston and M. K. Parmar, "Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials With a Time-To-Event Outcome," *BMC Medical Research Methodology* 13, no. 1 (2013): 152, https://doi.org/10.1186/1471-2288-13-152.

137. S. B. Vardeman, "What About the Other Intervals?," *American Statistician* 46, no. 3 (1992): 193–197, https://doi.org/10.2307/2685212.

138. K. Krishnamoorthy and T. Mathew, *Statistical Tolerance Regions: Theory, Applications, and Computation*, 1st ed. (Wiley, 2009).

139. I. C. Marschner, "Confidence Distributions for Treatment Effects in Clinical Trials: Posteriors Without Priors," *Statistics in Medicine* 43, no. 6 (2024): 1271–1289, https://doi.org/10.1002/sim.10000.

140. D. S. Robertson, T. Burnett, B. Choodari-Oskooei, et al., "Confidence Intervals for Adaptive Trial Designs II: Case Study and Practical Guidance." 2024, https://doi.org/10.48550/ARXIV.2411.08771.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Supporting Information.

Appendix A

Orderings of the Sample Space for Group Sequential Designs

As described in Jennison and Turnbull [20], let the pair (k, z) denote a (standardized) test statistic z observed at the stage k that a group sequential trial stops. Also let denote the (Fisher) information at stage k, while a_k and b_k denote the stopping boundary for futility and efficacy, respectively, at stage k. We write $(k', z') > (k, z)$ to denote that (k', z') is more extreme evidence against the null hypothesis than (k, z) in a given ordering of the sample space.

- *Stage-wise ordering*: $(k', z') > (k, z)$ if any one of the three conditions holds:
  i. $k' = k$ and $z' > z$
  ii. $k' < k$ and $z' \geq b_{k'}$
  iii. $k' > k$ and $z' \leq a_{k'}$
- *MLE ordering*: $(k', z') > (k, z)$ if $z'/\sqrt{I_{k'}} > z/\sqrt{I_k}$;
- *Likelihood ratio ordering*: $(k', z') > (k, z)$ if $z' > z$;
- *Score test ordering*: $(k', z') > (k, z)$ if $z'\sqrt{I_{k'}} > z\sqrt{I_k}$.

As a simple example, consider a two-stage group sequential design with early stopping only for efficacy and $b_1 = 2.8$, $b_2 = 2.0$. Consider the following observed trial results (k, z):

A. (1, 3.0)

B. (2, 3.5)

C. (2, 2.5)

Suppose also the information levels be $I_1 = 50$ and $I_2 = 100$. We would then have the following rankings of these trial results (from most to least extreme):

Stage-wise ordering: A > B > C.

MLE ordering: A > B > C.

Likelihood ratio ordering: B > A > C.

Score test ordering: B > C > A.

PRISMA Flowchart for the Systematic Review

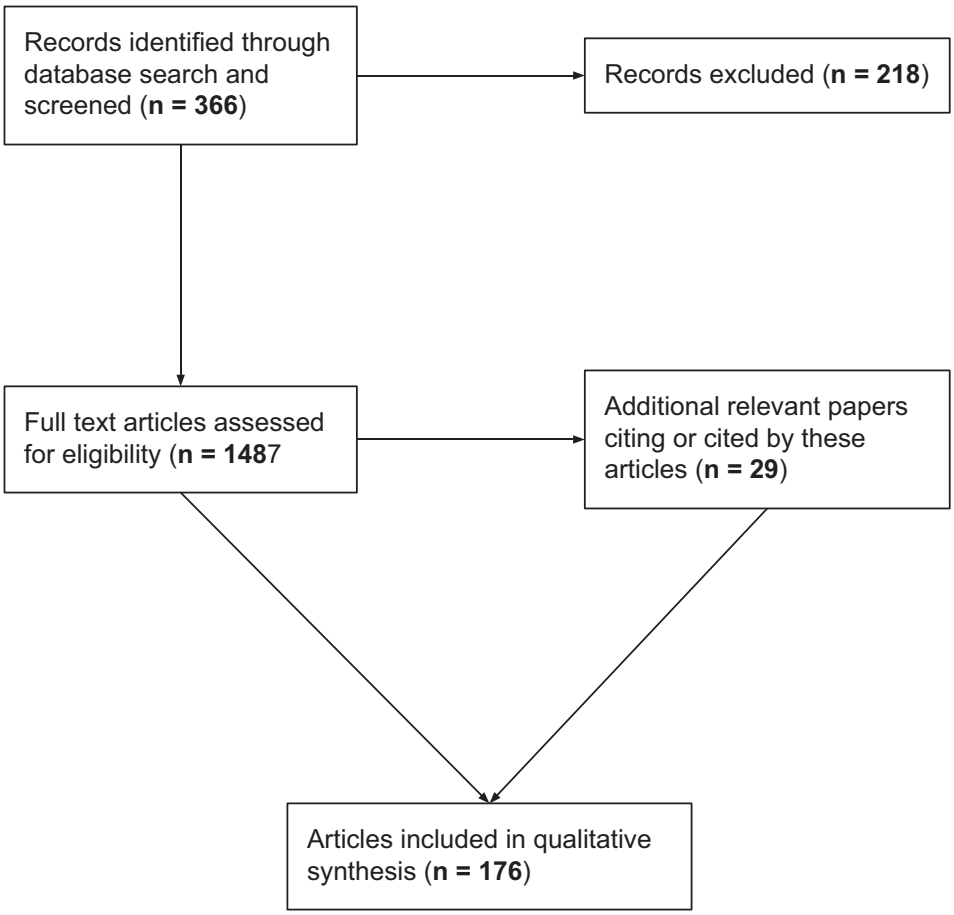

FIGURE A1 | PRISMA flowchart for the systematic review.