

RESEARCH ARTICLE

Group sequential designs for clinical trials when the maximum sample size is uncertain

Amin Yarahmadi¹ | Lori E. Dodd² | Thomas Jaki^{3,4} | Peter Horby⁵ |
Nigel Stallard¹

¹Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

²Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, USA

³Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

⁴MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

⁵Nuffield Department of Medicine, University of Oxford, Oxford, UK

Correspondence

Nigel Stallard, Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK.
Email: N.Stallard@warwick.ac.uk

Funding information

Medical Research Council, Grant/Award Numbers: MC_UU_00002/14, MC_UU_0040/03, MR/V038419/1

Motivated by the experience of COVID-19 trials, we consider clinical trials in the setting of an emerging disease in which the uncertainty of natural disease course and potential treatment effects makes advance specification of a sample size challenging. One approach to such a challenge is to use a group sequential design to allow the trial to stop on the basis of interim analysis results as soon as a conclusion regarding the effectiveness of the treatment under investigation can be reached. As such a trial may be halted before a formal stopping boundary is reached, we consider the final analysis under such a scenario, proposing alternative methods for when the decision to halt the trial is made with or without knowledge of interim analysis results. We address the problems of ensuring that the type I error rate neither exceeds nor falls unnecessarily far below the nominal level. We also propose methods in which there is no maximum sample size, the trial continuing either until the stopping boundary is reached or it is decided to halt the trial.

KEYWORDS

conditional error, group-sequential stopping boundary, sequential probability ratio test, spending function, underrunning analysis

1 | INTRODUCTION

The SARS-CoV-2 pandemic has demonstrated the ability of the worldwide scientific community to propose and evaluate potential new therapies for an emerging infectious disease rapidly under challenging conditions. A record of COVID-19 clinical trials was available on the website <https://www.covid-trials.org>,¹ and whilst no longer active, recorded over 2,500 clinical trials starting in 2020, with over 650 trials starting in April 2020 alone. This enormous and impressive level of activity has also demonstrated, however, the challenge of conducting clinical trials under pressures that exist in the setting of a newly emerging disease pandemic. The rapidly evolving and dynamic nature of such a setting imposes particular challenges that arise due to the considerable uncertainty regarding the natural history of the disease, potential treatment effects and number of affected patients. All of these uncertainties mean that conducting sample size calculations prior to starting a clinical trial is not straightforward. This was illustrated by the RECOVERY trial as described in the next section.

The difficulties in specifying a sample size at the start of the trial make sequential or adaptive designs attractive.² These approaches allow a number of interim analyses of accumulating data to be conducted as the trial proceeds, with the oppor-

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

tunity to stop as soon as sufficient evidence is obtained to demonstrate a significant treatment effect or if the trial appears to be futile. In particular, it may be desirable to plan a sequential trial in which there is no maximum sample size specified. This paper describes how one such approach can be implemented based on the sequential probability ratio test.^{3,4}

The considerable uncertainty in the course of an emerging disease epidemic, and in particular in the number of patients infected over time, also means that in some cases a reduction in patient numbers may lead to the necessity to stop the trial before the planned sample size is attained, or before either specified sample size is attained or a pre-defined stopping boundary is reached in a sequential testing procedure. This is illustrated by the RECOVERY-RS trial that is also described in the next section. In order to distinguish between this and termination of the trial when a stopping boundary is reached, we will refer to this as *halting* the trial. In this paper we will consider methods for the analysis of trials that have been halted in this way, with a particular focus on approaches that control but utilise the full type I error rate specified. Typically, although determined by external factors, the decision to halt a trial will be taken by members of the team conducting the study, for example by the trial steering committee. In some cases in which interim analyses have already been conducted this group may have seen interim analysis data, so that it is impossible to rule out the influence of these data on their decision. We will therefore consider methods in which the decision to halt the study is independent of interim analysis data and in which the interim analysis data may have been used.

The paper is structured as follows: Section 2 describes two trials of interventions for COVID-19 to provide motivation for the methods that we will propose and Section 3 briefly describes the statistical tools that will be used. Section 4 then proposes methods for the analysis of halted trials, with Section 4.1 focussing on the setting in which the decision to halt the trial is independent of any data from the trial, and Section 4.2 on the setting in which this decision may depend of data observed at interim analyses that have already been conducted. Section 5 describes the approach in which the group-sequential design has no maximum sample size. The methods proposed are applied to one of the examples of Section 2 in Section 6 and evaluated in a simulation study in Section 7. The paper concludes with discussion in Section 8.

2 | MOTIVATING EXAMPLES: THE RECOVERY AND RECOVERY-RS TRIALS

The ongoing RECOVERY trial⁵ is an international randomised controlled platform trial of treatments for COVID-19. One of the most successful trials in COVID-19, to date the study has identified four life-saving treatments as well as demonstrating that a number of other commonly used treatments are ineffective.⁶⁻⁸ The trial started early in the pandemic in spring 2020, when limited information on COVID-19 was available at the initial planning stage. This led the trial team to make the unusual decision not to specify a sample size for the trial, the protocol reporting that ‘appropriate sample sizes could not be estimated at the start of the trial’.⁹

The RECOVERY-RS trial^{10,11} was a randomised group sequential clinical trial of hospitalised adults with COVID-19-related acute hypoxic respiratory failure. This trial studied the comparison of clinical outcome improvements in non-invasive respiratory strategies: Continuous positive airway pressure (CPAP) and high-flow nasal oxygen (HFNO) with conventional oxygen therapy. The planned sample size was 4,002 with recruitment over a twelve month period. The trial started in April 2020, and in May 2021, after three interim analyses, was stopped early by the trial steering committee (TSC) due to a substantial reduction in hospitalised patients with COVID-19, and also the end of the funded recruitment period. The final analysis showed that the CPAP strategy led to a statistically significant reduction in tracheal intubation or mortality within 30 days over the standard oxygen therapy, whereas there was no statistically significant difference between the HFNO strategy and conventional oxygen therapy.

3 | BACKGROUND

3.1 | Group sequential design notation and specification

Group sequential methods can be used in clinical trials where data are evaluated as they accrue through a series of interim analyses performed at several points within the trial. In such a trial the sample size is not fixed in advanced as the trial is stopped when sufficient evidence become available to enable a decision to be made regarding treatment efficacy. Jennison and Turnbull¹² show how the interim analyses can be based on efficient score statistics for a parameter of interest, θ , measuring the treatment effect where this might, for example, be a log-odds ratio or log-risk ratio in a binary data, log-hazard ratio in survival data, or standardised difference in means for normal data.¹³

Let S_k denote the score statistic based on all data observed up to interim analysis k and let $\mathbf{S} = (S_1, \dots, S_k)^T$ denote the vector of statistics observed up to interim analysis k . Furthermore, let V_k indicate the observed Fisher's information for all data observed the parameter of interest based on data up to interim analysis k .

Jennison and Turnbull¹⁴ show that for a wide range of problems, it can be assumed that asymptotically the random vector $\mathbf{S}$ follows a multivariate normal distribution with

$$\mathbf{S} \sim N \left(\begin{pmatrix} \theta V_1 \\ \theta V_2 \\ \theta V_3 \\ \vdots \\ \theta V_k \end{pmatrix}, \begin{pmatrix} V_1 & V_1 & V_1 & \cdots & V_1 \\ V_1 & V_2 & V_2 & \cdots & V_2 \\ V_1 & V_2 & V_3 & \cdots & V_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ V_1 & V_2 & V_3 & \cdots & V_k \end{pmatrix} \right). \quad (1)$$

We wish to test the null hypothesis, $H_0 : \theta \leq 0$ against the one-sided alternative. We stop and reject the null hypothesis at look k if $S_k \geq u_k, k = 1, \dots$. Let

$$P_k = P_{H_0} \left(\bigcap_{\tilde{k}=1}^{k-1} \{S_{\tilde{k}} < u_{\tilde{k}}\}, S_k \geq u_k \right) \quad (2)$$

denote the probability under H_0 that the null hypothesis is rejected at look k the trial having not stopped at any early interim analysis. If up to a total of K interim analyses are conducted, the overall type I error rate is therefore equal to $P_1 + \dots + P_K$ and we wish to control this at some specified level α , that is to have

$$P_1 + \dots + P_K = P_{H_0} \left(\bigcup_{k=1}^K \{S_k \geq u_k\} \right) = \alpha. \quad (3)$$

We would also like to ensure that the test has specified power $1 - \beta$ under a given alternative hypothesis, $H_1 : \theta = \delta$, that is that

$$P_{H_1} \left(\bigcup_{k=1}^K \{S_k \geq u_k\} \right) = 1 - \beta \quad (4)$$

where the probabilities in Equations (3) and (4) can be calculated using the distribution given by (1).

Note that we will consider tests in which only an upper boundary is specified since we focus on a test of a one-sided hypothesis. In practice, a trial may also be stopped for futility at look k if S_k is sufficiently small, but in order to avoid specifying a binding futility boundary the lower boundary values are not used when calculating $u_1, \dots, u_K$ to satisfy (3). Since stopping for futility cannot lead to rejection of the null hypothesis, doing so when no lower boundary has been specified will lead to a test that is conservative in that the overall type I error rate will be less than α , but will also have power below $1 - \beta$. Futility boundaries are considered in more detail in the Discussion section below.

3.2 | The α -spending function approach

The idea of the α -spending function approach, which was initially introduced by Lan and DeMets,¹⁵ is used to set critical boundaries for group sequential designs. In this approach the amount of α spent up to each interim analysis is specified as a function of the information at that analysis since the condition (3) alone is insufficient to specify the boundary values $u_1, \dots, u_K$ for $K > 1$. Hence, we let $V_{\max}$ denote the maximum information, that is $V_{\max} = V_K$, and define $t_k = V_k/V_{\max}$ to be the *information time*, that is fraction of the maximum information observed, at look k . Suppose that we have some specified *spending function*, $\alpha^*(t)$, an increasing function on $[0, 1]$ with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$, the overall desired type I error rate. The group sequential test can then be constructed such that the type I error rate spent by look k is equal to $\alpha^*(t_k)$, that is, for $k = 1, \dots, K$,

$$P_{H_0} \left(\bigcup_{\tilde{k}=1}^k \{S_{\tilde{k}} \geq u_{\tilde{k}}\} \right) = \alpha^*(t_k), \quad (5)$$

or equivalently

$$P_k = \alpha^*(t_k) - \alpha^*(t_{k-1}) \quad 1 \leq k \leq K,$$

where P_k is as defined in (2) above and we define $t_0 = 0$. For given $V_{\max}$ and $V_1, \dots, V_K$, successive boundary values $u_1, \dots, u_K$ can be constructed to satisfy (5) and hence (3), using the distribution (1). If K and $t_1, \dots, t_K$ are specified in advance, the value of $V_{\max}$ can be obtained so that the power requirement (4) is satisfied.

For given $V_{\max}$, using (1) to calculate $u_1, \dots, u_K$ to satisfy (5) will control the overall type I error rate at level α for any K and $V_1, \dots, V_K = V_{\max}$ provided V_k does not depend on $(S_1, \dots, S_{k-1})$, though in general if V_k depends on $(S_1, \dots, S_{k-1})$ the overall type I error can be inflated above the nominal level as discussed by Proschan et al.¹⁶

A number of forms for the α -spending function, α^* have been proposed for example by Kim and Demets,¹⁷ and Hwang et al.¹⁸

3.3 | The conditional error approach

The idea of *conditional error function* was introduced by Proschan and Hunsberger¹⁹ in order to specify the critical boundaries in an adaptive two-stage design in which the sample size of the second stage can depend on the observed data from the first stage. Müller and Schäfer²⁰ proposed a general principle to allow for adaptations, such as sample size recalculation, modifying the number of interim analyses, and changing the spending function during the trial, whilst maintaining the overall type I error rate. As noted above and demonstrated by Proschan et al.,¹⁶ data-dependent changes to a group-sequential design, including modifying the timing of subsequent interim analyses on the basis of observed data, can lead to type I error rate inflation. Müller and Schäfer²⁰ define the *conditional error rate* (CE), the probability under the null hypothesis of rejecting the null hypothesis at any future interim looks given the observed value of the test statistic at an interim analysis. The remainder of the group sequential design can then be modified with full flexibility, including with data-dependent adaptations, provided this conditional error rate remains unchanged.

In detail, suppose that a group-sequential test is planned with up to K looks with $S_1, \dots, S_K$ distributed as given by (1) for specified $V_1, \dots, V_K$ with $u_1, \dots, u_K$ constructed so as to give overall type I error rate α . At some look r , suppose that we have observed data $(S_1 = s_1, \dots, S_{r-1} = s_{r-1})$ with $s_k \leq u_k, k = 1, \dots, r-1$ so that the test has not stopped. Given these data, the probability of a conditional error is equal to

$$P_{H_0} \left(\bigcup_{k=r}^K S_k \geq u_k \mid \bigcup_{k=1}^{r-1} S_k = s_k \right). \quad (6)$$

The test can be redesigned, possibly based on the values of $s_1, \dots, s_{r-1}$, to have a maximum number of looks K' with information levels $V'_r, \dots, V'_{K'}$ and score statistics $S'_r, \dots, S'_{K'}$ compared with critical values $u'_r, \dots, u'_{K'}$ without inflating the type I error rate above α for any set of critical values $u'_r, \dots, u'_{K'}$ satisfying

$$P_{H_0} \left(\bigcup_{k=r}^{K'} S'_k \geq u'_k \mid \bigcup_{k=1}^{r-1} S_k = s_k \right) = P_{H_0} \left(\bigcup_{k=r}^K S_k \geq u_k \mid \bigcup_{k=1}^{r-1} S_k = s_k \right). \quad (7)$$

The probabilities in (6) and (7) can be obtained from the conditional distribution of $S_{r+1}, \dots, S_K$ given $S_1 = s_1, \dots, S_{r-1} = s_{r-1}$. Since $S_1, \dots, S_K$ have multivariate normal distribution (1), this conditional distribution is multivariate normal with mean

$$\begin{pmatrix} \theta V_r \\ \vdots \\ \theta V_K \end{pmatrix} + \Sigma_{21} \Sigma_{11}^{-1} \begin{pmatrix} s_1 - \theta V_1 \\ \vdots \\ s_{r-1} - \theta V_{r-1} \end{pmatrix}$$

and variance-covariance matrix

$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

with

$$\Sigma_{11} = \begin{pmatrix} V_1 & V_1 & \cdots & V_1 \\ V_1 & V_2 & \cdots & V_2 \\ \vdots & \vdots & \ddots & \vdots \\ V_1 & V_2 & \cdots & V_{r-1} \end{pmatrix}, \quad \Sigma_{12} = \Sigma_{21}^T = \begin{pmatrix} V_1 & V_1 & \cdots & V_1 \\ \vdots & \vdots & \vdots & \vdots \\ V_{r-1} & V_{r-1} & \cdots & V_{r-1} \end{pmatrix} \quad \text{and} \quad \Sigma_{22} = \begin{pmatrix} V_r & V_r & \cdots & V_r \\ V_r & V_{r+1} & \cdots & V_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ V_1 & V_{r+1} & \cdots & V_K \end{pmatrix}$$

where Σ_{21} has $K - (r - 1)$ columns.

In Section 4.2.2 we show how this approach can be used to construct a critical value for the final analysis when the trial is halted early.

4 | EARLY HALTING OF A GROUP SEQUENTIAL DESIGN

4.1 | Halting time independent of observed interim data

We now consider a group sequential trial that is halted prior to a boundary being crossed. For example the trial might be halted due to a low or declining patient recruitment rate or financial limitations. We initially assume that the decision to halt the study is independent of any data already observed. This will be the case if the halt is due to some external factor and/or if the decision to halt the trial is made without knowledge of any interim data. The case when the decision to halt the trial could be related to the observed data is considered in Section 4.2 below.

Suppose that the halt occurs after some number, $r - 1$ with $r \geq 1$, interim analyses have been conducted using critical values obtained based on (5) for a pre-specified spending function and value of $V_{\max}$ and that at the point of halting the trial a final, r th, analysis is conducted with score statistic S_r and information V_r . Since $V_r < V_{\max}$, as otherwise the trial would have been completed, the test will be conservative and lack power since the critical value for the final analysis is based on the spending function value $\alpha^*(V_r/V_{\max})$ which is typically less than α (see Appendix A for details).

An alternative approach is to spend all remaining α at this final look, replacing the value $\alpha^*(V_r/V_{\max})$ used to calculate u_r with the full level α and to use critical value $\tilde{u}_r$ given by

$$P_{H_0} \left(\bigcup_{k=1}^{r-1} \{S_k < u_k\}, S_r \geq \tilde{u}_r \right) = \alpha - \alpha^*(t_{r-1}). \quad (8)$$

The overall type I error rate is then

$$\sum_{k=1}^{r-1} (\alpha^*(t_k) - \alpha^*(t_{k-1})) + \alpha - \alpha^*(t_{r-1}) = \alpha \quad (9)$$

so that the desired overall type I error rate is maintained at the level of α (see Appendix A for details when the halting time is considered random).

Although the halted test using the critical value $\tilde{u}_r$ for the final analysis will have full level α , the power of this test will, however, be reduced since the amount of data collected, that is, available information, is smaller than the planned sample size for which the trial has power $1 - \beta$. Typically the later the halting time, the lower the power loss will be. Some numerical results of the trade-off between the amount of power reduction, the level of type I error rate, and the timing of halting of the trial are presented in the Section 7 below.

It is worth noting that this method is analogous to *underrunning* in a group sequential trial as considered by Whitehead,²¹ who proposes a method for construction of p -values implicitly equivalent to specification of critical values as described above.

4.2 | Halting time possibly depending on observed interim data

Although driven by external factors, the decision to halt a trial will often be taken by those involved in the running of the study. In fact, when a decision is made with knowledge of interim data, one must assume the interim results influenced decision-making. In Section 4.1 and Appendix A, we showed that if the decision to halt a trial is independent of the observed data, the overall type I error rate is controlled at level of α if the critical boundary for the final test is adjusted to satisfy (8), spending all remaining α at this analysis. If the decision to halt the study is based on the observed data, however, the type I error can be inflated if this critical value is used for the final analysis. In this section we explore the impact of halting at a time that may depend on the interim data in two cases; (i) when the decision can be based on all data available up to and including the halting time, and (ii) when the decision can be based on all data observed up to the interim analysis preceding the halt time but excluding the halting time itself. We show that in the first case the type I error rate can be inflated by using any critical values smaller than u_r obtained for the original design, that is using (5) with the originally specified spending function, and propose an approach to obtaining an alternative critical value that controls the overall type I error rate in the second case.

4.2.1 | Making decision to halt at interim analysis r based on data including S_r

Suppose that at some look r , a decision is to be made to either halt the trial, conducting a final analysis, or to continue with the study. Further assume that this decision is based on observed data up to and including look r , summarised by the test statistic S_r .

This is illustrated by Figure 1 part (A). Let $u_1, \dots, u_r$ denote critical values calculated for the r interim analyses using (5). Suppose that, at look r , a decision is made to halt the trial using some alternative critical value $\bar{u}_r < u_r$. If the trial is not halted, then future interim analyses (ie, at times $r+1, \dots, K$) can continue with $V_K = V_{\max}$ and hence $t_K = 1$ and $\alpha^*(t_K) = \alpha$. Since if $S_r > \bar{u}_r$ the trial will be halted, and, since the final test will use the critical value $\bar{u}_r$, the null hypothesis will be rejected, the probability of stopping and rejecting the null hypothesis at this analysis will be

$$\bar{P}_r = P_{H_0} \left(\bigcup_{k=1}^{r-1} \{S_k < u_k\}, S_r \geq \bar{u}_r \right)$$

and the overall type I error rate will be

$$P_{H_0} (\{S_1 \geq u_1\} \cup \dots \cup \{S_{r-1} \geq u_{r-1}\} \cup \{S_r \geq \bar{u}_r\} \cup \{S_{r+1} \geq u_{r+1}\} \cup \dots \cup \{S_K \geq u_K\}) > P_{H_0} \left(\bigcup_{k=1}^K \{S_k \geq u_k\} \right) = \alpha.$$

The type I error rate can thus be inflated above the nominal level if any critical value $\bar{u}_r < u_r$ is used for the final analysis when the trial is halted with this decision based on the observed data including s_r . This is unsurprising as the critical values for the final analysis essential provide a group sequential boundary, so that any values more liberal than those originally planned can lead to type I error rate inflation.

4.2.2 | Making decision to halt at interim analysis r based on data $S_1, \dots, S_{r-1}$

Consider next the setting in which the decision of whether to halt the trial is made without the current data being observed, but when results from previous interim analyses are known. That is, at some look r the decision can be based on data observed up to the preceding look, that is on $s_1, \dots, s_{r-1}$, (see part (B) in Figure 1). Note that whilst the decision to halt the trial at the some stage r will be made based on $s_1, \dots, s_{r-1}$, the ultimate observed data, s_r , is compared with an updated critical boundary obtained from $s_1, \dots, s_{r-1}$. In turn, if the critical value $\bar{u}_r$ given by (8) is used for the final analysis and the decision is made to stop the trial then it is possible to base the decision on $s_1, \dots, s_{r-1}$ in such a way as to inflate the type I error. In detail, let A_r denote the conditional error rate given $S_1 = s_1, \dots, S_{r-1} = s_{r-1}$ for continuing with the original design, that is

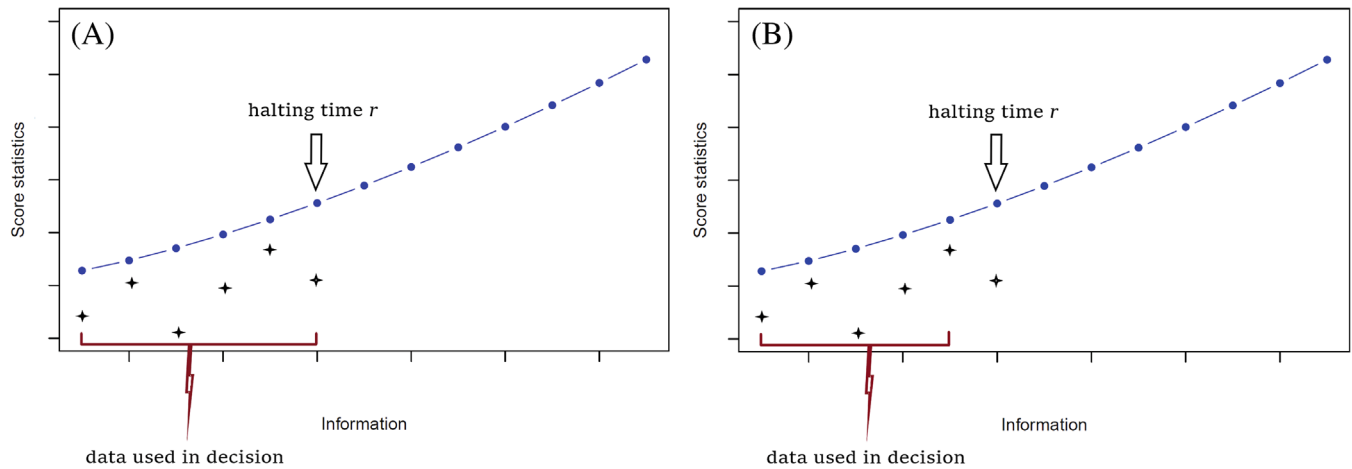


FIGURE 1 The schematic diagrams of halting a trial based on the decision made by: (A) interim data up to and including S_r , and (B) interim data up to including the penultimate look S_{r-1} .

$$A_r = P_{H_0} \left(\bigcup_{k=r}^K S_k \geq u_k \mid \bigcup_{\tilde{k}=1}^{r-1} S_{\tilde{k}} = s_{\tilde{k}} \right) = A(r)$$

and let $\tilde{A}_r$ denote the conditional error rate given $S_1 = s_1, \dots, S_{r-1} = s_{r-1}$ for the halted design with final critical value $\tilde{u}_r$, that is

$$\tilde{A}_r = P_{H_0} \left(S_r \geq \tilde{u}_r \mid \bigcup_{\tilde{k}=1}^{r-1} S_{\tilde{k}} = s_{\tilde{k}} \right).$$

The type I error can be inflated if the trial is halted if $s_1, \dots, s_{r-1}$ is such that $\tilde{A}_r > A_r$ and continues otherwise. The degree of this type I error rate inflation is demonstrated in Section 7 below. By the conditional error principle of Müller and Schäfer,²⁰ the type I error rate will be controlled at level α even if the decision to halt at look r is based on the data $s_1, \dots, s_{r-1}$ if the critical value for the final analysis is u_r^c where

$$P_{H_0} \left(S_r \geq u_r^c \mid \bigcup_{\tilde{k}=1}^{r-1} S_{\tilde{k}} = s_{\tilde{k}} \right) = A_r. \quad (10)$$

The adjusted critical boundary, u_r^c , depends on the observed data $s_1, \dots, s_{r-1}$, but since using u_r^c controls the overall type I error rate at level α , will be less conservative than using the boundary value obtained using the original spending function, u_r . It is worth mentioning that once u_r^c is obtained, we compare s_r with it to see whether the null hypothesis is rejected or not at the halting stage, r .

5 | EARLY HALTING OF GROUP SEQUENTIAL DESIGNS WITH NO MAXIMUM SAMPLE SIZE

In the setting of an emerging disease there can be a great deal of uncertainty about specifying the treatment effect at which power should be set, about nuisance parameters that affect sample size calculations, or about the likely achievable sample size for a clinical trial. Whilst having a smaller-than-expected patient cohort, especially amongst the most severely impacted individuals, is desirable, this uncertainty imposes challenges in the planning and execution of clinical trials that rely on a predetermined number of patients to achieve conclusive outcomes within a timeline determined by assumptions about recruitment rates. The RECOVERY trial described in Section 2 illustrates the difficulty in specifying the sample size in such a setting. A group sequential approach, in which the trial stops as soon as sufficient evidence is obtained to

reach a conclusion, is an attractive option in this case, though without specification of a maximum sample size it is not clear how the spending function approach could be implemented.

The sequential probability ratio test (SPRT) developed by Wald³ and Barnard⁴ is a sequential testing procedure of the null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1 (\theta_1 > \theta_0)$ based on the likelihood ratio, L_k obtained at a series of interim analyses, $k = 1, 2, \dots$. The test continues until $L_k < A$ or $L_k > B$ for some A and B with the null hypothesis rejected if $L_k > B$ or accepted if $L_k < A$. A test with no binding lower futility boundary can be obtained by taking $A = -\infty$. Note that unlike the group sequential tests introduced above, this test has no maximum sample size. In many settings this is considered a disadvantage. Under the uncertainty of an emerging disease, however, this allows the test to continue until either sufficient evidence is obtained to reject the null hypothesis or a decision is made to halt the study.

The test statistics $S_1, S_2, \dots$ with distribution given by (1) have the same distribution as points on a Brownian motion with drift θ at times $V_1, V_2, \dots$. We will write this Brownian motion as $S(t), t > 0$. In this case the test with no lower bound, that is with $A = -\infty$ has type I error rate α when $B = \alpha^{-1}$. This SPRT of the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta = \theta_1$ with $\theta_1 > 0$ stops at look k with the null hypothesis rejected if

$$S_k > \frac{-\log \alpha}{\theta_1} + \frac{V_k \theta_1}{2}. \quad (11)$$

For *continuous monitoring*, that is for the Brownian motion, $S(t)$, when the increments in information, $V_{k+1} - V_k$ tend to zero, the SPRT given by (11) has type I error rate α . For discrete monitoring the test is conservative. With equally spaced information values, with $V_{k+1} - V_k = I$, Siegmund²² proposes a modification of the best boundary to use

$$S_k > \frac{-\log \alpha}{\theta_1} + \frac{V_k \theta_1}{2} - 0.583 I^{1/2} \quad (12)$$

to give a less conservative test with type I error rate approximately equal to α .

For discrete monitoring when information values at interim values are not equally spaced, it is possible to construct critical values corresponding to the SPRT based on a modification of the α -spending function approach described in Section 3.2. Siegmund²² shows that for the test defined by 11, the probability of stopping the test and rejecting the null hypothesis with information less than V is

$$1 - \Phi(bV^{-1/2} - \theta^*V^{1/2}) + e^{2\theta^*b}\Phi(-bV^{-1/2} - \theta^*V^{1/2})$$

where $\theta^* = \theta - \theta_1/2$, $b = \log(B)/\theta_1 = -\log \alpha/\theta_1$ and Φ denotes the standard normal distribution function. Under the null hypothesis with $\theta = 0$ this becomes

$$1 - \Phi(\theta_1 V^{1/2}/2 + bV^{-1/2}) + \alpha \Phi(\theta_1 V^{1/2}/2 - bV^{-1/2}). \quad (13)$$

Denoting this function by $\tilde{\alpha}^*(V)$, given interim analyses at information levels $V_1, V_2, \dots$, we can obtain critical values $u_1, u_2, \dots$ to satisfy

$$P_{H_0} \left(\bigcup_{\tilde{k}=1}^k \{S_{\tilde{k}} \geq u_{\tilde{k}}\} \right) = \tilde{\alpha}^*(V_k), \quad (14)$$

or equivalently

$$P_k = \tilde{\alpha}^*(V_k) - \tilde{\alpha}^*(V_{k-1}). \quad (15)$$

This will allow an unlimited series of interim analyses with the type I error rate spent at each equal to that of the SPRT. We note that as $\tilde{\alpha}^*(V) \rightarrow \alpha$ as $V \rightarrow \infty$, the overall type I error rate is controlled at level α as desired.

If a trial that is planned as a SPRT is halted at look r , with information V_r , the design can be modified in one of the ways proposed in Sections 4.1 or 4.2. If the decision to halt the trial is independent of the observed data, all remaining type I error rate can be spent, with the final analysis using the critical value $\tilde{u}_r$ with

$$P_{H_0} \left(\bigcup_{k=1}^{r-1} \{S_k < u_k\}, S_r \geq \tilde{u}_r \right) = \alpha - \tilde{\alpha}^*(V_{r-1}). \quad (16)$$

If the decision to halt the trial can depend on observed value of $s_1, \dots, s_{r-1}$, a critical value for the final analysis, u_r^c , can be obtained to preserve the conditional error rate, that is with

$$P_{H_0} \left(S_r \geq u_r^c \mid \bigcup_{k=1}^{r-1} S_k = s_k \right) \quad (17)$$

equal to the conditional error rate for the SPRT given $S_1 = s_1, \dots, S_{r-1} = s_{r-1}$.

For continuous monitoring, the conditional error rate for the test given by 11 given $S_{r-1} = s_{r-1}$ is equal to

$$\alpha \exp \left[-\theta_1 (\theta_1 V_{r-1} / 2 - s_{r-1}) \right] \quad (18)$$

(see Appendix B), enabling the critical value u_r^c to be obtained from (17).

6 | EXAMPLE

In this section, we apply the discussed and proposed methods of adjusting critical boundaries for a halted trial to the RECOVERY-RS trial that was introduced in Section 2. The RECOVERY-RS was a randomised group sequential clinical trial of hospitalised adults with COVID-19 related acute hypoxic respiratory failure. The trial compared two experimental interventions, continuous positive airway pressure (CPAP) and high-flow nasal oxygen (HFNO), with conventional oxygen therapy. Depending on therapies available, patients were either randomised equally between all three arms or between one active arm and the control. For the latter randomisation case, each of the CPAP and HFNO group was compared with those control patients who could have received this treatment. As this leads to some control patients being excluded from each comparison, the total planned sample size was inflated to allow for this. Here we will focus on the comparison of CPAP with conventional therapy. The primary endpoint was the composite outcome of either mortality or the need for tracheal intubation within 30 days.

Whitehead¹³ gives expressions for test statistics for a group sequential trial comparing two treatments in terms of a binary outcome. Let p_A and p_C denote, respectively, success probabilities for patients randomly allocated to either intervention arm, A, or standard therapy, C. Defining θ to be the log-odds ratio as a measure of treatment difference, under the null hypothesis, $H_0 : p_A = p_C$, the S_k score statistic and estimated information level $V_k, k = 1, \dots, K$, at the k^{th} interim analysis are

$$S_k = \frac{n_{C,k}x_{A,k} - n_{A,k}x_{C,k}}{(n_{A,k} + n_{C,k})} \quad (19)$$

$$V_k = \frac{n_{A,k}n_{C,k}(x_{A,k} + x_{C,k}) \left[(n_{A,k} + n_{C,k}) - (x_{A,k} + x_{C,k}) \right]}{(n_{A,k} + n_{C,k})^3},$$

where $n_{A,k}$ and $n_{C,k}$ are the cumulative number of patients allocated to arms A and C respectively and $x_{A,k}$ and $x_{C,k}$ are the cumulative number of the success responses from arms A and C at the stage k^{th} , respectively, with $x_{A,k} \sim \text{Bin}(n_{A,k}, p_A)$ and $x_{C,k} \sim \text{Bin}(n_{C,k}, p_C)$. Based on the general theory proposed by Jennison and Turnbull¹² the vector of test statistics $(S_1, \dots, S_K)^T$ asymptotically follows multivariate normal distribution with $E(S_k) = \theta V_k$, $\text{Var}(S_k) = V_k$ and $\text{Cov}(S_k, S_{k'}) = V_k; 1 \leq k \leq k' \leq K$ as in (1). Note that basing stopping boundaries on the observed values $V_1, \dots, V_K$, automatically adjusts the sample size to maintain power for different values of the true success probability.

The study was designed as a group-sequential trial on the assumption of 12 equally space interim analyses. In consultation with the clinical trial team, an α -spending function of the form $\alpha^*(t) = \alpha t$ with $\alpha = 0.025$. $\alpha^*(t) = 0.025t$ was used, though other forms could be appropriate in other settings.¹⁸ To give the desired power of at least 0.9 to detect an improvement in the rate of the composite outcome of either mortality or intubation from 0.15 on the control arm to 0.1 on each experimental arm using a one-sided level 0.025 test, the maximum required information was calculated at the level of 57.26, corresponding to approximately 1000 patients per arm. This was inflated to a recruitment target of 1,334 patients

per arm, that is 4,002 patients overall, to allow for drop-outs and for control patients who were included in only one of the comparisons.

Although interim analyses were planned to be taken with equal increments in information, in practice, three unequally spaced interim analyses were conducted prior to the study being halted and a final analysis being performed. The information observed at the first two interim analyses was less than that for the first originally planned interim look, whilst the information observed at the third interim analysis was between that of the ninth and tenth originally planned interim analyses. The trial was halted after one year with a recruitment rate lower than anticipated. At this point, patients had been included in the trial to enable 377 patients in the CPAP group to be compared with 356 in the conventional oxygen therapy group. The decision to halt the trial was taken by the Trial Steering Committee (TSC) who had not seen any interim analysis data.

Table 1 summarises the data obtained from the RECOVERY-RS trial at each interim analysis. The information and test statistics were calculated based using Equation (19). The table expresses the test statistic in both score statistics, S_k , and the standardised value $Z_k = \frac{S_k}{\sqrt{V_k}}$. Stopping boundaries obtained using the α -spending approach via the Equation (5) allowing for the information available at each interim analysis are shown by the plus signs in Figure 2, part (A). The observed test statistics are shown by the stars and indicate that the critical boundary was not exceeded at any of the first three interim analyses. At the final analysis, after the trial had been halted, the observed test statistic, as given in Table 1 is 14.726. The critical value for this analysis obtained using the original spending function is $u_4 = 14.053$. As indicated above, the use of this critical value would be conservative since the decision to halt the trial was taken independent of the observed data. The alternative critical value obtained from spending all remaining α , $\tilde{u}_4 = 13.208$, was therefore used in the trial analysis. This is shown by the circle in Figure 2, part (A). In this case as the observed test statistic is above $\tilde{u}_r$, the conclusion is that the null hypothesis can be rejected with the trial indicating a statistically significant effect of CPAP compared to standard care.

Although in this case the decision to halt the trial was made without any information on the data observed at the interim analyses already conducted, had the decision been influenced by these data, as discussed above, the type I error

TABLE 1 Summary of the observed data on CPAP and standard care (SC) for the RECOVERY-RS trial.

Interim look	CPAP (s_k, n_k)	SC (s_k, n_k)	V_k	S_k	Z_k
First	(12, 22)	(11, 23)	2.8	0.756	0.45
Second	(19, 38)	(15, 28)	4.027	-0.576	-0.287
Third	(83, 126)	(81, 122)	13.883	-0.323	-0.087
Fourth	(240, 377)	(198, 356)	44.032	14.726	2.22

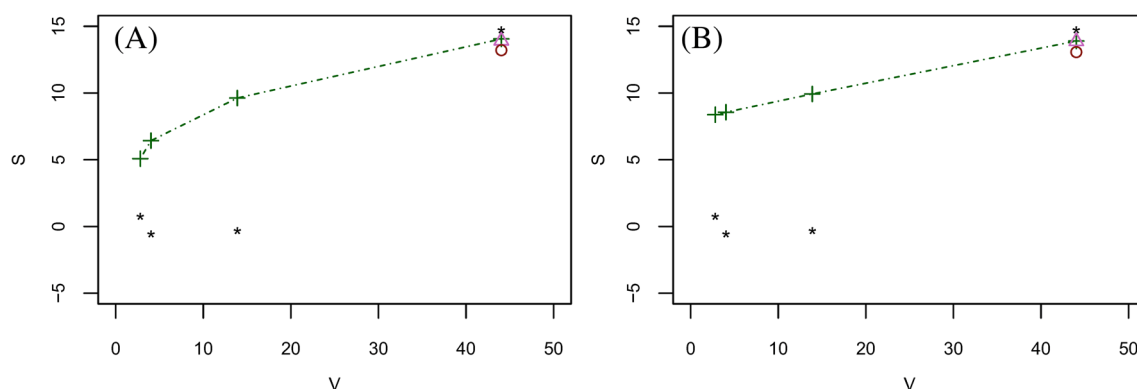


FIGURE 2 Observed test statistic values (*) for the RECOVERY-RS trial together with critical values (+) constructed using the spending function $\alpha^*(t) = at$ for (A) and sequential probability ratio test (SPRT) for (B), and alternative critical values for the final analysis when the trial was halted obtained by either spending all remaining α , $\tilde{u}_4$, using Equation (8) for (A) and (16) for (B) denoting by (O), or based on the conditional error approach, u_4^c , using Equation (10) for (A) and (17) for (B) showing by ($\triangle$).

rate could be inflated if the critical value $\tilde{u}_4$ was used. If this was the case the critical boundary value, u_4^c , obtained using the conditional error approach could alternatively be used. In this case $u_4^c = 13.936$. This is shown by the triangle in Figure 2, part (A). The value of u_4^c depends on the observed data in the penultimate interim look, that is, the third interim analysis, but in this case falls between the critical value for the original test, u_4 , and that obtained by spending all remaining α , $\tilde{u}_4$, though here the observed test statistic would have led to rejection of the null hypothesis irrespective of whether $\tilde{u}_4$, u_4^c or u_4 was used as the critical value at the final analysis.

In order to illustrate the approach proposed in Section 5, we also present a group-sequential design that could have been used for the RECOVERY-RS trial without specification of a maximum sample size using the sequential probability ratio test with one-sided type I error rate 0.025 and power 0.9 at $\theta_1 = \delta$ corresponding to an improvement in the composite outcome from 0.15 to 0.1 as in the original trial. Critical values obtained using Equation (14) are shown by the plus signs in Figure 2, part (B). As above, when the trial is halted alternative critical values for the final analysis can be obtained by spending all remaining α , ($\tilde{u}_4$), and the conditional error approach, (u_4^c), using the Equations (16) and (17) respectively. Critical boundary values at final interim analysis are $u_4 = 13.897$, $\tilde{u}_4 = 13.066$ and $u_4^c = 13.869$. These are shown by the plus sign, circle, and triangle respectively in the figure. The observed score statistics are shown by the stars in the figure. The test statistic at the final analysis would indicate a statistically significant difference between CPAP and the standard therapy if this design had been used for any proposed critical boundaries.

Comparison of parts (A) and (B) in Figure 2 show how the critical values for the test with spending function $\alpha^*(t) = \alpha t$ differ from those for the sequential probability ratio test. It can be seen that critical values for the SPRT are much greater than those for the spending function test at early interim analyses. The triangles in the left-hand side column of plots in Figure C1 show the spending function for the SPRT given by Equation (13), illustrating how the type I error rate is spent in the SPRT (dashed lines) in comparison with a linear spending function (solid lines). It can be seen that the type I error spent at early in the SPRT is much smaller than that for the linear spending function, leading to the higher critical values observed. Later in the trial the critical values are much more similar.

7 | SIMULATION STUDY

In this section, the properties of the methods proposed above are explored through a simulation study. We consider a two-arm group-sequential trial to compare an experimental treatment with a standard therapy with ten equally spaced interim analyses simulating $S_1, \dots, S_{10}$ from the multivariate normal distribution given by (1). As in the example above, we use a group sequential test with linear spending function with one-sided type I error rate of 0.025 and 90% power at $\delta = 0.5$ and a SPRT with type I error rate 0.025 and $\theta_1 = \delta$ as explained in Section 5. In the simulation study we assume that interim analyses occur as planned for the group-sequential trial with linear spending function, and the time on which the trial is halted coincides with one of the planned interim analysis. In turn, we evaluate the type I error rate and power by utilising three different strategies in choosing stopping boundary at the halt time: originally planned boundaries, spending all remaining α , $\tilde{u}_r$, and the conditional error approach, u_r^c , at the halting time r .

We initially consider trials halted independent of the data observed, with three distributions for the halting times. In the first case we take the halting time to be at each of the K interim analyses including the final one with equal probability, 1/10. In order to simulate earlier and later halt times we take the probability of stopping at each of the ten analyses to be given by (0.35, 0.25, 0.2, 0.1, 0.05, 0.03, 0.02, 0, 0, 0) and (0, 0, 0, 0.02, 0.03, 0.05, 0.1, 0.2, 0.25, 0.35) respectively.

In contrast to the cases in which halting times are randomly generated independent of the data, we also considered a data-dependent halting rule deliberately constructed to demonstrate that the type I error rate can be inflated using the critical boundaries obtained with all remaining α spent. With this rule, the decision to halt the trial at interim analysis r ($r > 1$) is based on the comparison of the two conditional error rates mentioned in Section 4.2.2; $\tilde{A}_r > A_r$, calculated using the observed data at interim analysis, $r - 1$, so that the trial is halted or not so as to maximise the conditional error rate. Table 2 gives simulated type I error rate and power values with the trial halted using each of the rules described above with the final critical value given by u_r , $\tilde{u}_r$ or u_r^c if the trial is halted. Each estimated probability is based on 100,000 simulated datasets giving a simulation standard error of 0.0005 so that estimated type I error rates are accurate to three decimal places.

When the halting time is independent of the interim data regardless of whether linear spending function or SPRT approach is used, as we show in Section 4.1, the overall type I error rate is maintained at or below the nominal level for all approaches. As would be expected, the test is most conservative, and the power lowest, when the original boundary value is used and least conservative, and most powerful, when all remaining α is spent in the final analysis if the trial is

TABLE 2 Simulated type I error rate and power obtained from utilising a linear spending function and SPRT approaches in designing the trial where the proposed methods in adjusting the critical boundary at the halt time is used.

Linear spending function		Halting time			
		Early	Equal probability	Late	Data-dependent
Originally planned boundaries	Type I error	0.0063	0.0138	0.0212	0.0084
	Power	0.198	0.525	0.816	0.2560
Spending all remaining α	Type I error	0.0258	0.0255	0.0252	0.0353
	Power	0.38	0.6457	0.8538	0.5678
Conditional error approach	Type I error	0.0257	0.0256	0.0252	0.0245
	Power	0.365	0.6342	0.8511	0.4787
SPRT					
Originally planned boundaries	Type I error	0.0045	0.0129	0.0196	0.0057
	Power	0.183	0.521	0.8	0.2160
Spending all remaining α	Type I error	0.0257	0.0255	0.0254	0.0356
	Power	0.384	0.6503	0.857	0.5564
Conditional error approach	Type I error	0.0253	0.0253	0.0249	0.0242
	Power	0.369	0.635	0.847	0.4656

Note: We assume a one-sided test with ten equally spaced interim analyses with a power of 90% and $\alpha = 0.025$.

halted. In all cases the power is below the planned level due to the early halting of the study, with the power lowest when the trial is halted earlier and the sample size is smallest.

When the halting time can depend on the interim data using the rule described above the type I error rate is substantially inflated above the nominal level if the critical value, $\tilde{u}_r$, obtained spending all remaining α is used. As expected, the type I error rate is controlled in this case when the critical value u_r^c , based on the conditional error approach, is used.

Details of the simulated probabilities under the null and alternative hypotheses of rejecting the null hypothesis at each interim analysis are given in Appendix C.

8 | DISCUSSION

In an emerging disease there may be considerable uncertainties regarding the number of infected patients, potential treatment effect, and natural history of the disease. These uncertainties give rise to difficulty in specifying an appropriate sample size in the planning stage of a clinical trial. In this case, adaptive sequential designs can be attractive, as they allow the accumulating data to be analysed at several interim analyses so that the trial can stop for either futility or efficacy as soon as the data can support such a decision. A reduction in the number of infected patients due to a waning epidemic, however, lead to a necessity to halt the trial before a specified sample size is attained or a pre-defined stopping boundary is crossed in the sequential testing procedure. In this paper we have investigated methods for the analysis of such halted trials, with a particular focus on controlling the type I error rate.

The methods we propose distinguish between settings in which the decision to halt a trial are made with no knowledge of the trial data and when the decision-maker has observed some data giving an indication of the treatment effect so that the influence of these data on the decision cannot be ruled out. In the former case the type I error is controlled by a spending function approach in which all remaining α is spent in the final analysis if the trial is halted without a stopping boundary being crossed. If such an approach is taken in the latter case the type I error rate can be inflated. In this case we propose using a critical value for the final analysis obtained so as to maintain the conditional error rate given the observed data at the same level as if the original group-sequential design, that is without halting, was used.

A slightly different setting to that considered arises if the decision-maker(s) responsible for deciding whether or not the trial should be halted have not seen efficacy data but have seen safety data from earlier interim analyses. In this case, as the efficacy and safety data may be correlated, an approach that ignores the influence of these data may not control the

type I error rate. A conditional error approach conditioning on the efficacy data from those patients for whom safety data are observed should control the type I error rate and be conservative in this case. Typically this will be the efficacy data from the interim analysis preceding the decision to stop the study, though this may not be the case if there is a delay in observing efficacy data with no delay, or a shorter delay, in the observation of safety data,²³ or if a time-to-event efficacy endpoint is used that may be correlated with earlier observed safety data.²⁴

The problem of the analysis of a halted group-sequential trial that we have considered is essentially that termed *under-running* by Whitehead.²¹ The analysis approach proposed by Whitehead in this setting is based on construction of a corrected p-value, but is equivalent to the approach we propose of spending all remaining α . The problem of underrunning due to the trial being halted based on a decision influenced by the data is not, however, considered by Whitehead other than to state that the methods that he proposes would then be inappropriate.

If the trial is halted on the basis of data available at the time at which the decision is taken, that is in the setting of Section 4.2.1 above, further follow-up data may be observed after the trial is halted. Inclusion of these data in the final analysis is discussed by Whitehead²¹ and Hampson and Jennison.²⁵ In the setting of Section 4.2.2, where only data from earlier interim analyses can be used in the decision to halt the trial, the final analysis can be made when full follow-up data are available.

In the work above only the upper boundaries for the group-sequential tests have been considered. In practice a lower boundary can be used to allow stopping for futility. A number of approaches have been proposed for the construction of this futility boundary,²⁶ for example based on the estimated treatment effect, the value of a test statistic for testing the null hypothesis,^{3,13} the conditional power²⁷⁻²⁹ or Bayesian alternatives,³⁰ or the value of a test statistic for testing the alternative hypothesis.³¹ While it is possible to adjust the upper boundary to control the type I error rate allowing for the lower boundary¹² this means that the lower boundary is *binding* in that the type I error rate can be inflated if the trial is not stopped when the lower boundary is crossed. If greater flexibility over stopping for futility is desired, the upper boundary can be obtained assuming no lower boundary will be used. This ensures type I error control since stopping for futility does not lead to rejection of the null hypothesis and so cannot give a type I error. The approach can, however, lead to conservatism and a resulting loss of power if a futility stopping rule is then used. The inclusion of a simple futility stopping rule in the approaches described above would be straightforward.

We have proposed the use of the sequential probability ratio test (SPRT)³ in order to allow a group-sequential trial to be conducted without any maximum sample size being specified, implementing this via a modified spending function approach. A similar approach could be used replacing the spending function for the SPRT with any increasing function that approaches α as $V \rightarrow \infty$. Different forms of such spending functions will have different properties in terms of the power and the distribution of the sample size under different true treatment effect values. The optimality properties of the SPRT³² may make this an attractive choice if halting is unlikely, though depending on the anticipate distribution of the halting time other spending functions might lead to smaller expected sample sizes for the same power. Further research or more extensive simulation studies could be conducted to explore such designs.

ACKNOWLEDGEMENTS

The authors are grateful to two anonymous referees for their helpful comments. This work was supported by the Medical Research Council (MRC)—grant numbers: MR/V038419/1, MC_UU_00002/14 and MC_UU_00040/03.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Amin Yarahmadi  <https://orcid.org/0009-0003-7435-6632>

Lori E. Dodd  <https://orcid.org/0000-0002-3433-5429>

Thomas Jaki  <https://orcid.org/0000-0002-1096-188X>

Peter Horby  <https://orcid.org/0000-0002-9822-1586>

Nigel Stallard  <https://orcid.org/0000-0001-7781-1512>

REFERENCES

1. Thorlund K, Dron L, Park J, Hsu G, Forrest J, Mills E. A real-time dashboard of clinical trials for COVID-19. *Lancet Digital Health*. 2020;2:e286.

2. Stallard N, Hampson L, Benda N, et al. Efficient adaptive designs for clinical trials of interventions for COVID-19. *Stat Biopharmaceut Res*. 2020;12(4):483-497.
3. Wald A. *Sequential Analysis*. 1st ed. New York: John Wiley and Sons; 1947.
4. Barnard GA. Sequential tests in industrial statistics. *J R Stat Soc Suppl*. 1964;8(1):1-26.
5. The RECOVERY collaborative group. RECOVERY trial website. www.recoverytrial.net
6. The RECOVERY collaborative group. Dexamethasone in hospitalized patients with Covid-19. *New Engl J Med*. 2021;384:693-704.
7. The RECOVERY collaborative group. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet*. 2021;397:1637-1645.
8. The RECOVERY collaborative group. Casirivimab and imdevimab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. medRxiv. doi:[10.1101/2021.06.15.21258542v1](https://doi.org/10.1101/2021.06.15.21258542v1)
9. The RECOVERY collaborative group. RECOVERY trial protocol. <https://www.recoverytrial.net/files/protocol-archive/recovery-protocol-v1-0-2020-03-18-final.pdf> 2020.
10. Perkins GD, Ji C, Connolly BA, et al. Effect of noninvasive respiratory strategies on intubation or mortality among patients with acute hypoxemic respiratory failure and COVID-19: the RECOVERY-RS randomized clinical trial. *Jama*. 2022;327(6):546-558.
11. Perkins GD, Couper K, Connolly B, et al. RECOVERY- respiratory support: respiratory strategies for patients with suspected or proven COVID-19 respiratory failure; continuous positive airway pressure, high-flow nasal oxygen, and standard care: a structured summary of a study protocol for a randomised controlled trial. *BMC Trials*. 2020;21:687. doi:[10.1186/s13063-020-04617-3](https://doi.org/10.1186/s13063-020-04617-3)
12. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: CRC Press; 1999.
13. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester: John Wiley & Sons; 1997.
14. Jennison C, Turnbull BW. Group-sequential analysis incorporating covariate information. *J Am Stat Assoc*. 1997;92(440):1330-1341.
15. Lan GKK, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663.
16. Proschan MA, Follmann DA, Waclawiw MA. Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*. 1992;48(4):1131-1143.
17. Kim K, Demets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*. 1987;74(1):149-154.
18. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. *Stat Med*. 1990;9(12):1439-1445.
19. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995;51(4):1315-1324.
20. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Stat Med*. 2004;23(16):2497-2508.
21. Whitehead J. Overrunning and underrunning in sequential clinical trials. *Control Clin Trials*. 1992;13(2):106-121.
22. Siegmund D. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer Science & Business Media; 1985.
23. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Stat Med*. 2010;29(9):959-971.
24. Bauer P, Posch M. Letter to the editor. *Stat Med*. 2004;23:1333-1334.
25. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *J R Stat Soc Series B Stat Methodology*. 2013;75(1):3-54.
26. Chang Y, Song T, Monaco J, Ivanova A. Futility stopping in clinical trials, optimality and practical considerations. *J Biopharm Stat*. 2020;30(6):1050-1059.
27. Lan KG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Sequent Anal*. 1982;1(3):207-219.
28. Lan KG, Wittes J. The B-value: a tool for monitoring data. *Biometrics*. 1988;44(2):579-585.
29. Lachin JM. A review of methods for futility stopping based on conditional power. *Stat Med*. 2005;24(18):2747-2764.
30. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Control Clin Trials*. 1986;7(1):8-17.
31. He P, Lai TL, Liao OYW. Futility stopping in clinical trials. *Stat Interface*. 2012;5(4):415-423.
32. Wald A, Wolfowitz J. Optimum character of the sequential probability ratio test. *Ann Math Stat*. 1948;19(3):326-339.

How to cite this article: Yarahmadi A, Dodd LE, Jaki T, Horby P, Stallard N. Group sequential designs for clinical trials when the maximum sample size is uncertain. *Statistics in Medicine*. 2024;43(24):4667-4683. doi: [10.1002/sim.10203](https://doi.org/10.1002/sim.10203)

APPENDIX A. CALCULATION OF TYPE I ERROR RATE SPENT AT A GIVEN HALT TIME WHEN USING ORIGINALLY PLANNED BOUNDARIES OR SPENDING ALL REMAINING α WHEN THE HALT TIME IS INDEPENDENT OF INTERIM DATA

Consider a group sequential trial with up to K looks, halted at interim analysis r , with the halt time is independent of the data collection procedure. Let the probability that $r = i$ be denoted by p_i , $i = 1, \dots, K$, with $\mathcal{P} = (p_1, \dots, p_K)$.

Assuming the interim data are compared with the originally planned boundary value, u_r , at the halt time, it is easy to show by mathematical induction that the probability under the null hypothesis of rejecting the null hypothesis by look k is

$$\sum_{i=1}^{k-1} \alpha^*(t_i) p_i + \alpha^*(t_k) \left(1 - \sum_{i=1}^{k-1} p_i \right). \quad (\text{A1})$$

Setting $k = K$ and noting that $p_1 + \dots + p_K = 1$, gives the overall type I error rate to be $\sum_{i=1}^K \alpha^*(t_i) p_i$ which, since $\alpha^*(t_i)$ is increasing in t_i and hence in i is equal to $\alpha^*(K) = \alpha$ when $p_1 = \dots = p_{K-1} = 0$ and $p_K = 1$, that is when there is no early halting, and is less than $\alpha^*(K) = \alpha$, showing that the test is conservative, for any other $p_1, \dots, p_K$.

In the case of comparing interim data with spending all remaining α boundaries, $\tilde{u}_r$, at the halt time, it can be similarly shown that the probability under the null hypothesis of rejecting the null hypothesis by look k is

$$\alpha^*(t_k) \left(1 - \sum_{i=1}^k p_i \right) + \alpha \sum_{i=1}^k p_i. \quad (\text{A2})$$

Setting $k = K$ in this case gives overall type I error rate α irrespective of the choice of $\mathcal{P}$.

APPENDIX B. DERIVATION OF CONDITIONAL ERROR RATE FOR CONTINUOUSLY MONITORED SEQUENTIAL PROBABILITY RATIO TEST

Consider the SPRT with upper bound only given by (11) in the main text, and suppose that this is continuously monitored. That is, we have a Brownian motion with drift θ , which will be denoted $S(t)$, $t > 0$, and we stop and reject the null hypothesis, $H_0 : \theta = 0$ if

$$S(t) > \frac{-\log \alpha}{\theta_1} + \frac{t\theta_1}{2}. \quad (\text{B1})$$

We wish to evaluate the conditional type I error rate given $S(V_1) = s_1, \dots, S(V_{r-1}) = s_{r-1}$ for given $V_1, \dots, V_{r-1}$, that is

$$P_{H_0} \left(S(t) > -\log \alpha / \theta_1 + t\theta_1 / 2, \text{ some } t > V_{r-1} | S(V_1) = s_1, \dots, S(V_{r-1}) = s_{r-1} \right). \quad (\text{B2})$$

Since $S(t)$ is a Markov process, the conditional distribution of $S(t)$, $t > V_{r-1}$ given $S(V_1) = s_1, \dots, S(V_{r-1}) = s_{r-1}$ only depend s_{r-1} .

Write $t' = t - V_{r-1}$ and let $S'(t')$ denote $S(t) - s_{r-1}$, $t > V_{r-1}$, then $S'(t')$ is also a Brownian motion with drift θ and the probability given by (B2) is equal to

$$P_{H_0} \left(S'(t') > -s_{r-1} - \log \alpha / \theta_1 + (t' + V_{r-1})\theta_1 / 2, \text{ some } t' > 0 \right). \quad (\text{B3})$$

This is the type I error rate for a test in which H_0 is rejected if

$$S'(t') > -s_{r-1} - \log \alpha / \theta_1 + (t' + V_{r-1})\theta_1 / 2$$

that is

$$S'(t') > -\log \alpha' / \theta_1 + t'\theta_1 / 2 \quad (\text{B4})$$

where

$$\frac{-\log \alpha'}{\theta_1} = \frac{-\log \alpha}{\theta_1} + \frac{V_{r-1}\theta_1}{2} - s_{r-1}$$

or equivalently

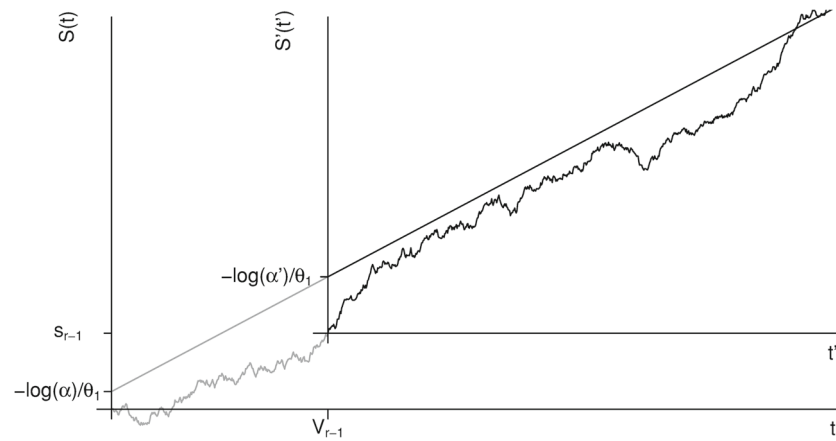


FIGURE B1 Boundary for original SPRT and for test conditional on $S_{r-1} = s_{r-1}$ with $-\log \alpha' / \theta_1 = -\log \alpha / \theta_1 + V_{r-1} \theta_1 / 2 - s_{r-1}$.

$$\log \alpha' = \log \alpha + ((V_{r-1}/2 - s_{r-1})\theta_1)$$

so that

$$\alpha' = \alpha \exp(-\theta_1(V_{r-1}\theta_1/2 - s_{r-1})).$$

Figure B1 illustrates a possible realisation of $S(t)$ and $S'(t')$ relative to the stopping boundary given by (B1). Since under the null $S'(t')$ is a Brownian motion with zero drift and (B4) defines a SPRT, this probability given by (B3), and hence by (B2) is equal to $\alpha' = \alpha \exp(-\theta_1(V_{r-1}\theta_1/2 - s_{r-1}))$ as given by (18) in the main text.

APPENDIX C. ADDITIONAL SIMULATION RESULTS

Figure C1 shows simulated probabilities under the null and alternative hypotheses for the probability of rejecting the null hypothesis by look k for $k = 1, \dots, K$ for the designs simulated in Section 7 are shown in Figure C1. The left-hand side column of plots in the figure give probabilities under the null hypothesis, showing the type I error rate spent at each look, in trials with early halting time (A), halting with equal probability at each look (B), and late halting time (C) and the right-hand side column gives similar results under the alternative hypothesis, showing the power. The results for trials designed using a linear spending function and SPRT approach are denoted by solid and dashed lines respectively in all plots. The triangles in the left-hand side column of plots show the linear spending function $\alpha^*(t) = \alpha t$ and spending function for the SPRT given by Equation (13). The type I error rates (left-hand side column) and the power (right-hand side column) of using different critical values for the halting analysis time are shown by $\circ$ for originally planned boundaries, u_r , $+$ for spending all remaining α , $\tilde{u}_r$, and ∇ for the conditional error approach, u_r^c .

The simulated results shown in the left hand panels of Figure C1 are in agreement with the probabilities given by expressions (A1) and (A2) with $\mathcal{P} = (0.35, 0.25, 0.2, 0.1, 0.05, 0.03, 0.02, 0, 0, 0)$, $\mathcal{P} = (1/10, \dots, 1/10)$ or $\mathcal{P} = (0, 0, 0, 0.02, 0.03, 0.05, 0.1, 0.2, 0.25, 0.35)$.

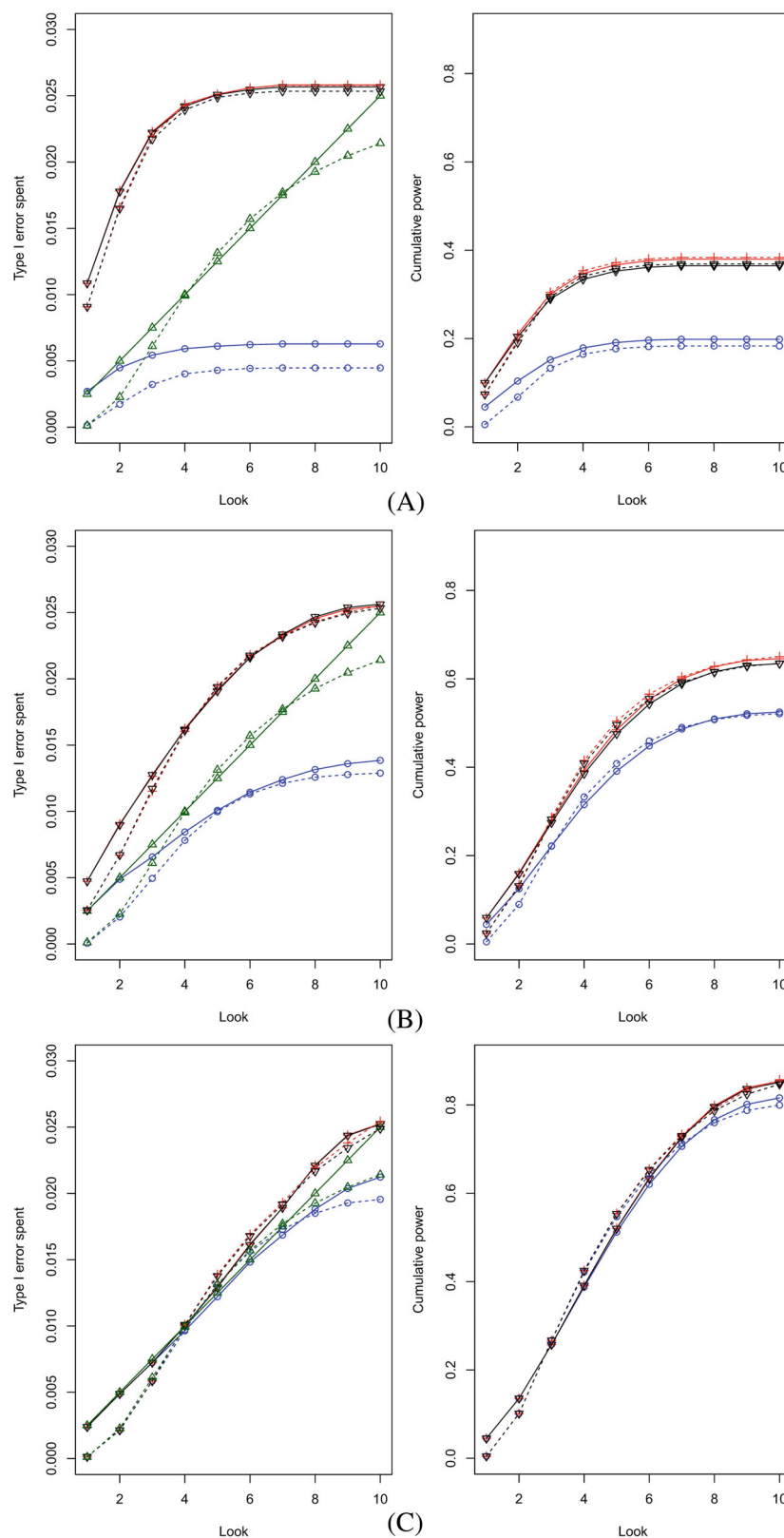


FIGURE C1 Comparison of type I error rate (heft-hand side column) and power (right-hand side column) in trials with early halting time (plots in A), halting with equal probability at each look (plots in B), and late halting time (plots in C).