Journal of Clinical Epidemiology 153 (2023) 45–54

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

An overview of methodological considerations regarding adaptive stopping, arm dropping, and randomization in clinical trials

Anders Granholm[a,*], Benjamin Skov Kaas-Hansen[a,b], Theis Lange[b], Olav Lilleholt Schjørring[c,d], Lars W. Andersen[e,f,g], Anders Perner[a], Aksel Karl Georg Jensen[b], Morten Hylander Møller[a]

[a]*Department of Intensive Care, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark*
[b]*Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[c]*Department of Anaesthesia and Intensive Care, Aalborg University Hospital, Aalborg, Denmark*
[d]*Department of Clinical Medicine, Aalborg University, Aalborg, Denmark*
[e]*Research Center for Emergency Medicine, Department of Clinical Medicine, Aarhus University and Aarhus University Hospital, Aarhus, Denmark*
[f]*Department of Anesthesiology and Intensive Care, Aarhus University Hospital, Aarhus, Denmark*
[g]*Prehospital Emergency Medical Services, Central Denmark Region, Aarhus, Denmark*

Accepted 2 November 2022; Published online 17 November 2022

Abstract

Background and Objectives: Adaptive features may increase flexibility and efficiency of clinical trials, and improve participants' chances of being allocated to better interventions. Our objective is to provide thorough guidance on key methodological considerations for adaptive clinical trials.

Methods: We provide an overview of key methodological considerations for clinical trials employing adaptive stopping, adaptive arm dropping, and response-adaptive randomization. We cover pros and cons of different decisions and provide guidance on using simulation to compare different adaptive trial designs. We focus on Bayesian multi-arm adaptive trials, although the same general considerations apply to frequentist adaptive trials.

Results: We provide guidance on 1) interventions and possible common control, 2) outcome selection, follow-up duration and model choice, 3) timing of adaptive analyses, 4) decision rules for adaptive stopping and arm dropping, 5) randomization strategies, 6) performance metrics, their prioritization, and arm selection strategies, and 7) simulations, assessment of performance under different scenarios, and reporting. Finally, we provide an example using a newly developed *R* simulation engine that may be used to evaluate and compare different adaptive trial designs.

Conclusion: This overview may help trialists design better and more transparent adaptive clinical trials and to adequately compare them before initiation.

Keywords: Clinical trials; Adaptive trials; Randomization; Simulation; Response-adaptive randomization; Trial design

Declarations of interest: The Department of Intensive Care at Rigshospitalet has received funds for other research projects from the Novo Nordisk Foundation, Fresenius Kabi, and Pfizer, and conducts contract research for AM-Pharma.

Funding: This study was conducted as part of the Intensive Care Platform Trial (INCEPT) program (www.incept.dk), which has the primary purpose of initiating a platform trial investigating commonly used interventions in critically ill adults acutely admitted to an intensive care unit. The INCEPT program has received funding from Sygeforsikringen "danmark", Grosserer Jakob Ehrenreich og Hustru Grete Ehrenreichs Fond, and Dagmar Marshalls Fond, which had no influence on planning, conduct, analyses, or reporting of this project.

Author contributions: Conceptualization: AG, BSKH, OLS, LWA, AKGJ, MHM. Methodology: AG, BSKH, AKGJ, MHM. Software: AG, BSKH, TL, AKGJ. Formal analysis: AG. Writing—Original Draft: AG. Writing—Review & Editing: all authors.

Supporting information, code, and data availability: Example along with additional methodological details, results and further discussion is included in Supplement 1, which also provides instructions for installing the *"adaptr"* R-package [28] used for these simulations. Full annotated analysis code (used to generate and analyze all simulated data) is provided in Supplement 2.

* Corresponding author: Department of Intensive Care, Copenhagen University Hospital – Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. Tel.: +45 3545 4131; fax: +45 3545 2736.

E-mail address: andersgran@gmail.com (A. Granholm).

https://doi.org/10.1016/j.jclinepi.2022.11.002

What is new?

Key Points

- Adaptive clinical trials are flexible and adaptive features may increase trial efficiency and individual participants' chances of being allocated to superior interventions.
- Adaptive trials come with increased complexity and not all adaptive features may always be beneficial.

What this adds to what is known?

- This manuscript provides an overview of and guidance on key methodological considerations for clinical trials employing adaptive stopping, adaptive arm dropping, or response-adaptive randomization.
- In addition, a simulation engine and example on how to compare adaptive trial designs using simulation is provided.

What is the implication and what should change now?

- This guidance paper may help trialists design and plan adaptive clinical trials.

1. Introduction

Most randomized clinical trials (RCTs) compare two interventions at a time and run until a prespecified sample size (or event number) has been accrued, with few or no interim analyses [1]. For economic and logistic reasons, sample size estimations are often based on anticipated intervention effects substantially larger than the minimal clinically important differences or plausible effect sizes expected by peers [2]. Consequently, RCTs may fail to accept or firmly reject hypotheses about clinically important intervention effects, and lack of statistical significance is often erroneously interpreted as ''no difference'' in interventional trials [3−6]. Ultimately, this may lead to premature abandonment of potentially relevant interventions. Further, with few or no interim analyses (especially if the criteria for early stopping are very strict [7]), there is a risk that trials continue longer than necessary if intervention effects are underestimated in sample size calculations. Conventional RCTs are thus somewhat inflexible, which may lead to ineffectiveness, inconclusiveness, suboptimal use of research resources, incorrect promises to trial participants, and ultimately slower improvements in quality of care.

For these reasons, there is increased interest in more elaborate, flexible, and adaptive trial designs [1], including adaptive multi-arm and platform trials [8]. Adaptive trials are trials where results from adaptive (interim) analyses are used to modify aspects of the trial [9,10], by, for example, including adaptive stopping rules or sample sizes, adaptive arm dropping, and response-adaptive randomization (RAR) [8,11]. Adaptive trials use adaptive (interim) analyses to, for example, adjust the target sample size, update allocation ratios, drop inferior intervention arms early, or stop the trial if a prespecified statistical decision rule is met [8−10]. This may lead to increased efficiency and more conclusive trials [12−14], especially if multiple interventions are compared simultaneously or against the same control group, with inferior arms dropped early [8,11]. However, adaptive features increase complexity and may not always increase efficiency, and thus may not always be ideal [15−22].

Given these considerations, it is paramount that trialists planning adaptive trials carefully consider the ideal design, including the advantages and disadvantages of different methodological choices. In this manuscript, we provide an overview of the key methodological considerations when planning adaptive trials and provide practical guidance on how to use simulation to make informed comparisons and decide between multiple possible adaptive trial designs. We focus on adaptive multi-arm trials conducted using Bayesian statistical methods and employing adaptive stopping, adaptive arm dropping and RAR, although the decisions and considerations discussed also apply to other adaptive trials.

2. Overview

2.1. Scope

We describe the key methodological considerations when planning adaptive clinical trials, focusing on adaptive stopping, adaptive arm dropping, and RAR. While we focus on adaptive multi-arm (>2 arms), late-stage pragmatic trials of interventions already in clinical use, similar considerations apply to two-arm adaptive trials and to some earlier-phase trials. Of note, we do not specifically cover adaptive dose-finding studies or specific regulatory requirements for approval of new drugs or devices. General design considerations applicable to all RCTs (including more conventional RCTs) are discussed elsewhere [1], and trial features not directly related to the adaptive features themselves (for example, allocation concealment and blinding) are not covered here.

We focus on adaptive trials using Bayesian statistical methods, as these are commonly applied in complex adaptive trials [8], and as Bayesian probabilities from the most recent adaptive analysis in a trial have the same interpretation regardless of the intended or actual number of adaptive analyses conducted [23]. Despite methodological differences, similar concerns apply as to adaptive trials using frequentist methods. These include an increased risk of erroneous conclusions due to random fluctuations when an increased

number of analyses are conducted, all else being equal. Thus, despite being frequentist concepts, Bayesian analogues to power and type 1 and 2 errors are commonly evaluated for Bayesian adaptive trial designs [16,21,23], as recommended by regulatory authorities [24,25]. In a Bayesian context, power may be defined as the overall probability of a conclusive result (or of declaring an arm superior); type 1 error risk may be defined as the probability of declaring an intervention superior when no differences between intervention arms exist (including in simulations); and type 2 error risk may be defined as the risk of an inconclusive adaptive trial (or a trial that conclusively claims that there is no difference) despite differences being present.

2.2. Simulation-based example and simulation engine

To evaluate and compare complex adaptive clinical trial designs, simulations are required [26,27]. Thus, in addition to a comprehensive review of the key methodological considerations, we have developed a simulation engine—the *"adaptr"* R statistical software package [28]—to aid fellow trialists. The package (installation instructions are provided in Supplement 1) allows specification and simulation of adaptive trials and calculation of performance metrics. We provide a simulation-based example comparing different adaptive trial designs along with additional methodological details and complete analysis code in Supplement 1 and 2, where additional technical details and discussion of the simulations conducted are also provided.

While the *"adaptr"* package provides a freely accessible, well-documented, open-source simulation engine, it comes with some limitations [28]. These include the limited number of outcome types supported out-of-the-box (binary, binomially distributed outcomes and continuous, normally distributed outcomes only, although the package supports user-written function for analyzing other outcomes), the lack of the ability to simulate losses to follow-up and varying inclusion rates, and the selected forms of RAR available [29], and others. These limitations are further discussed in the Supplement 1. Of note, other software options for planning and comparing adaptive trials exist [30–32], all with different advantages and disadvantages. Importantly, the considerations outlined in this manuscript are relevant regardless of the software package used.

2.3. Considerations

The key methodological considerations specific to adaptive trials are summarized in Figure 1, with each point discussed in detail in the following sections and further scrutinized in the simulation-based worked example included in Supplement 1. Importantly, the considerations are intertwined and may not always be made in the specified order. The process is thus iterative, with possible changes in decisions based on simulation results until a final design is chosen.

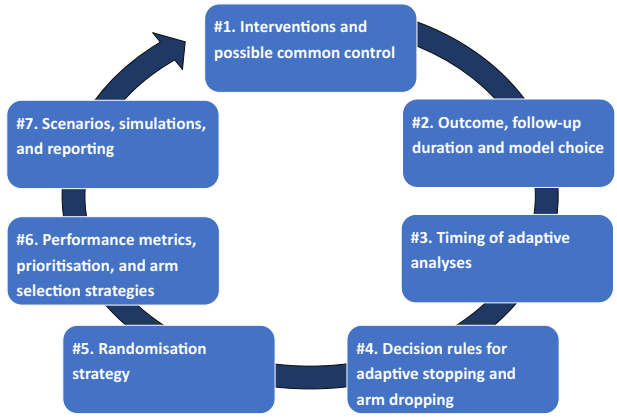

Fig. 1. Key methodological considerations in trial designs with adaptive stopping, adaptive arm dropping, or adaptive randomization. Decisions do not necessarily have to be made in this order, and iteratively refining trial designs according to simulation results will often be necessary, hence the circular design. As different design decisions interact, simulation-based comparisons evaluate all decisions simultaneously. Additional details on each consideration are provided in sections 3-9 in the text.

3. Consideration #1: interventions and possible common control

3.1. Interventions

Select interventions (treatments) to be compared. The number of intervention arms should match the clinical question but also consider expected recruitment rates, as more arms require larger total sample sizes and increase clinical and logistical complexity. If the total number of interventions of interest is higher than what is considered feasible to study simultaneously, adaptive platform trial designs where new arms can be introduced during trial conduct may be considered [8,33], although not discussed further here.

3.2. Control arm

If one intervention can be considered standard of care, this may be used as a common control against which all other arms are compared, which may increase efficiency. Otherwise, all arms may be compared with each other simultaneously. This will affect trial design performance metrics and several of the choices discussed below.

4. Consideration #2: outcome, follow-up duration, and model choice

4.1. Outcome and follow-up duration

Select the outcome used to guide the adaptive analyses and consider duration of follow-up and expected data completeness. Longer durations of follow-up will make trials slower to adapt [9], and increase the risk of different estimates when follow-up for all randomized patients (including those without outcome data available at the time

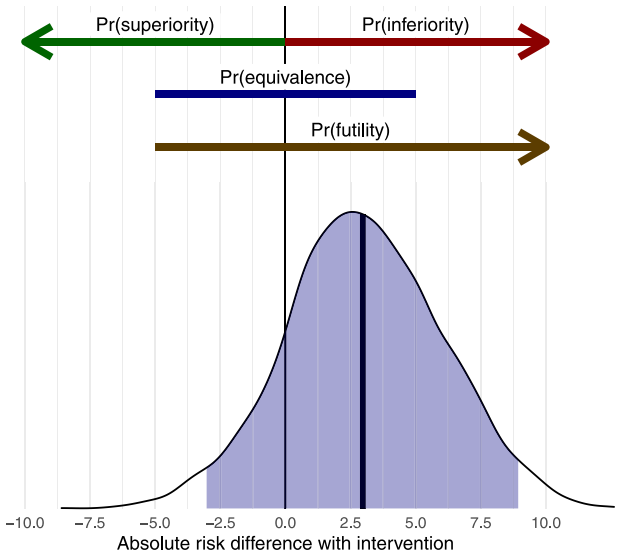

Fig. 2. Decision rules in trial designs with a common control. This figure illustrates different decisions rules for a single comparison of a noncontrol arm with a common control and an undesirable binary outcome. The posterior probability distribution of absolute differences (as percentage points, with values > 0 favoring the control arm) is illustrated. The probabilities (Pr) of superiority and inferiority correspond to the probability mass below/above no difference, respectively (the probability of inferiority equals 100 - the probability of superiority; thus, stopping rules for inferiority are often specified as a low probability of an arm being superior). The probability of practical equivalence corresponds to the probability mass between limits defined by the minimal clinically important difference in either direction (in this example defined as 5 percentage points). Finally, the probability of futility corresponds to the probability mass above the limit defined as the minimal clinically important difference in the beneficial direction.

of the adaptive analysis) concludes compared to when trials are stopped at an adaptive analysis. Longer follow-up durations may lead to more missing data due to loss to follow-up, which may increase complexity of analysis and interpretation. If the outcome of primary interest has a long follow-up duration, an intermediate outcome [33] may be considered to guide adaptation. This can be the same outcome assessed at an earlier time-point or a different outcome that is adequately correlated with the outcome of primary interest. Caution must be taken to detect and handle unexpected lack of correlation between such outcomes, especially if a surrogate and possibly less patient-important outcome is chosen to guide adaptation. For simplicity, we are assuming that a binary, undesirable outcome (for example, mortality) is selected in the rest of the text, but the same concerns apply for other types of outcomes.

4.2. Model choice

Specify the primary statistical model used to assess the primary outcome including the priors used in Bayesian analyses; often weakly- or non-informative priors are used for the primary analyses of adaptive trials, but other choices may be used as well [1,28,34,35]. A detailed discussion of statistical models and priors is beyond the scope of this text, but the statistical model (including priors) used in simulations should roughly correspond to the model that will be used in the primary analysis of the trial. Simplifications may be made during simulations (as in conventional sample size calculations), for example, covariate adjustment may be omitted as simulating covariance patterns is difficult and time-consuming when running simulations. As covariate adjustment generally increases power [36], using simpler analyses may make simulations slightly conservative. Importantly, adaptive and final analyses of actual trials may employ as complex models as required, even if simpler models are used in simulations.

5. Consideration #3: timing of adaptive analyses

5.1. Start and "burn-in"

Consider a *"burn-in"* period at the beginning of the trial with no adaptive analyses conducted until a minimum number of patients have been enrolled in order to limit the risk of inappropriate adaptations due to random fluctuations when limited data have been collected [16,21,33] and to ensure that an acceptable number of subjects have been enrolled to allow for more complex statistical analyses (for example, adjustment for multiple covariates).

5.2. Frequency of adaptive analysis

Specify the frequency of adaptive analyses according to follow-up rates (for example, each time 200 patients complete follow-up) or time intervals (for example, monthly); the latter may ease coordination of data collection across centers in practice. All else being equal, the risk of random errors grows with the number of adaptive analyses, while fewer adaptive analyses limit the benefits of conducting an adaptive trial [21]. Of note, adaptive analyses on the way to a predefined maximum sample size may be called "interim analyses" as in conventional RCTs, while the term "adaptive analysis" is more appropriate in adaptive trials without a prespecified target sample size. For feasibility reasons (for example, economical or logistic constraints), a maximum sample size may be specified a priori.

6. Consideration #4: decision rules for adaptive stopping and arm dropping

6.1. Decision rules

Multiple probabilistic decision rules are described in this section; these are illustrated in Figure 2 and Figure S1 in Supplement 1. In addition to the probabilistic decision rules described below, trials may be stopped at a prespecified maximum sample size.

6.2. Superiority

Define a decision rule for superiority as the probability threshold for an intervention being better than the common control or the best of all arms. The desired level of evidence for declaring superiority for an intervention (or a specific outcome) may vary. Weaker evidence may be sufficient for choosing between well-known interventions already in clinical use, while stronger evidence may be required for new, expensive, or invasive interventions [37]; this may be based on cost-effectiveness analyses [38]. In designs with a common control group, an intervention declared superior to the common control may become the new control against which all remaining arms are subsequently compared [33,34]. Controlling the type 1 error rate is generally important and may be required in regulatory settings and for approval of new interventions [24,25]. The superiority threshold in adaptive trials may thus be chosen to match a desired type 1 error rate [16,34] under a null scenario as discussed under considerations #6-7 (in sections 8-9).

6.3. Inferiority

Inferior interventions may be dropped from multi-arm adaptive trials before an overall superior arm has been identified. Inferiority thresholds can be specified as a low probability of each arm being the best or better than a common control. Dropping inferior arms will increase allocation to the remaining arms and thus increase overall trial efficiency and power for the remaining comparisons. Temporary arm dropping may also be considered, with allocation temporarily paused to interventions if their probabilities of being the best/better than the control drop below a certain threshold, and resumed if probabilities change in a later analysis, for example, due to more undesirable outcomes in the other arms [16,21]. However, clinical complexity may ensue with this modality, especially in multicenter trials. Like superiority, an inferiority threshold may be chosen according to the probability of incorrectly dropping noninferior arms based on simulation.

6.4. Practical equivalence and futility

A stopping rule for practical equivalence may be defined as a certain probability of the differences between noncontrol arms and the control arm being less than a specified difference (for example, the minimal clinically important difference) or in trials without a common control, as an adequately high probability of differences between the best and worst arms being less than a similar threshold. Similarly, a stopping rule for futility—a substantially low probability of an experimental arm being better than the control arm by, for example, the minimal clinically important difference—can be specified in trial designs with a common control group [39]. Interventions may thus be dropped for futility without fulfilling the inferiority criteria (Fig. 2). Typically, somewhat lower probability thresholds are used for equivalence and futility assessments for practical reasons, as this generally requires substantially more patients than assessments of superiority/inferiority [39]. Equivalence and futility thresholds may be specified on the absolute or relative scale, with absolute differences preferred as they are more clinically relevant and easier to interpret [40]. In adaptive designs with a common control group where superior arms are allowed to become the new control group, it must be decided whether equivalence or futility testing is only done for the first control arm or also for any superior arms that become the new control.

7. Consideration #5: randomization strategy

7.1. Initial allocation

Specify initial allocation ratios; initial equal randomization to all arms is the most efficient choice if no common control arm is used. If a common control arm is used and all other arms are pair-wise compared to the control, relatively higher allocation to the control may increase power in some design (for example, some designs using RAR), while power may be highest with equal allocation in other designs (for example, designs with fixed allocation ratios). Importantly, increased control group allocation may come at the expense of a higher total number of patients experiencing an undesirable outcome if an experimental arm is better than the control [16,21,33]. An allocation ratio based on the square root of the number of noncontrol arms to one (for each of the remaining arms) has been recommended where increased control arm allocation is desired to optimize power relative to other nonequal allocation ratios [33,41]. In a four-arm trial with a common control arm, this corresponds to 1.73:1:1:1 allocation or 36.6% allocation to the control arm and 21.1% allocation to each noncontrol arm.

7.2. Fixed randomization, RAR, combinations, and limitations

Decide on the use of fixed randomization, RAR, or combinations. If RAR is not used, the initial allocation ratios will remain unchanged until arms are dropped and the remaining allocation probabilities are normalized (scaled to sum to 100%).

When RAR is used, allocation ratios are updated using the accrued data at the time of an adaptive analysis; multiple specific variants of RAR exist [29]. For example, allocation ratios may be updated according to some function of the current probabilities of each arm being the best [8,15,29], and more patients may thus be allocated to interventions more likely to be superior at the end of the trial. Compared to fixed, equal allocation, RAR may increase power and trial efficiency (that is, lower total sample sizes) in some trial designs, but the opposite may also be the case in other trial designs [15−19,29]. RAR may thus lead to more

patients in the trial being allocated to a superior intervention and less total undesirable events (for example, deaths) within the trial [14], at the potential cost of decreasing overall efficiency and requiring larger samples in some trial designs. These designs include two-arm trials [20] and some multi-arm trial without a common control group where fixed, equal randomization may be more efficient [17], and where RAR may delay treatment improvements for patients external to the trial. Consequently, the use of RAR has been debated [18−21,29]. Another limitation of RAR is that updating allocation ratios makes advanced randomization schemes (for example, stratified block randomization or minimization [42]) difficult, and, therefore, simple, unstratified randomization is generally employed. Thus, adequate balance between important stratification variables in smaller trials may be better facilitated with fixed, stratified randomization. Finally, if RAR is used, modeling time drift, that is, changes in control group event rates due to temporal changes in case-mix or usual care may be necessary [8], which further increases complexity. Thus, comparing trial designs using different randomization strategies is recommended.

If RAR is used, specify for which arms and whether a fixed or special allocation fraction (possibly larger than the others) is used for the control arm in relevant designs. In addition to the square-root-based ratio discussed above, matching the control group allocation fraction to that of the best-performing noncontrol arm may be used to maximize power in comparisons between these two arms [12,16]. In addition to the randomization strategy, specify whether a superior arm that becomes the new control should use its original allocation ratio or change to another (for example, a higher) allocation ratio.

If RAR is used, consider limitations to decrease extreme allocation ratios, especially for control arms, where low allocation ratios may decrease power for all comparisons [16,21]. Employing unrestricted RAR may cause a) the probability of an arm appearing best/better than the control to be lower than the true probability in an early analysis (due to chance) leading to low allocation to that arm that may take time to reverse; and b) low allocation to an active arm with a relatively low probability of being the best, but without low enough probability to be deemed inferior, increasing the time until the arm can be dropped due to uncertainty and insufficient power. Thus, RAR may be restricted to avoid extreme allocation ratios [12,21], by either capping allocation fractions at lower and/or upper bounds, or softening them (raising the probability of each arm being the best to a power between 0 and 1 and normalizing; 1 leaves probabilities unchanged, 0 leads to equal allocation, and 0.5 corresponds to the square root transformation) [12,35]. Different softening factors may be used at different times during the trial, to restrict RAR more at earlier times when less data are available and the risk of adapting to random, extreme fluctuations is higher [16].

8. Consideration #6: performance metrics, prioritization, and arm selection strategies

8.1. Select and prioritize performance metrics

Different trial designs may be preferred depending on which performance metrics (Table 1) are considered most important [16,21]. This decision will typically involve considerations related to efficiency for economical/logistic reasons, benefits to internal patients, benefits to external or future patients, and accuracy of estimated intervention effects, as detailed in Table 1 [16]. Previous arguments have been made in favor of designs optimizing benefits to internal patients [43] as well as designs optimizing efficiency and value to external/future patients [18−20]; these decisions are complex, and the relative importance of different metrics must be considered in each trial. If multiple performance metrics are considered relevant, a utility function that simultaneously considers multiple metrics and their relative importance may be explicitly defined a priori to help guide the selection of the optimal, final trial design.

8.2. Calculation of performance metrics

To calculate several performance metrics (Table 1), it must be specified which arm will be selected (that is, used in practice) if the trial ends without declaring a single intervention superior. A common control arm is the obvious choice, but in the absence of such—or if it has been dropped early with no conclusive evidence found for the remaining arms—an intervention may be selected based on the final probabilities of each remaining arm being the best (despite no stopping thresholds reached), cost, inconvenience to patients, availability or practical use, akin to health technology assessments or clinical practice guidelines based on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach [44]. If no intervention of choice can be specified in such cases, performance metrics may be calculated for trials ending with a superiority decision only.

9. Consideration #7: scenarios, simulations, and reporting

9.1. Scenarios

Scenarios with different, but reasonable, expected true events rates in each arm must be specified for the simulation-based comparisons of trial designs. As the true event rates will be unknown at trial initiation, multiple scenarios with different expected event rates should be considered. A null scenario can be included to assess the risk of type 1 errors (as discussed in sections 2 and 6) [16,21,23], despite assumptions of exactly no difference often being implausible in practice. Other scenarios with different between-arm differences based on, for example, a previously established baseline event rate and the most

Table 1. Performance metrics

#	Metric	Description	Presentation[a]
1.	Sample size	Total sample size, that is, the total number of patients enrolled when the trial is stopped, regardless of reason (superiority, practical equivalence, futility, or maximum sample size reached).	Mean, SD, median, IQR, range
2.	Total event count[b]	Total number of events across all arms in the trial.	Mean, SD, median, IQR, range
3.	Total event rate[b]	Total event rate across all arms in the trial (total event count divided by the total number of patients). This corresponds to the expected (mean) event rate for patients in the trial.	Mean, SD, median, IQR, range
4.	Probability of conclusiveness (power)	Probability of conclusiveness; defined as the probability of stopping for any other reason than inconclusiveness at the maximum sample size (that is, stopping for superiority, practical equivalence or futility before or at the maximum sample size). Power may be defined as the probability of conclusiveness or as the probability of stopping for superiority only (see metric #5).	Proportion or percentage
5.	Decision probabilities	Probabilities of stopping trials with different final decisions—superiority, practical equivalence of all remaining arms, futility, or inconclusiveness (if a maximum sample size is reached before a stopping rule is reached).	Proportions or percentages
6.	Probabilities of selecting[c] each arm	Probabilities of selecting each intervention arm[c] in the trial (see footnote[c] regarding arm selection in inconclusive trials).	Proportions or percentages
7.	RMSE[d] of the selected[c] arm's effect	Root mean squared error (RMSE) of the estimate (for example, the event probability) in the selected arms across simulations compared to the ''true'' simulated value.	RMSE
8.	RMSE[d] of the intervention effect	Root mean squared error (RMSE) of the intervention effect for designs selecting[c] a different arm than the common control arm (or another defined standard of care arm if applicable, in designs where all arms are compared to each other). Calculated based on the differences between the estimated effect estimates (for example, the event probability) in the selected[c] vs. the reference arm, compared to the assumed ''true'' differences in effect estimates between these arms. Smaller numbers are preferable, as this indicates that the estimated intervention effects are closer to the assumed ''true'' intervention effects, meaning that the design is less likely to overestimate intervention effects due to stopping at random, extreme fluctuations.	RMSE
9.	Ideal design percentage (IDP)[e]	A combined measure of arm selection probabilities and the importance or consequences of selecting an inferior arm (for example, incorrectly selecting an arm with a 1 percentage point absolute higher mortality rate than the best arm is less severe than selecting an arm with a 5 percentage point higher mortality rate).	Percentage (or proportion)

Abbreviations: IDP, ideal design percentage [16]; IQR, interquartile range; RMSE, root mean squared error; SD, standard deviation.

Overview of different performance metrics, as discussed elsewhere [16,21]. All are calculated across multiple simulations of the same trial design. Performance metrics may be prioritized according to economic/logistic reasons (that is, total sample sizes); benefits to those included in the trial, that is, internal patients (total event counts/rates); benefits to future patients and those not included in the trial, that is, external patients (probability of conclusiveness/superiority, arm selection probabilities, ideal design percentage); and accuracy of the estimated event rate of the selected arm and accuracy of the intervention effects (RMSE of the event rate of the selected arm and of the intervention effect, if a noncontrol arm is chosen in designs with a common control arm).

[a] For metrics where multiple summary statistics may be used (for example, means and medians), trialists will have to select the summary statistic of primary importance for comparing different designs.

[b] This metric is only directly applicable for binary outcomes. For non-binary outcomes, other measures summarising outcomes across all arms in the trial may be used instead, e.g., for the number of days alive and out of hospital, the overall sum or mean value across all arms may be used.

[c] For the performance measures calculated according to the selected arms, different options for handling trials not stopped for superiority are possible. If a common control arm is used or if one arm can be defined as the standard of care, it may be reasonable to consider this arm as selected (unless the arm is dropped at an adaptive analysis before the final analysis; in this case, no arm or the best remaining arm [highest probability of being the best in the final analysis] may be selected instead). This will likely reflect clinical practice, which is unlikely to change based on an inconclusive trial. If no arm can be considered standard of care, an arm may still be selected based on cost, convenience, or other considerations. These performance metrics can also be calculated for trials ending with a superiority decision only (or either superiority or practical equivalence compared to a common control), as is also possible for the other performance metrics. If multiple selection strategies may be considered reasonable for inconclusive trials, performance metrics may be calculated using multiple selecting strategies based on the same simulations.

[d] Smaller numbers are preferable, as this indicates that the effect estimates of the selected arms are closer to the assumed ''true'' effect sizes, meaning that the trial design is less likely to stop at random, extreme fluctuations.

[e] Defined according to Viele and colleagues [16]; formulas for calculating the IDP are included in Supplement 1. IDPs closer to 100% are preferred.

likely expected effects, the minimal clinically important difference, and clinically unimportant differences in both the expected and opposite directions should also be considered.

9.2. Simulations and reporting

Conduct simulations for all design choices (as per considerations #1-6) and for each scenario, followed by iterative revision of trial designs and decisions as necessary. A smaller number of simulations (for example, 1,000) may be used to obtain approximate estimates of relative performance that can be used to adjust specifications or abandon clearly inferior designs, while a higher number of simulations (for example, 10,000 or more) should be preferred in the later stages of trial design, to increase accuracy of the final simulation-based estimates. Additional, detailed guidance on how to use simulations to evaluate adaptive trial designs are provided elsewhere [45−48].

Report performance metrics for at least the final design with clear specification of all choices and adaptation rules in the trial protocol or appendices.

10. Discussion

10.1. Summary and discussion

We have outlined the key methodological considerations when designing and comparing adaptive trial designs using adaptive stopping, adaptive arm dropping, and adaptive randomization strategies with explanations, rationales, and discussion of pros and cons of different methodological choices. In addition, we have developed a new simulation engine (the *"adaptr"* R package [28]) for conducting simulations and comparing different adaptive trial designs using adaptive stopping, arm-dropping and RAR. Finally, we have exemplified its use (in Supplements 1 and 2) by comparing multiple designs for a four-arm example trial under different assumptions regarding intervention effects, including detailed code to replicate and amend these simulations. This should serve as a thorough introduction and reference for trialists considering these adaptive methodological features.

While adaptiveness in the form of adaptive stopping rules, adaptive arm dropping and adaptive randomization may increase trial efficiency and offer several other benefits (such as potentially decreased costs and higher chances of conclusive trials), these features also increase complexity. Additionally, some adaptive features (e.g., RAR) may in some cases lead to substantial worsening of several performance metrics [16−20], as also illustrated in the provided example (Supplements 1 and 2). It is, thus, essential to recognize that trials employing these adaptive features require thorough planning, additional methodological knowledge, and extensive simulation-based comparison. Importantly, the ideal trial design depends on the prioritization of different performance metrics; accordingly, different designs may be preferred in different scenarios [16,21] and several similarly performing designs may be reasonable choices (Supplements 1 and 2). While no general recommendation can be made for prioritizing performance metrics, total sample sizes, total event counts/rates, probability of conclusiveness or superiority (power) and the ideal design percentage have straightforward interpretations. In addition, these metrics cover important aspects including internal and external benefit, probability of conclusiveness, and a combined measure of the risk of selecting inferior arms including its consequences [16]. Further, as different designs may be preferable under different scenarios with different true event rates, performance metrics should ideally be evaluated under multiple plausible clinical scenarios including a null scenario, if assessment of the type 1 error risk is desired [16]. Crucially, all adaptation rules must be clearly prespecified before actual trial initiation.

10.2. Strengths and limitations

This manuscript has several strengths. First, it thoroughly describes the necessary steps needed to assess adaptive trial designs under different scenarios and features a simulation-based example. Second, it discusses important performance metrics [16,21] and emphasizes that prioritization varies with context and should be considered carefully for each individual trial. Third, it transparently includes all analysis code and the simulation engine used is freely available.

There are limitations too. First, we have primarily focused on multi-arm late-phase trials and have not covered issues specifically related to earlier-phase trials, including dose-finding studies. Second, adding intervention arms after trial initiation, as may be done in adaptive platform trials [8,33,49], increases complexity and has not been covered here. Third, while this manuscript and the example focus on a single, undesirable, binary outcome guiding adaptations, the same considerations apply to other adaptive designs guided by single outcomes, regardless of type. Multiple outcomes may be considered simultaneously in adaptation rules, and as long as they can be combined to a single probability of each arm being the best, the same general considerations apply. This, however, increases complexity and is beyond the scope of most adaptive trial designs and this manuscript. Fourth, we have not considered the handling of different subgroups in regard to different allocation ratios, separate adaptive decisions, and adaptive enrichment [50]. Fifth, while we provide complete simulation code that may be adapted to other scenarios, there are some methodological features that are not supported by the simulation engine used, as discussed above, in Supplement 1, and elsewhere [28]. Finally, as for sample size calculations in conventional trials, all simulation-based comparisons of adaptive trial designs require certain

assumptions that are unverifiable before trial conduct, as well as several simplifications compared to the complex clinical reality. The benefit of this, however, is that such assumptions are made explicit which facilitates scrutiny. As such, evaluating adaptive trial designs should be done under multiple, different assumed scenarios, as emphasized in this text.

11. Conclusion

In conclusion, we have described the key methodological considerations when planning and comparing adaptive trials designs using adaptive stopping rules, adaptive arm dropping, and adaptive randomization. This work may help trialists design better and more transparent adaptive clinical trials and to adequately compare them before trial initiation.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.11.002.

References

[1] Granholm A, Alhazzani W, Derde LPG, Angus DC, Zampieri FG, Hammond NE, et al. Randomised clinical trials in critical care: past, present and future. Intensive Care Med 2022;48:164−78.

[2] Ridgeon EE, Bellomo R, Aberegg SK, Sweeney RM, Varughese RS, Landoni G, et al. Effect sizes in ongoing randomized controlled critical care trials. Crit Care 2017;21:132.

[3] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.

[4] Scott Braithwaithe R. EBM's six dangerous words. JAMA 2020;323:1676−7.

[5] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305−7.

[6] Hemming K, Javid I, Taljaard M. A review of high impact journals found that misinterpretation of non-statistically significant results from randomised trials was common. J Clin Epidemiol 2022;145:112−20.

[7] Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. BMC Med Res Methodol 2020;20:4.

[8] Adaptive Platform Trials Coalition. Adaptive platform trials: definition, design, conduct and reporting considerations. Nat Rev Drug Discov 2019;18:797−807.

[9] Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 2018;16:29.

[10] Burnett T, Mozgunov P, Pallmann P, Villar SS, Wheeler GM. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. BMC Med 2020;18:352.

[11]] Thorlund K, Haggstrom J, Park JJ, Mills EJ. Key design considerations for adaptive clinical trials: a primer for clinicians. BMJ 2018;360:k698.

[12] Ryan EG, Lamb SE, Williamson E, Gates S. Bayesian adaptive designs for multi-arm trials: an orthopaedic case study. Trials 2020;21:83.

[13] Ryan EG, Stallard N, Lall R, Ji C, Perkins GD, Gates S. Bayesian group sequential designs for phase III emergency medicine trials: a case study using the PARAMEDIC2 trial. Trials 2020;21:84.

[14] Ryan EG, Bruce J, Metcalfe AJ, Stallard N, Lamb SE, Viele K, et al. Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. BMC Med Res Methodol 2019;19:99.

[15] Wason JMS, Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. Stat Med 2014;33:2206−21.

[16] Viele K, Broglio K, McGlothlin A, Saville BR. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. Clin Trials 2020;17:52−60.

[17] Wathen JK, Thall PF. A simulation study of outcome adaptive randomization in multi-arm clinical trials. Clin Trials 2017;14:432−40.

[18] Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? J Clin Oncol 2011;29:771−6.

[19] Thall PF, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann Oncol 2015;26:1621−8.

[20] Hey SP, Kimmelman J. Are outcome-adaptive allocation trials ethical? Clin Trials 2015;12:102−6.

[21] Viele K, Saville BR, McGlothlin A, Broglio K. Comparison of response adaptive randomization features in multiarm clinical trials with control. Pharm Stat 2020;19:602−12.

[22] Wason JMS, Brocklehurst P, Yap C. When to keep it simple − adaptive designs are not always useful. BMC Med 2019;17:152.

[23] Ryan EG, Brock K, Gates S, Slade D. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? BMC Med Res Methodol 2020;20:150.

[24] U.S. Food & Drug Administration. Interacting with the FDA on complex innovative trial designs for drugs and biological products - guidance for industry. 2020. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/interacting-fda-complex-innovative-trial-designs-drugs-and-biological-products. Accessed February 22, 2022.

[25] U.S. Food & Drug Administration. Guidance for the use of bayesian statistics in medical device clinical trials. 2010. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials. Accessed February 22, 2022.

[26] U.S. Food & Drug Administration. Adaptive design clinical trials for drugs and biologics - guidance for industry. 2019. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry. Accessed September 20, 2022.

[27] European Medicines Agency. Complex clinical trials − questions and answers. 2022. Available at: https://health.ec.europa.eu/latest-updates/questions-and-answers-complex-clinical-trials-2022-06-02_en. Accessed September 20, 2022.

[28] Granholm A, Jensen AKG, Lange T, Kaas-Hansen BS. adaptr: an R package for simulating and comparing adaptive clinical trials. J Open Source Softw 2022;7:4284.

[29] Robertson DS, Lee KM, Lopez-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations [preprint]. arXiv 2020.

[30] Meyer EL, Mesenbrink P, Mielke T, Parke T, Evans D, König F. Systematic review of available software for multi-arm multi-stage and platform clinical trial design. Trials 2021;22:183.

[31] Chandereng T, Musgrove D, Haddad T, Hickey G, Hanson T, Lystig T. bayesCT: simulation and analysis of adaptive bayesian clinical trials [R package]. 2020. Available at: https://CRAN.R-project.org/package=bayesCT. Accessed September 20, 2022.

[32] Parsons N. asd: simulations for adaptive seamless designs [R package]. 2016. Available at: https://CRAN.R-project.org/package=asd. Accessed September 20, 2022.

[33] Park JJH, Harari O, Dron L, Lester RT, Thorlund K, Mills EJ. An overview of platform trials with a checklist for clinical readers. J Clin Epidemiol 2020;125:1−8.

[34] Angus DC, Berry S, Lewis RJ, Al-Beidh F, Arabi Y, van Bentum-Puijk W, et al. The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design. Ann Am Thorac Soc 2020;17:879−91.
[35] Thorlund K, Golchi S, Haggstrom J, Mills E. Highly Efficient Clinical Trials Simulator (HECT): software application for planning and simulating platform adaptive trials. Gates Open Res 2019;3:780.
[36] Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 2014;15:139.
[37] Young PJ, Nickson CP, Perner A. When should clinicians act on non−statistically significant results from clinical trials? JAMA 2020;323:2256−7.
[38] Willan AR, O'Brien BJ. Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. Health Econ 1999;8:203−11.
[39] REMAP-CAP Investigators. Interleukin-6 receptor antagonists in critically ill patients with covid-19. N Engl J Med 2021;22:1491−502.
[40] Kent DM, Paulus JK, Van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2020;172:35−45.
[41] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096−121.
[42] Altman DG, Bland JM. How to randomise. BMJ 1999;319:703−4.
[43] Meurer W, Lewis R, Berry D. Adaptive clinical trials: a partial remedy for the therapeutic misconception? JAMA 2012;307:2377−8.
[44] Granholm A, Alhazzani W, Møller MH. Use of the GRADE approach in systematic reviews and guidelines. Br J Anaesth 2019;123:554−9.
[45] Mayer C, Perevozskaya I, Leonov S, Dragalin V, Pritchett Y, Bedding A, et al. Simulation practices for adaptive trial designs in drug and device development. Stat Biopharm Res 2019;11:325−35.
[46] Hummel J, Wang S, Kirkpatrick J. Using simulation to optimize adaptive trial designs: applications in learning and confirmatory phase trials. Clin Invest 2015;5:401−13.
[47] Friede T, Stallard N, Parsons N. Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: methods, simulation model and their implementation in R. Biom J 2020;62:1262−83.
[48] Wathen JK. Simulation for bayesian adaptive designs - step-by-step guide for developing the necessary R code. In: Lakshminarayanan M, Natanegara F, editors. Bayesian Applications in Pharmaceutical Development. New York: Chapman and Hall/CRC; 2019:267−85.
[49] Park JJH, Detry MA, Murthy S, Guyatt G, Mills EJ. How to use and interpret the results of a platform trial: users' guide to the medical literature. JAMA 2022;327:67−74.
[50] Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics 2013;14:613−25.