Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R

Tim Friede1 | Nigel Stallard2 | Nicholas Parsons2

1Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany
2Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK

Correspondence
Tim Friede, Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
Email: tim.friede@med.uni-goettingen.de

Funding information
Medical Research Council, Grant/Award Number: G1001344

Abstract
Adaptive seamless designs combine confirmatory testing, a domain of phase III trials, with features such as treatment or subgroup selection, typically associated with phase II trials. They promise to increase the efficiency of development programmes of new drugs, for example, in terms of sample size and/or development time. It is well acknowledged that adaptive designs are more involved from a logistical perspective and require more upfront planning, often in the form of extensive simulation studies, than conventional approaches. Here, we present a framework for adaptive treatment and subgroup selection using the same notation, which links the somewhat disparate literature on treatment selection on one side and on subgroup selection on the other. Furthermore, we introduce a flexible and efficient simulation model that serves both designs. As primary endpoints often take a long time to observe, interim analyses are frequently informed by early outcomes. Therefore, all methods presented accommodate interim analyses informed by either the primary outcome or an early outcome. The R package asd, previously developed to simulate designs with treatment selection, was extended to include subgroup selection (so-called adaptive enrichment designs). Here, we describe the functionality of the R package asd and use it to present some worked-up examples motivated by clinical trials in chronic obstructive pulmonary disease and oncology. The examples both illustrate various features of the R package and provide insights into the operating characteristics of adaptive seamless studies.

KEYWORDS
adaptive design, clinical trials, closed test procedure, combination test, Dunnett test

1 | INTRODUCTION

There is a long history of application of sequential methods in clinical trials to allow the monitoring of accumulating data at a series of interim analyses (Jennison & Turnbull, 1999; Whitehead, 1997). Whilst most early work focussed on the aim of
stopping the trial as soon as sufficient evidence has been obtained, this body of work has rapidly expanded to include the use of interim data for other design adaptations, including sample size reestimation (Friede & Kieser, 2006; Proschan, 2009). Over the past years, there has been considerable interest in using interim analyses for selection of treatments, with less effective treatments dropped from the study, or for selection of patient subgroups, with recruitment following an interim analysis limited to subgroup(s) in which a promising effect is indicated (Bauer, Bretz, Dragalin, König, & Wassmer, 2016; Pallmann et al., 2018).

A major statistical challenge in the development of such methods is the control of the type I error rate when adaptations are made on the basis of data that will also be included in the final analysis. This can be achieved by the combination test approach of Bauer and Köhne (1994), which yields flexible designs for treatment selection (see, e.g., Bauer and Kieser, 1999; Posch et al., 2005; Bretz, Schmidli, König, Racine, and Maurer, 2006) and subgroup selection (see Brannath et al., 2009; Jenkins, Stone, and Jennison, 2011; Wassmer and Dragalin, 2015).

Earlier work has assumed that the adaptations would be informed by the pre-specified primary outcome, which is also used for hypothesis testing. From a practical perspective, however, this can be a strong limitation. In particular, in chronic diseases clinically meaningful endpoints might take some time to observe, which means that most or all patients are recruited by the time the primary outcome is observed for the first patients. This is illustrated, for example, by Chataway et al. (2011) in the context of secondary progressive multiple sclerosis. As a consequence, adaptations need to be based on early outcomes for adaptive designs to be feasible in these situations. Therefore, some adaptive seamless designs have been extended to allow the use of short-term endpoint data for decision-making at interim while the pre-specified primary endpoint is used for hypothesis testing (Friede et al., 2011; Friede, Parsons, & Stallard, 2012; Jenkins et al., 2011; Kunz, Friede, Parsons, Todd, & Stallard, 2014, 2015; Stallard, 2010; Stallard, Kunz, Todd, Parsons, & Friede, 2015).

It is acknowledged that these more complex designs require intensive simulation studies in the planning to evaluate their operating characteristics (Benda, Branson, Maurer, & Friede, 2010; Friede et al., 2010). A limitation to the use of adaptive methods in practical applications is often the availability of software to enable construction and evaluation of appropriate study designs and to conduct the final analysis. A number of commercial software packages including ADDPLAN and EAST are available for this purpose. Although some R packages for group-sequential designs including including gsDesign and gscounts and adaptive group-sequential designs such as rpkg are available from CRAN, there is still a shortage of comprehensive freely available software for adaptive seamless designs with treatment or subgroup selection, with the exception of the R package asd developed to plan clinical trials with treatment selection (Parsons et al., 2012).

Using the combination test approach, the aim of this paper is to present, to our knowledge for the first time, designs for treatment and subgroup selection in a unified notation and to present an efficient simulation model for their evaluation. By expressing the subgroup selection problem in a similar setting to that of treatment selection, the R package asd, originally developed for treatment selection, was extended to include designs of both types, that is, with subgroup or treatment selection. The methods implemented are based on the combination testing approach with at most two design stages. The designs obtained are fully flexible, controlling the type I error rate for any data-driven adaptation including treatment or subgroup selection as well as sample size adaptation.

Although the methods we propose can be based on $p$-values obtained from either summary statistics or individual patient data, in order to provide a general and efficient simulation model, we base this on the simulation of standardized sufficient statistics. The simulation model assumes multivariate normal distributions for the test statistics, but not for the individual observations. Therefore, the simulation model is not only widely applicable but the simulations are also fast. Furthermore, early outcomes informing the interim decisions are incorporated, since this is often important in practice as explained above. The application of the methods using the R package asd will be illustrated by clinical trials in chronic obstructive pulmonary disease (COPD) and oncology with treatment and subgroup selection, respectively.

2 | METHODS

2.1 | Notation and hypotheses

We consider first the setting of treatment selection designs. The study is conducted in up to $J \geq 2$ stages. In the first stage, patients are randomized between $K$ experimental treatments and a control treatment. In a general setting, suppose that observations of the pre-specified primary outcome from treatment group $k$, $k = 0, \ldots, K$, where $k = 0$ corresponds to the control treatment, have a distribution depending on some parameter $\mu_k$, and that it is desired to test the family of null hypotheses $H_k : \theta_k = 0$, $k = 1, \ldots, K$, where $\theta_k = \mu_k - \mu_0$. Let $p_{kj}$ denote a $p$-value for the test of the null hypothesis $H_k$ based on the data from patients.
first observed in stage $j$. These data may not be available at the time of the interim analysis following stage $j$, but become available only later on in the trial (Friede et al., 2011). As we will explain in Section 2.2 below, the interim decisions need not necessarily be based on the primary outcome but could, in principle, make use of any available outcome while still maintaining control of the type I error rate.

Next we consider the setting of subgroup selection designs. Suppose that a single experimental treatment is to be compared with a control treatment but that patients can be categorized as belonging to one or more predefined subgroups. Interest is focused on the treatment effect in the full population and each subgroup, in particular in the subgroup with the largest treatment effect.

As part of the aim of this paper is to make explicit the similarities between methodology for treatment and subgroup selection designs, we will use similar notation for both settings when this does not cause confusion. Thus suppose that observations come from patients in $K$ subgroups (including the full population as a special case) labeled $k, k = 1, \ldots, K$. Suppose that data for patients in subgroup $k$ receiving treatment $r, r = 0, 1$ have a distribution depending on some parameter $\mu_{k,r}$. Setting $\theta_k = \mu_{k,1} - \mu_{k,0}$, it is desired to test the family of hypotheses $H_k : \theta_k = 0, k = 1, \ldots, K$. As above, $p_{kj}$ denotes a $p$-value for the test of null hypothesis $H_k$ based on the data from patients with an outcome first observed in stage $j$.

### 2.2 Interim selection rules

Although in both the treatment and subgroup selection settings, the testing strategies described here control the type I error rate for any selection rule, it is good practice to specify selection rules in advance to enable calculation of operating characteristics including sample size and power. In the following, we introduce some possible examples of interim selection rules based on the primary outcome, though these could equally be applied to an early outcome as we will explain below.

One obvious way to proceed would be to select the treatment or subgroup which performs best in terms of some statistic $Z_{j,k}$ (where $Z_{j,k}$ is some estimator of $\theta_k$ following stage $j$). However, this might not be wise in situations where sample sizes are relatively small given the differences between the treatments or subgroups, since there is a rather high risk of picking some treatment or subgroup that is not optimal. The so-called $e$-rule proposed by Kelly et al. (2005) selects all treatments or subgroups with statistics $Z_{j,k}$ for which $Z_{j,k} \geq \max_j Z_{j,j} - e$ (assuming that larger values of $Z_{j,k}$ are better). For $e = 0$, this rule reduces to selecting the maximum only. For large $e$, no selection takes place as all treatments or subgroups are carried forward. Otherwise varying numbers of treatments or subgroups are carried forward into the next stage. For practical applications, it is advisable to study the operating characteristics of the design for a range of values for $e$. Comments on the optimal choice of $e$ are provided in Section 4 of Friede and Stallard (2008). The authors conclude that “selecting only the best treatment is a simple selection rule which is often optimal or close to optimal” (Friede & Stallard, 2008).

Multi-arm studies including several doses of an experimental drug motivate another selection rule. In some indications, it is not uncommon to select not one but two doses for confirmatory testing in phase III. The COPD study discussed in more detail in Section 5 is a good example for this. A more generalized version of this rule would be to select the best $K^*$ out of $K$ treatments or subgroups where $K^*$ would be specified in advance.

The selection of treatments or subgroups in interim analyses could be informed by the primary outcome or, if this is not feasible, by an early outcome. In situations where an early outcome informs an interim adaptation and the primary endpoint is used for hypothesis testing, we refer to the primary endpoint also as the final outcome to distinguish it more clearly from the early outcome. The early outcome need not necessarily fulfill all requirements of a surrogate endpoint (Burzykowski, Molenberghs, & Buyse, 2005) as weaker conditions might suffice. Chataway et al. (2011) used the phrase of a “biologically plausible” outcome that “gives some indication as to whether the mechanism of action of a test treatment is working as anticipated”. Nevertheless, for the operating characteristics of the adaptive seamless design the correlation between the early and the final outcomes on an individual patient level as well as the treatment effects on both the early and the final outcome (population level) are relevant as we will see below.

### 2.3 Error rate control via the closed testing procedure

In either the treatment selection or the subgroup selection setting, the problem has been posed in such a way that it is desired to test the null hypotheses $H_k : \theta_k = 0, k = 1, \ldots, K$. It is desirable to conduct these hypotheses tests so as to control the familywise error rate in the strong sense, that is to control the probability of rejection of any true null hypothesis within this family, at some specified level, $\alpha$. Strong error rate control may be achieved through a closed testing procedure in which, denoting by $H_{\bar{k}}$ the
intersection hypothesis \( \cap_{k \in \mathcal{K}} H_k \), all hypotheses \( H_\mathcal{K} \) for \( \mathcal{K} \subseteq \{1, \ldots, k\} \) are tested at nominal level \( \alpha \), and \( H_k \) rejected if and only if \( H_\mathcal{K} \) is rejected at this level for all \( \mathcal{K} \ni k \) (Marcus, Peritz, & Gabriel, 1976).

Application of the closed testing procedure requires a test of the intersection hypothesis \( H_\mathcal{K} \) for each \( \mathcal{K} \subseteq \{1, \ldots, k\} \). These hypotheses tests must also combine evidence from the different stages in the trial. This may be achieved through the use of a combination testing method, as described in detail in the next subsection.

### 2.4 The combination testing method and early stopping

Although the treatment or subgroup selection might be informed by an early outcome, hypothesis testing is for the primary outcome. Extending the notation introduced above, let \( p_{\mathcal{K}j} \) denote a \( p \)-value for a test of the null hypothesis \( H_\mathcal{K} \) based on data from patients with an outcome first observed at stage \( j = 1, \ldots, J \). By construction, under \( H_\mathcal{K} \), \( p_{\mathcal{K}j} \sim U[0, 1] \), or if a conservative test is used, \( p_{\mathcal{K}j} \) is stochastically larger than or equal to \( U[0, 1] \), for all \( j \) and \( \mathcal{K} \) (Branath, Posch, & Bauer, 2002). We also assume that the conditional distribution of \( p_{\mathcal{K}j} \) given \( p_{\mathcal{K}1}, \ldots, p_{\mathcal{K}j-1} \) is stochastically no smaller than a \( U[0, 1] \) for all \( p_{\mathcal{K}1}, \ldots, p_{\mathcal{K}j-1} \) for all \( j \), which is also referred to as the \( p \)-cloud condition (Branath et al., 2002). The condition is satisfied if the \( p \)-values from different stages are independent. In practical applications, the \( p \)-cloud condition is only satisfied asymptotically and referred to as asymptotically \( p \)-cloud (Branath et al., 2009). In the following, we assume that all \( p \)-values are (at least asymptotically) \( p \)-cloud.

The \( p \)-values from the different stages can be combined using a number of combination functions (Bauer & Köhne, 1994; Lehmacher & Wassmer, 1999) to give test statistics \( C_\mathcal{K}(p_{\mathcal{K}1}, \ldots, p_{\mathcal{K}J}) \) which, under the assumptions above regarding the distributions of the \( p \)-values, have known distributions under \( H_\mathcal{K} \) irrespective of adaptations made to the study design. These test statistics may thus be used to test hypothesis \( H_\mathcal{K} \) (Bauer & Kieser, 1999; Bretz et al., 2006).

Although a number of combination functions have been proposed, here we use the inverse normal combination function (Lehmacher & Wassmer, 1999), which is equivalent to the method of Cui, Hung, and Wang (1999). This gives test statistics

\[
C_\mathcal{K} = \sum_{j=1}^{J} w_j \Phi^{-1}(1 - p_{\mathcal{K}j}) \quad \text{where} \quad w_1, \ldots, w_J \quad \text{are specified in advance and the sum of their squares is equal to 1},
\]

that is,

\[
\sum_{j=1}^{J} w_j^2 = 1.
\]

Given the distributional assumptions under \( H_\mathcal{K} \), \( C_\mathcal{K} \) are distributed as, or are stochastically no larger than, a multivariate normal distribution with mean zero, \( \text{var}(C_\mathcal{K}) = w_1^2 + \cdots + w_J^2 \) and \( \text{cov}(C_{\mathcal{K}j}, C_{\mathcal{K}j'}) = w_1^2 + \cdots + w_{\min(j,j')}^2 \).

Following Posch et al. (2005), we assume that hypotheses cannot be rejected once they are dropped, resulting in conservative tests. Furthermore, applying the closed testing principle outlined in Section 2.3 in an adaptive design the \( p \)-value \( p_{\mathcal{K}j} \) is replaced by \( p_{\mathcal{K}j \cap \mathcal{J}_j} \) where \( \mathcal{J}_j \) is the set of hypotheses carried forward into stage \( j \) (Posch et al., 2005).

If interim analyses are used for adaptation of the design but not for stopping the trial, a final test of \( H_\mathcal{K} \) may be based on \( C_{\mathcal{K}J} \). More generally, a sequential test of \( H_\mathcal{K} \) may be conducted based on the joint distribution of \( C_{\mathcal{K}1}, \ldots, C_{\mathcal{K}J} \), rejecting \( H_\mathcal{K} \) if \( C_{\mathcal{K}j} \geq \gamma_{\mathcal{K}j} \) for some critical values \( \gamma_{\mathcal{K}j} = (u_{\mathcal{K}j1}, \ldots, u_{\mathcal{K}j})^T \). To simplify the notation, \( u_{\mathcal{K}j} \) and \( u_{\mathcal{K}j} \) will generally be written as \( u \) and \( u \) when it is clear which hypothesis is tested. The single constraint that the overall error rate should be at most \( \alpha \) is insufficient to determine \( u \) uniquely. A common approach in sequential analysis is to specify the type I error to be spent at each interim analysis, and to find \( u_{\mathcal{K}1}, \ldots, u_{\mathcal{K}J} \) to satisfy

\[
\text{pr}_{H_\mathcal{K}}(C_{\mathcal{K}j'} \geq u_{\mathcal{K}j'}, \text{ some } j' \leq j) = \alpha_{j'},
\]

where \( \alpha_{1} = \cdots = \alpha_{J} = \alpha \) are either specified in advance (Slud & Wei, 1982) or depend on the observed information in some predetermined way (Lan & DeMets, 1983). Critical values \( u_{\mathcal{K}1}, \ldots, u_{\mathcal{K}J} \) satisfying (1) can be found recursively with \( u_{\mathcal{K}1} \), found directly from the joint distribution of \( C_{H_{\mathcal{K}1}, \ldots, C_{H_{\mathcal{K}J}}} \) via a numerical search once \( u_{\mathcal{K}1}, \ldots, u_{\mathcal{K}J-1} \) are known. Computational details are given by, for example, Jennison and Turnbull (1999).

Construction of the sequential test statistics \( C_{\mathcal{K}1}, \ldots, C_{\mathcal{K}J} \) requires specification of \( p \)-values \( p_{\mathcal{K}1}, \ldots, p_{\mathcal{K}J} \) for testing \( H_\mathcal{K} \). For elementary hypotheses, that is when \( |\mathcal{K}| = 1 \), \( p_{\mathcal{K}1} \) can be obtained from a standard test, such as a \( t \)-test for normally distributed data or a chi-squared test for binary data. When \( |\mathcal{K}| > 1 \), \( p_{\mathcal{K}j} \) should be calculated so as to allow for the multiple comparisons implicit in testing \( H_\mathcal{K} \). In the treatment selection setting, as comparisons are with a common control, a Dunnett test (Dunnett, 1955) may be used. In the subgroup selection setting, the simple Bonferroni procedure, Simes’ procedure (Branath et al., 2009) or the Spiessens–Debois test (Spiessens & Debois, 2010), a Dunnett-type test with a generalized covariance structure, may be used. In each case, the level of adjustment depends on the size of \( \mathcal{K} \).
3 | SIMULATION MODEL

As eluded to in Section 1, simulations are often required in the planning phase of a study to choose design options such as sample sizes or interim selection rules. Here, we propose a simulation model that is efficient in the sense that population statistics are generated rather than individual patient data. Therefore, computation times do not increase with larger sample sizes.

In a wide variety of settings, it is possible to obtain statistics that are, at least asymptotically, normally distributed with known variance. When a series of interim analyses is conducted, these statistics follow a multivariate normal distribution with non-zero correlations, since data obtained at earlier stages are also used at later ones. In the setting of treatment or subgroup selection, these multivariate normal distributions may be extended to give the joint distribution of statistics corresponding to different treatment comparisons or for treatment effects in different subgroups. To be clear, these statistics are not actually necessarily the test statistics used to test the hypotheses but are rather some estimators of $\hat{\theta}_k$. Similarly, the distributional forms assumed in the simulations need not be used as the basis of hypotheses tests, since these, as described in Section 2.4, can use the combination testing approach.

When considering adaptations to the trial design based on the observation of short-term endpoint data, it is helpful to further extend the simulation models to include test statistics calculated based on different endpoints. The distributions of the resulting test statistics are described briefly in this section. Additional detail is given in Appendix.

3.1 | Treatment selection

Considering treatment selection designs, let $\hat{\theta}_{kj}$ denote the estimate of the treatment effect, $\theta_k$, for treatment $k$ based on the data available at the $j$th interim analysis, and $I_{kj}^{-1}$ denote the variance of this estimate. As these estimates are often, at least asymptotically normally distributed, our model will be based on an assumption of this distributional form. Assume that $I_{kj}$ does not depend on $k$, and so may be denoted by $I_j$, and that the correlation between different treatment comparisons with a common control group is $1/(1+\lambda)$, as will often be the case if we have 1:1 randomization. Setting $S_{kj} = \hat{\theta}_{kj} I_{kj}$, we then have

$$(S_{11}, \ldots, S_{K1}, \ldots, S_{1J}, \ldots, S_{KJ})' \sim N(I \otimes \theta, \Sigma^{(GS)}(I) \otimes \Sigma^{(CS)}_k \left(\frac{1}{1+\lambda}\right)),$$

where $I$ and $\theta$ denote, respectively, the vectors $(I_1, \ldots, I_J)'$ and $(\theta_1, \ldots, \theta_K)'$ and

$$\Sigma^{(CS)}_k(r) = \begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r \\ r & \cdots & r & 1 \end{pmatrix}$$

and

$$\Sigma^{(GS)}(I)^{(1)} = \begin{pmatrix} I^{(1)}_1 & I^{(1)}_1 & \cdots & I^{(1)}_1 \\ I^{(1)}_1 & I^{(1)}_2 & \cdots & I^{(1)}_2 \\ \vdots & \vdots & \ddots & \vdots \\ I^{(1)}_1 & I^{(1)}_2 & \cdots & I^{(1)}_J \end{pmatrix}$$

denote variance matrices of the $K \times K$ complex symmetric form and of that obtained in the usual group-sequential setting.

3.2 | Subgroup selection

In the subgroup selection setting, let $\hat{\theta}_{kj}$ denote the estimate of the treatment effect, $\theta_k$, for subgroup $k$ based on the data available at the $j$th interim analysis. Again, assume that these estimates are normally distributed, let $I_{kj}^{-1}$ denote the variance of this estimate and assume that $I_{kj}$ does not depend on $k$, and so may be denoted by $I_j$. In the simplest case, in which $K = 2$ and subgroup 2 is the whole population and subgroup 1 a proportion of size $\tau$, setting $S_{kj} = \hat{\theta}_{kj} I_{kj}$, we have

$$(S_{11}, S_{21}, \ldots, S_{1J}, S_{2J})' \sim N(I \otimes \theta, \Sigma^{(GS)}(I) \otimes \Sigma^{(GS)} \left(\frac{\tau}{1}\right)).$$
Extensions to the case when a short-term endpoint is also considered, and to more complex subgroup selection settings are discussed in the Appendix.

4 | SOFTWARE IMPLEMENTATION IN R

4.1 | R package asd

The simulation models for two-stage treatment selection and subgroup selection designs, described in Section 3, can be implemented in the R (R Development Core Team, 2009) package asd (Parsons et al., 2012), which is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/package=asd). This package comprises a number of functions that allow the properties of seamless phase II/III clinical trial designs, potentially using early outcomes, for treatment or subgroup selection to be explored and evaluated prior to a study commencing. An earlier version of this package, without the extension to subgroup designs, enhanced options for a wider range of outcome measures or more complete output model description was described previously by Parsons et al. (2012). The general structure of the code comprises a set of base functions that implement lower level tasks such as hypothesis testing, treatment selection, and closed testing procedures. The base functions, which have been described previously in the setting of treatment selection designs (Parsons et al., 2012), have been modified in the latest version of the asd package to work with both subgroup and treatment selection designs. The higher level user facing function asd.sim has been replaced by two new functions treatsel.sim and subpop.sim that are called directly by the user to implement hypothesis testing and simulations. The more general functions, gtreatsel.sim and gsubpop.sim can in principle be called directly by the user, although due to the more complex input structure this is generally not recommended.

4.2 | Function subpop.sim

Friede et al. (2012) extended the previously described combination test (CT) approaches, for co-primary analyses in a single pre-defined subgroup and the full population, using the methods proposed by Spiessens and Debois (2010) to control the familywise error rate (FWER) in the subgroup and the full population, and also proposed a novel method to obtain a critical value for the definitive test using a conditional error function (CEF) approach; full details of these methods are given by Friede et al. (2012). The function subpop.sim implements all the methods for subgroup selection in adaptive clinical trials reported by Friede et al. (2012) and subsequent correspondence (Friede, Parsons, & Stallard, 2013). The authors described and explored the performance of a number of methods in the setting described here, distinguishing between two distinct approaches to control the FWER, a CT method (Brannath et al., 2009; Jenkins et al., 2011) and a CEF method.

4.2.1 | Input arguments

An overview and brief description of the input arguments available for subpop.sim is shown in Table 1. The CT methodology described by Friede et al. (2012) can be implemented in subpop.sim using either the Spiessens and Debois (2010) (SD), Simes or Bonferroni testing procedures to control the FWER. These and the CEF approach can be selected using the method argument to subpop.sim. The following options are available: (i) CT-SD (method="CT-SD"), (ii) CT-Simes (method="CT-Simes"), (iii) CT-Bonferroni (method="CT-Bonferroni"), and (iv) CEF (method="CEF").

The syntax providing the group sample sizes, for stages 1 and 2, for a putative trial design and effect sizes for early and final outcomes is consistent across the available outcome types and comprises a list of the selected options. For instance, a design where the required sample size per treatment arm is 100 for stage 1 and 200 for stage 2 would be implemented using the following expression, n=list(stage1=100,stage2=200). The assumption, in the current version of asd, is that the sample size in the control arm of the study is the same as in the treatment arm. The default setting is that if the subgroup only is selected at the interim analysis at stage 1, then the subgroup prevalence remains the same after stage 1. If an increase in the sample size in the subgroup was planned, if this group only was selected (enrichment), then this can be implemented by adding an additional item to the list to indicate this. For instance, if we wanted a sample size of 200 in the subgroup in this setting, irrespective of the subgroup prevalence, then we would modify the previous argument to the following n=list(stage1=100,stage2=200, enrich=200).

Effect sizes for early and final outcomes are also given as a list using expressions of the following structure, where the first element of each vector is the effect size in the subgroup and the second element is the effect size in the full population. Setting effect=list(early=c(0.3,0.1),final=c(0.3,0.1)) specifies an effect size of 0.3 in the subgroup and 0.1 in the full population; the effect size in the control group is set by default to be zero. The default setting is for normal outcomes for both early and final outcomes, outcome=list(early="N",final="N"), and the effects are interpreted given these options. The
### TABLE 1  Brief description and available input arguments to R functions `subpop.sim` and `treatsel.sim`

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>Sample sizes for each treatment group at stage 1 (interim) and stage 2 (final) analyses</td>
<td>List of sample sizes; for example, 32 in each group in stage 1 and 64 in stage 2, <code>list(stage1=32, stage2=64)</code></td>
</tr>
<tr>
<td><strong>effect</strong></td>
<td>Effect sizes for early and final outcomes</td>
<td>List of effect sizes for the control (first) and each treatment group; for example, for control (0) and two effects of size 0.1 and 0.2 for early and 0.2 and 0.3 for final outcome, <code>list(early=c(0,0.1,0.2), final=c(0,0.2,0.3))</code></td>
</tr>
<tr>
<td><strong>outcome</strong></td>
<td>Outcome type for early and final outcomes</td>
<td>List of outcome types, options for normal (N), time-to-event (T), and binary (B) are currently available; for example, normal for early and final outcomes <code>list(early=&quot;N&quot;, final=&quot;N&quot;)</code></td>
</tr>
<tr>
<td><strong>nsim</strong></td>
<td>Number of simulations</td>
<td>An integer $&lt; 1 \times 10^7$</td>
</tr>
<tr>
<td><strong>sprev</strong></td>
<td>Subgroup prevalence</td>
<td>The prevalence of the subgroup in the main population; $0 &lt; sprev &lt; 1$. Subgroup prevalence can be either fixed or allowed to vary at each simulation; default is <code>sprev.fixed=TRUE</code></td>
</tr>
<tr>
<td><strong>corr</strong></td>
<td>Correlation between outcomes</td>
<td>Correlation between early and final outcomes; $-1 &lt; corr &lt; 1$</td>
</tr>
<tr>
<td><strong>seed</strong></td>
<td>Seed number</td>
<td>Seed number to ensure repeatability of simulations</td>
</tr>
<tr>
<td><strong>select</strong></td>
<td>Method for treatment selection</td>
<td>Seven available options (see Section 4.3.1), with default, <code>select=0</code>, to select all treatments. For <code>select=4</code>, the epsilon rule, can be used to allow more flexibility, with default <code>epsilon=1</code> and for <code>select=6</code>, all treatments greater than a threshold can be selected, with default <code>thresh=1</code></td>
</tr>
<tr>
<td><strong>ptest</strong></td>
<td>Treatments for counts of the number of rejections</td>
<td>A vector of valid treatment numbers for treatment specific counts of rejections; for example, for four treatments, <code>ptest=c(1), ptest=c(1, 2)</code> or <code>ptest=c(1,2,3,4)</code> are all valid options</td>
</tr>
<tr>
<td><strong>method</strong></td>
<td>Methodology used for simulations; and for subpop.sim either</td>
<td>Either <code>method=&quot;invnorm&quot;</code> or &quot;fisher&quot; to select inverse normal or Fisher combination test</td>
</tr>
<tr>
<td><strong>fu</strong></td>
<td>Follow-up options</td>
<td>Subjects in the dropped treatment groups can followed-up (<code>fu=TRUE</code>), with default <code>fu=FALSE</code> (see Section 4.3.1.)</td>
</tr>
<tr>
<td><strong>weight</strong></td>
<td>Stage 1 weight</td>
<td>Stage 1 weight be defined with this option ($0 \leq weight \leq 1$); default <code>weight=NULL</code></td>
</tr>
<tr>
<td><strong>level</strong></td>
<td>Test level</td>
<td>Test level must be set ($0 \leq level \leq 1$); default <code>level=0.025</code></td>
</tr>
<tr>
<td><strong>file</strong></td>
<td>File name for output</td>
<td>If unset will default to R console. For example, set <code>file=&quot;output.txt&quot;</code>, direct output to file in R working directory (<code>getwd()</code>)</td>
</tr>
</tbody>
</table>
available options for outcome types are normal (N), time-to-event (T), and binary (B), and all combinations of these are allowed for early and final outcome measures. Generally, it is assumed that higher means (N) and lower event rates (B or T) are better. A detailed description of these options is left for the following section describing function treatsel.sim; the options described there are analogous to those available for subpop.sim. In the simpler setting where group selection is based purely on the final outcome, this can be implemented by setting the effect sizes for the early and final outcomes to be equal and the correlation between the early and final outcomes to one (i.e., corr=1).

The subgroup prevalence is set by a single argument, namely (sprev). For instance, sprev=0.5 indicates that the subgroup comprises half of the full population. The function subpop.sim randomly generates test statistics (with a seed number set using seed) and accumulates results from usually a large number of simulations that must be set using the nsim option (default setting, nsim=1000). The prevalence can be either fixed at the set value (sprev.fixed=TRUE) or allowed to vary (sprev.fixed=FALSE) using a single realization of the binomial random variate generation function rbinom, at the set values for the sample size and subgroup prevalence, at each simulation.

Subgroup selection at interim is implemented using the so-called threshold selection rule (Friede & Stallard, 2008; Friede et al., 2012) (select="thresh"). If for the difference Δ of the test statistics for the full population and the subgroup holds Δ ≤ l1, then the subgroup only is tested at the end of stage 2. If Δ > l2 the full population only is tested. Otherwise both subgroup and full populations are tested. The thresholds (l1, l2) are set using the argument selim, which is a vector of standard deviation multiples. If, for instance, large limits are set (e.g., selim=c(-10,10)) then both subgroup and full populations will always be tested at the trial endpoint, whereas if selim=c(0,0), only the test regarding the population with the largest test statistic at interim is taken into stage 2. Intermediate values for selim between these extremes provide more flexible selection options. The weight for the CT approaches, if unset, is given by nstage1/(nstage1 + nstage2) with ndage1 and ndage2 referring to the (full) population sizes at stage 1 and stage 2, respectively. The test level is set by default to 0.025 (level=0.025).

4.2.2 Output

The output to subpop.sim first gives a summary of the simulation model, including expected values for the test statistics at each stage of the study. The main summary table reports the number of times that hypotheses H0^S, H0^F, H0^[S,F] were selected for testing and rejected when the subgroup (S), the full population (F), or both were tested. Output from subpop.sim is by default directed to the usual R console, but to save more detailed summaries of the simulation model, output can be directed to a file using this as an argument to the file function (e.g., file="output.txt"). Section 5 shows how subpop.sim is used in a practical setting with example data.

4.3 Function treatsel.sim

Function treatsel.sim replaces and generalizes the previous function asd.sim, with the name change made to make it much more explicit that the code implements only simulations for treatment selection designs for multi-arm studies. Much of the syntax and model setup is consistent between treatsel.sim and subpop.sim.

4.3.1 Input arguments

An overview and brief description of the input arguments available for treatsel.sim is shown in Table 1. One aspect of the design setup that differs considerably between treatsel.sim and subpop.sim is the coding of the treatment effects. Treatment effect sizes are given for the control group \( \mu_0 \) and the test treatment or treatments \( \mu_k \) and as vectors for both early and final outcomes; for instance, effect=list(early=c(0,0.1,0.2,0.1), final=c(0,0.1,0.2,0.3)) indicates that there are three test treatments with effect sizes 0.1, 0.2, and 0.1 for the early outcome and 0.1, 0.2, and 0.3 for the final outcome, respectively; the control is set to 0 for both outcomes. The null hypotheses tested are \( H_k : \theta_k = \mu_k - \mu_0 = 0 \). There is no limit to the number of test treatment groups \( K \). However, in practice, our experience is that the code runs slowly for designs with eight or more treatment groups, that is, \( K \geq 8 \). The setting of the effect sizes can be clarified further by considering the available options for the outcome types (normal N, time to event T, and binary B), set using the option outcome, with the default being to have both early and final outcomes normal (outcome=list(early="N", final="N")), with all nine combinations available. Generally, it is assumed that higher means (N) and lower event rates (B or T) are better.

For normal outcomes, the test statistics for the simulation model for the \( K \) test treatments, relative to the control group, are given by \( \sqrt{n/2} \times (\mu_k - \mu_0) \). For time-to-event outcomes, effects are interpreted as minus log hazard rates, with the control \( \mu_0 \) set to zero and test statistics are given by \( \sqrt{o_k/4} \times (\mu_k - \mu_0) \), where the expected total number of events in the control and treatment groups \( o_k \) for treatment \( k \) is calculated under an assumed exponential model to be \( n \times (1 - \exp(-\exp(-\mu_0))) + n \times \)
Clinical trial in COPD with treatment selection

Output

The method argument to treatsel.sim allows either the inverse normal (invnorm) or Fisher’s (fisher) combination test and the logical follow-up argument (fu) determines whether (i) patients in the dropped treatment groups are removed from the trial and unknown test statistics in the dropped treatments are set to $-\infty$ at stage 2 (fu=FALSE) or (ii) patients are kept in the trial and followed-up to the final outcome, in the same manner as the patients recruited in stage 1 in the selected treatment groups (fu=TRUE); Friede et al. (2011) called option (i), the default setting, discontinued follow-up and option (ii) complete follow-up. Seven treatment selection rules based on stage 1 test statistics are available in treatsel.sim, and are chosen with the select argument: (i) select all treatments (select=0), (ii) select the maximum (select=1), (iii) select the maximum two (select=2), (iv) select the maximum three (select=3), (v) flexible treatment selection using the ε-rule (Friede & Stallard, 2008; Friede et al., 2011), with additional argument (epsilon) (select=4), (vi) randomly select a single treatment (select=5) or (vii) select all treatments greater than a threshold, with the additional argument (thresh) (select=6).

The only additional argument available for treatsel.sim, that has not been covered in the section describing subpop.sim, is ptest. This is a vector of valid treatment numbers for determining specific counts for the number of simulations that reject the null hypothesis; for instance, for three test treatments and $\text{ptest} = \{1,3\}$, treatsel.sim will count and report the number of rejections of one or both hypotheses for testing treatments 1 and 3 against the control, in addition to the number of rejections of each of the elementary hypotheses.

4.3.2 | Output

The output from treatsel.sim first gives a summary of the simulation model, including expected values for the test statistics at each stage of the study. The main summary tables report (i) the number of treatments selected at stage 1, (ii) treatment selection at stage 1, that is, how often each treatment was selected, (iii) counts of hypotheses rejected at study endpoint, for each of the elementary hypotheses ($H_{0}^{[1]}, H_{0}^{[2]}, \ldots, H_{0}^{[K]}$), and (iv) the number of times that one or more than one of the hypotheses identified in ptest are rejected. Section 5 shows how treatsel.sim is used in a practical setting with example data.

5 | EXAMPLES

In this section, we illustrate the methods described above by two example studies using the R package asd. The first is a multi-arm randomized controlled trial in COPD with treatment selection; which will be considered in Section 5.1. As a second example, we consider trials in oncology with time-to-event outcomes and subgroup selection; this will be considered in Section 5.2.

5.1 | Clinical trial in COPD with treatment selection

Barnes et al. (2011) and Donohue et al. (2010) report a seamless adaptive design with dose selection, which was also discussed elsewhere (Cuffe, Lawrence, Stone, & Vandemeulebroecke, 2014). Patients were randomized to four doses of indacaterol (75 μg, 150 μg, 300 μg and 600 μg), active controls, and placebo control. For the purpose of illustration, we ignore the active controls in the following and use the observed results to illustrate the design process. The primary outcome was the percentage of days of poor control over 26 weeks. As recruitment for the entire study took only 6 months, the interim treatment selection could not be based on the primary endpoint. Trough forced expiratory volume in 1 s (FEV1) at 15 days was identified as a suitable early outcome to inform the interim analysis. From Figure 1 of Barnes et al. (2011), difference in trough FEV1 compared to placebo at 15 days is 150 ml, 180 ml, 210 ml, and 200 ml for the indacaterol doses 75 μg, 150 μg, 300 μg, and 600 μg, respectively. Reported 95% confidence intervals suggest two standard errors of treatment difference is approximately 60 ml, so assuming equal stage 1 sample sizes $n_{1} = 110$, then the standard deviation of the measurements is approximately 220 ml. Standardized effect sizes are thus approximately 0.68, 0.82, 0.95, and 0.91 for the indacaterol doses 75 μg, 150 μg, 300 μg, and 600 μg, respectively. Regarding the final outcome days of poor control (%) over 26 weeks, we gather from the report of stage 2 results at http://clinicaltrials.gov/show/NCT00463567 that the placebo rate was 35%. Let us assume rates of 31%, 30%, 28%, and 29% for the four doses of indacaterol (75 μg, 150 μg, 300 μg and 600 μg). Based on reported standard errors, we estimate the standard deviation to be approximately 30%. Hence, the standardized effect sizes are approximately 0.13 (75 μg), 0.17 (150 μg), 0.23 (300 μg), and 0.20 (600 μg). The approximate sample sizes per arm were 100 patients in stage 1 and 300 patients in stage 2.
Since the aim was to select two doses of indacaterol at the interim analysis to take into stage 2, we consider an overall sample size for the two stage of $5 \times 100 + 3 \times 300 = 1,400$ patients. Furthermore, a moderate positive correlation between early and final outcomes of 0.4 is assumed.

In the following, we will consider three settings: (i) continuous early and final outcomes and selecting always two doses for the second stage, as in the original study; (ii) continuous early and final outcomes, as in the original study, but for the purpose of illustration, we vary the selection rule using the threshold rule to allow varying numbers of doses being taken forward for confirmatory testing in the second stage; (iii) as final outcomes are non-normal and early and final outcomes are on different scales, we assume in another setting that the final is binary and not continuous.

### 5.1.1 Continuous early and final outcomes: Selecting always two doses for the second stage

As an example, we consider a fixed total sample size of 1,400 patients, which we can allocate between the two stages, with either more or less resources for each stage. We consider the following combinations of $n_1 = 10, 25, \ldots, 175$ and $5 \times n_1 + 3 \times n_2 = 1,400$. Each one of these sample size options can be tested using the following implementation of the `treatsel.sim` function.
for the setting with 100 patients per arm in stage 1 and 300 patients per arm in stage 2:

treatsel.sim(n=list(stage1=100,stage2=300), effect=list(early=c(0,0.68,0.82,0.95,0.91), final=c(0,0.13,0.17,0.23,0.20)), outcome=list(early="N",final="N"), nsim=10000,corr=0.4,seed=145514,select=2, level=0.025,ptest=c(3,4))

This code sets the sample sizes for each stage, and provides the effect estimates as described above, for a normal early outcome ("N") and a normal ("N") final outcome. The correlation is set to 0.4, and the number of simulations to 10,000. The select=2 option implements the rule that chooses the two treatments with the largest test statistics an interim. The test level is set to 0.025 and the ptest options allows us to count rejections for either or both of the doses 300 µg and 600 µg. Results from running this code are as follows (omitting the parts describing the setup):

simulation of test statistics:
expectation early = 4.8 5.8 6.7 6.4
expectation final stage 1 = 0.9 1.2 1.6 1.4 and stage 2 = 1.6 2.1 2.8 2.4
weights: stage 1 = 0.5 and stage 2 = 0.87

number of treatments selected at stage 1:

<table>
<thead>
<tr>
<th>n</th>
<th>1 0 0.00</th>
<th>2 10000 100.00</th>
<th>3 0 0.00</th>
<th>4 0 0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>10000 100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

treatment selection at stage 1:

<table>
<thead>
<tr>
<th>n</th>
<th>1 383 3.83</th>
<th>2 3282 32.82</th>
<th>3 8661 86.61</th>
<th>4 7674 76.74</th>
</tr>
</thead>
</table>

hypothesis rejection at study endpoint:

<table>
<thead>
<tr>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
</tr>
</thead>
<tbody>
<tr>
<td>183</td>
<td>2067</td>
<td>7206</td>
<td>5541</td>
</tr>
<tr>
<td>1.83</td>
<td>20.67</td>
<td>72.06</td>
<td>55.41</td>
</tr>
</tbody>
</table>

reject H3 and/or H4 = 8469 : 84.69%

The first part of the output provides a summary of the model setup, and the second part values of the test statistics used in the simulations. The squared weights are calculated, in this case, as 100/400 and 300/400. The results are summarized in the three lower tables. The first indicates that two treatments were always selected at stage 1, the second gives the number of simulations in which each treatment was selected and the third gives the number of simulations in which the elementary hypotheses were rejected. The final statement gives the number of simulations in which at least one of the treatments picked using the ptest options were rejected. Assigning the output for this function to an object that we for the sake of illustration call simply output, then the summaries described here can be accessed directly, for instance, for plotting data or other analysis, using the syntax output$count.total, output$select.total, output$reject.total, and output$sim.reject.

The results of the simulations suggest that given the large treatment effects on the early outcome (trough FEV1 at 15 days), and the modest effects on the final outcome (days of poor control (%) over 26 weeks), the greatest power would have been achieved by using a sample size of around 50 per group in stage 1 (see Figure 1).

5.1.2 Continuous early and final outcomes: Selecting varying numbers of doses for the second stage

As an alternative to always selecting the best two performing treatments at interim, we now consider a threshold rule, where all treatments with test statistics at interim analysis above a fixed threshold are taken into stage 2. If no treatments reach the
threshold, the study is stopped for futility. Simulations are implemented using the same effect sizes as in setting 1, a stage 1 sample size of 40, and a stage 2 sample size of 400 in the following code for a threshold of 3:

\[
\text{treatsel.sim(n=list(stage1=40,stage2=400),}
\]
\[
\text{effect=list(early=c(0,0.68,0.82,0.95,0.91),}
\]
\[
\text{final=c(0,0.13,0.17,0.23,0.20)),}
\]
\[
\text{outcome=list(early="N",final="N"),}
\]
\[
\text{nsim=10000,corr=0.4,seed=145514,select=6,}
\]
\[
\text{thresh=3,level=0.025,ptest=c(3,4))}
\]

The `select=6` option implements the threshold rule, with the fixed early outcome test statistic threshold set using the `thresh` option. Results from running this code are as follows (omitting the parts describing the setup):

Simulation of test statistics:

expectation early = 3 3.7 4.2 4.1
expectation final stage 1 = 0.6 0.8 1.0 0.9 and stage 2 = 1.8 2.4 3.3 2.8
weights: stage 1 = 0.3 and stage 2 = 0.95

number of treatments selected at stage 1:

<table>
<thead>
<tr>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>800</td>
</tr>
<tr>
<td>2</td>
<td>1634</td>
</tr>
<tr>
<td>3</td>
<td>3098</td>
</tr>
<tr>
<td>4</td>
<td>4175</td>
</tr>
<tr>
<td>Total</td>
<td>9707</td>
</tr>
</tbody>
</table>

treatment selection at stage 1:

<table>
<thead>
<tr>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5083</td>
</tr>
<tr>
<td>2</td>
<td>7469</td>
</tr>
<tr>
<td>3</td>
<td>8914</td>
</tr>
<tr>
<td>4</td>
<td>8596</td>
</tr>
</tbody>
</table>

hypothesis rejection at study endpoint:

<table>
<thead>
<tr>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>2480</td>
</tr>
<tr>
<td>H2</td>
<td>4882</td>
</tr>
<tr>
<td>H3</td>
<td>7769</td>
</tr>
<tr>
<td>H4</td>
<td>6642</td>
</tr>
</tbody>
</table>

reject H3 and/or H4 = 8600 : 86%

Although the probability for futility stopping is not given explicitly, it can be easily derived. The considered design was stopped for futility only 3% (100% - 97.07%) of the time; with all four experimental treatments being taken into stage 2 more than 41% of the time. Running the above code for thresholds in the range 0 to 6 (at intervals of a half) and extracting output for each option gives the results summarized in Figure 2.

The expected overall sample sizes for the scenarios in Figure 2, based on the simulated number of treatments selected at stage 1 and the fixed stage 1 and stage 2 sample sizes of 40 and 400, are as follows: 2199.5, 2197.3, 2188.1, 2164.3, 2109.0, 1991.2, 1802.5, 1548.2, 1264.0, 1004.2, 807.9, 688.2, and 631.0. The power drops off rapidly as the threshold increases from 3 to 5, as futility stopping increases from 3% to 65%. When, on average, two test treatments are taken into stage 2, that is when the fixed threshold is somewhere between 3.5 and 4 (overall sample size between 1548.2 and 1264.0), the power is lower (between 78.7% and 65.4%) than the analogous setting in Figure 1 (86.6%). The early outcome effect sizes and distributions are such that a fixed threshold that on average picks two treatments at interim, also stops for futility so often that it reduces the power considerably compared to a design that always takes exactly two treatments.
5.1.3 | Continuous early and binary final outcome

The implementation described here has focused on normal outcome measures for both early and final outcomes. However, if one or other outcome were binary, the changes necessary to implement this new scenario are relatively straightforward. For instance, instead of using days of poor control over 26 weeks as the final outcome measure, we might use a threshold based on this outcome. If this was the case, then success or failure of the treatment could be determined for each study participant, based on some a priori threshold for the number of days that one might expect to maintain control over the 26 week period. For the sake of example, let us consider the failure rate to be 50% in the control group, and 45% (75 µg), 45% (150 µg), 40% (300 µg), and 40% (600 µg), at each dose, respectively. Given that every other aspect of the design is the same as the first setting in the COPD example, then this design can be implemented using the following code:

```r
# Hypotheses to be rejected
Hs = c(0.50, 0.45, 0.45, 0.40, 0.40)
Hf = c(0.6, 0.9)

# Early and final effect sizes
effect = list(early = c(0.6, 0.9), final = c(0.6, 0.9))
sprev = 0.3
outcome = list(early = "N", final = "B")
nsim = 10000
seed = 145514
select = 2
level = 0.025
ptest = c(3, 4)

treatsel.sim(n = list(stage1 = 100, stage2 = 300),
            effect = effect,
            sprev = sprev,
            outcome = outcome,
            nsim = nsim,
            seed = seed,
            select = select,
            level = level,
            ptest = ptest)
```

This gives an overall rejection probability of 76.99%, by changing the outcome argument for the treatsel.sim function of Section 5.1.2 to list(early = "N", final = "B") and the vector of final effect sizes to c(0.50, 0.45, 0.45, 0.40, 0.40).

5.2 | Clinical trials in oncology with subgroup selection

Jenkins et al. (2011) suggested designs for adaptive seamless phase II/III designs for oncology trials using correlated survival endpoints. Designs of this type can be implemented relatively straightforwardly using the subpop.sim function. Here, we explore some design properties for a typical scenario from amongst the many that Jenkins et al. (2011) explored. We assume early and final time-to-event outcomes with a hazard ratio of 0.6 in the subgroup and 0.9 in the full population for both, and a correlation between endpoints of 0.5; in the setting of an oncology trial, the endpoints might be progression free and overall survival. We set the stage 1 sample size to 100 patients per arm and the stage 2 sample size to 300 patients per arm, if we progress in the full population, and to 200 patients per arm if we progress in the subgroup only. The subgroup prevalence is fixed at 0.3. Using the futility rule for selection at interim, with limits for the subgroup and full population both set to 0, this scenario can be implemented using the following code:

```r
# Hypotheses to be rejected
Hs = c(0.50, 0.45, 0.45, 0.40, 0.40)
Hf = c(0.50, 0.45, 0.45, 0.40, 0.40)

# Early and final effect sizes
effect = list(early = c(0.6, 0.9), final = c(0.6, 0.9))
sprev = 0.3
outcome = list(early = "N", final = "B")
nsim = 10000
seed = 1234
select = "futility"
level = 0.025
method = "CT-SD"

treatsel.sim(n = list(stage1 = 100, stage2 = 300),
            effect = effect,
            sprev = sprev,
            outcome = outcome,
            nsim = nsim,
            seed = seed,
            select = select,
            level = level,
            method = method)
```

The method="CT-SD" option implements the combination test method with Spiessens and Debois testing procedure. Given test statistics at interim of $S_1$ and $S_2$ for the subgroup and the full population, respectively, and selection rule limits ($l_1$, $l_2$), the futility rule implements the following options: (i) continue with a co-primary analysis if $S_1 < l_1$ and $S_2 < l_2$, (ii) continue in the subgroup alone if $S_1 < l_1$ and $S_2 \geq l_2$, (iii) continue in the full population alone if $S_1 \geq l_1$ and $S_2 < l_2$, and (iv) stop for futility if $S_1 \geq l_1$ and $S_2 \geq l_2$. Results from running this code are as follows (omitting the parts describing the setup):

```
simulation of test statistics:
expectation early: sub-pop = -1.46 : full-pop = -0.58
expectation final stage 1: sub-pop = -1.46 : full-pop = -0.58
expectation final stage 2: sub-pop only = -3.76 : full-pop only = -1.01
expectation final stage 2, both groups selected: sub-pop = -2.52 : full-pop = -1.01
weights: stage 1 = 0.5 and stage 2 = 0.87

hypotheses rejected and group selection options at stage 1 (n):
   Hs  Hf  Hs+Hf  Hs+f  n  n%
sub  2225  0    0     2225  2309 23.09
full  0   48    0     54    227  2.27
both 5370 1658 1636  5407  6987 69.87
total 7595 1706 1636  7686  9523 -
%
75.95 17.06 16.36 76.86 95.23 -
reject Hs and/or Hf = 76.65%
```
TABLE 2  Selection probabilities, probability of futility stopping and power (probability of rejecting at least one elementary null hypothesis) for a range of values of \( l_s \) and \( l_F \)

<table>
<thead>
<tr>
<th>( l_F )</th>
<th>( l_s )</th>
<th>Selection (%)</th>
<th>Futility (%)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Subgroup</td>
<td>Full</td>
<td>Both</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>23.1</td>
<td>2.3</td>
<td>69.9</td>
</tr>
<tr>
<td>0</td>
<td>−1</td>
<td>11.4</td>
<td>16.2</td>
<td>55.8</td>
</tr>
<tr>
<td>0</td>
<td>−2</td>
<td>2.3</td>
<td>45.1</td>
<td>26.5</td>
</tr>
<tr>
<td>0</td>
<td>−3</td>
<td>0.1</td>
<td>66.0</td>
<td>6.0</td>
</tr>
<tr>
<td>−1</td>
<td>0</td>
<td>60.0</td>
<td>0.4</td>
<td>32.3</td>
</tr>
<tr>
<td>−1</td>
<td>−1</td>
<td>37.4</td>
<td>4.0</td>
<td>29.7</td>
</tr>
<tr>
<td>−1</td>
<td>−2</td>
<td>12.3</td>
<td>16.5</td>
<td>16.9</td>
</tr>
<tr>
<td>−1</td>
<td>−3</td>
<td>1.5</td>
<td>28.4</td>
<td>4.8</td>
</tr>
<tr>
<td>−2</td>
<td>0</td>
<td>84.9</td>
<td>0.0</td>
<td>7.4</td>
</tr>
<tr>
<td>−2</td>
<td>−1</td>
<td>60.1</td>
<td>0.3</td>
<td>7.2</td>
</tr>
<tr>
<td>−2</td>
<td>−2</td>
<td>24.1</td>
<td>2.4</td>
<td>5.6</td>
</tr>
<tr>
<td>−2</td>
<td>−3</td>
<td>4.4</td>
<td>5.3</td>
<td>2.1</td>
</tr>
<tr>
<td>−3</td>
<td>0</td>
<td>91.6</td>
<td>0.0</td>
<td>0.7</td>
</tr>
<tr>
<td>−3</td>
<td>−1</td>
<td>66.7</td>
<td>0.0</td>
<td>0.7</td>
</tr>
<tr>
<td>−3</td>
<td>−2</td>
<td>28.6</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>−3</td>
<td>−3</td>
<td>5.6</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

The output reports that in 75.95%, 17.06%, and 16.36% of the simulations, the null hypothesis was rejected in the subgroup, full population and both, respectively. The final two columns give a breakdown of the selections made at the interim analysis; 23.09% of the simulations were continued in the subgroup only, 2.27% in the full population only, 69.87% in both, and 4.77% were stopped for futility. A concise summary of this table can be obtained for further analysis, by assigning to an output object and accessing the results using the syntax `output$results`.

It is informative in understanding the futility rule to run the above code for a grid of futility rule limits in the range 0 to −3; the results of this for each option are summarized in Table 2, where the notation \( l_s \) and \( l_F \) indicate the limits for the subgroup and full population, respectively.

From Table 2, it is clear that as \( l_s \) becomes more negative, the subgroup is selected progressively less often and similarly as \( l_F \) becomes more negative, it is selected progressively more often. The balance between the two limits determines overall power in this setting. With a strong effect in the subgroup and a much weaker effect in the full population, the best strategy is to always, unless stopping for futility, test in the subgroup at the final analysis. For the grid of values tested here, this is best achieved when \( l_F = −3 \) and \( l_s = 0 \).

6 | DISCUSSION

Adaptive seamless designs are recognized as a tool to increase the efficiency of clinical development programs by combining features of learning and confirming in a single trial, while traditional development programs would have investigated these in separate trials. However, their implementation is more involved than traditional designs (see, e.g., Quinlan and Krams (2006) for a discussion). One aspect is the planning which is more complex, often requiring extensive Monte Carlo simulations (Benda et al., 2010; Friede et al., 2010). Here, we presented a unified framework for adaptive seamless designs with treatment or subgroup selection. Furthermore, we developed a flexible and yet efficient simulation model. This, as all other methods discussed, can accommodate interim selection informed by an early outcome rather than the final one. Furthermore, we demonstrate how the R package `asd`, freely available from CRAN, can be used to evaluate and compare operating characteristics of various designs.

Here, we employed the combination test approach to achieve type I error rate control. For treatment selection, alternative approaches build on the work of Thall, Simon, and Ellenberg (1988, 1989) using the group-sequential method, but require that a single treatment continues along with a control beyond the first stage (Stallard & Todd, 2003) or that the number of treatments at each stage is specified in advance (Stallard & Friede, 2008). Methods based on the combination test approach are more flexible (see, e.g., Bauer and Kieser (1999), Posch et al. (2005), and Bretz et al. (2006)), but may be less powerful in some settings.
The authors have declared no conflict of interest.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.
OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Tim Friede https://orcid.org/0000-0001-5347-7441
Nigel Stallard https://orcid.org/0000-0001-7781-1512
Nicholas Parsons https://orcid.org/0000-0001-9975-888X

REFERENCES


APPENDIX: DERIVATION OF DATA SIMULATION MODELS

We consider first treatment selection designs. Let $\hat{\mu}^{(i)}_{k,j}$ be the estimate for treatment $k$, $k = 0, 1, \ldots, K$ with $k = 0$ corresponding to control, at stage $j$, $j = 1, \ldots, J$, for endpoint $i$, $i = 1, 2$.

Assume, as often holds at least asymptotically, that $\hat{\mu}^{(i)}_{k,j}$ is normally distributed with

$$
\hat{\mu}^{(i)}_{k,j} \sim N(\mu_k^{(i)}, (\mathcal{F}^{(i)}_{k,j})^{-1}),
$$

with

$$
cov(\hat{\mu}^{(i)}_{k,j}, \hat{\mu}^{(i)}_{k',j'}) = 0, k \neq k',
$$

$$
cov(\hat{\mu}^{(i)}_{k,j}, \hat{\mu}^{(i)}_{k',j'}) = (\mathcal{F}^{(i)}_{k,\max\{j,j'\}})^{-1}
$$

and

$$
cov(\hat{\mu}^{(i)}_{k,j}, \hat{\mu}^{(i')}_{k,j'}) = \rho\sqrt{(\mathcal{F}^{(i)}_{k,\max\{j,j'\}})^{-1}(\mathcal{F}^{(i')}_{k,\max\{j,j'\}})^{-1}}, i \neq i'.
$$

For independent normal random variables with known variance $\sigma^2$, we have $\mathcal{F}^{(i)}_{k,\max\{j,j'\}} = \sigma^2/\hat{n}^{(i)}_{k,j}$ where $\hat{n}^{(i)}_{k,j}$ is the number of observations at look $j$ for treatment $k$ on endpoint $i$.

Let $\hat{\beta}^{(i)}_{k,j} = \hat{\beta}^{(i)}_{k,j} - \hat{\beta}^{(i)}_{0,j}$, $k = 1, \ldots, K$, $\hat{I}^{(i)}_{k,j} = (\mathcal{F}^{(i)}_{k,j})^{-1} + (\mathcal{F}^{(i)}_{0,j})^{-1}$ and $\hat{S}^{(i)}_{k,j} = \hat{\beta}^{(i)}_{k,j} \hat{I}^{(i)}_{k,j}$.

Let $\lambda^{(i)}_{k,j} = (\mathcal{F}^{(i)}_{k,j})^{-1}/(\mathcal{F}^{(i)}_{0,j})^{-1}$. If $\hat{\theta}^{(i)}_{k,j}$ is constant, say equal to $\lambda$, and $\hat{I}^{(i)}_{k,j} = \hat{I}^{(i)}_{k,j}$ for all $k$ (for a single sample of known-variance normals, this is equivalent to assuming that sample sizes are the same for all experimental treatments). In this case, considering first a single endpoint, we get

$$
\begin{pmatrix}
S^{(1)}_{11} & \cdots & \theta^{(1)}_{1,1} \\
\vdots & \ddots & \vdots \\
S^{(1)}_{K1} & \cdots & \theta^{(1)}_{K,1}
\end{pmatrix}
\sim N
\begin{pmatrix}
\theta^{(1)}_{1,1} & \cdots & 1 \\
\vdots & \ddots & \vdots \\
\theta^{(1)}_{K,1} & \cdots & 1
\end{pmatrix}, \Sigma
$$

(A1)
where the variance matrix is given by

\[
\Sigma = \begin{pmatrix}
I_1 \frac{1}{1+\lambda} & I_1(1) \frac{1}{1+\lambda} & \cdots & I_1(1) \frac{1}{1+\lambda} & I_1(1) \frac{1}{1+\lambda} & \cdots & I_1(1) \frac{1}{1+\lambda} \\
I_1(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
I_1(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} \\
I_1(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
I_1(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda} & I_2(1) \frac{1}{1+\lambda} & \cdots & I_2(1) \frac{1}{1+\lambda}
\end{pmatrix}
\]

Note that we can rewrite the right hand side of (A1) as

\[
N \left( I^{(1)} \otimes \theta^{(1)}, \Sigma^{(GS)}(I^{(1)}) \otimes \Sigma^{(CS)}_K \left( \frac{1}{1+\lambda} \right) \right)
\]

where \( I^{(1)} \) and \( \theta^{(1)} \) denote, respectively, the vectors \( (I_i^{(1)} \ldots, I_J^{(1)})' \) and \( (\theta_1^{(1)} \ldots, \theta_K^{(1)})' \) and

\[
\Sigma^{(CS)}_K(r) = \begin{pmatrix}
1 & r & \cdots & r \\
r & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & r \\
r & \cdots & r & 1
\end{pmatrix}
\]

and

\[
\Sigma^{(GS)}(I^{(1)}) = \begin{pmatrix}
I_1^{(1)} & I_1^{(1)} & \cdots & I_1^{(1)} \\
I_1^{(1)} & I_2^{(1)} & \cdots & I_2^{(1)} \\
\vdots & \vdots & \ddots & \vdots \\
I_1^{(1)} & I_2^{(1)} & \cdots & I_J^{(1)}
\end{pmatrix}
\]

denote variance matrices of the \( K \times K \) complex symmetric form and of that obtained in the usual group-sequential setting.

When both endpoints are considered, the vector

\[
(S_{11}^{(1)} \ldots, S_{K1}^{(1)} \ldots, S_{1j}^{(1)} \ldots, S_{Kj}^{(1)} \ldots, S_{11}^{(2)} \ldots, S_{K1}^{(2)} \ldots, S_{1j}^{(2)} \ldots, S_{Kj}^{(2)})'
\]
is normally distributed with mean
\[
\begin{pmatrix}
\theta^{(1)} & \otimes & I^{(1)} \\
\theta^{(2)} & \otimes & I^{(2)}
\end{pmatrix}
\]
and variance
\[
\begin{pmatrix}
\Sigma^{(GS)}(I^{(1)}) & \rho \Sigma^{(GS)}(I^{(12)}) \\
\rho \Sigma^{(GS)}(I^{(12)}) & \Sigma^{(GS)}(I^{(2)})
\end{pmatrix}
\otimes
\Sigma^{(CS)}
\begin{pmatrix}
1 \\
1 + \lambda
\end{pmatrix}
\]
where is a \(J\)-dimensional vector with \(I_{12}^{(i)} = \sqrt{I_{1}^{(i)} I_{2}^{(i)}}, j = 1, \ldots, J\).

We consider next designs with subgroup selection. In analogy to the notation used above, now let \(\hat{\theta}^{(i)}_{kj}\) be the estimate of the treatment effect in subgroup \(k\), \(k = 1, \ldots, K\) at stage \(j\), \(j = 1, \ldots, J\), for endpoint \(i\), \(i = 1, 2\). We will typically consider the case of \(K = 2\), with \(k = 2\) corresponding to the entire population and \(k = 1\) corresponding to a subgroup comprising a proportion \(\tau\) of the entire population.

Let \((I^{(i)}_{kj})^{-1}\) denote the variance of \(\hat{\theta}^{(i)}_{kj}\) and let \(S^{(i)}_{kj} = \hat{\theta}^{(i)}_{kj} I^{(i)}_{kj}\). As above, we will assume that the \(\hat{\theta}^{(i)}_{kj}\) are normally distributed with mean \(\theta^{(i)}_k\) and variance \((I^{(i)}_{kj})^{-1}\) with
\[
\text{cov}(\hat{\theta}^{(i)}_{kj}, \hat{\theta}^{(i)}_{k'j'}) = (I^{(i)}_{\max\{k,k'\} \max\{j,j'\}})^{-1}
\]
and
\[
\text{cov}(\hat{\theta}^{(i)}_{kj}, \hat{\theta}^{(i')}_{k'j'}) = \rho (I^{(i)}_{\max\{k,k'\} \max\{j,j'\}})^{-1/2}, i \neq i'.
\]

Denote \(I^{(i)}_{j2}\) by \(I^{(i)}_j\) and assume further that \(I^{(i)}_{j1} = \tau I^{(i)}_j\) for \(i = 1, 2, j = 1, \ldots, J\) (see Spiessens and Debois), then
\[
(S^{(1)}_{11}, S^{(1)}_{21}, \ldots, S^{(1)}_{1j}, S^{(1)}_{2j}, S^{(2)}_{11}, S^{(2)}_{21}, \ldots, S^{(2)}_{1j}, S^{(2)}_{2j})'
\]
is normally distributed with mean
\[
\begin{pmatrix}
\theta^{(1)} & \otimes & I^{(1)} \\
\theta^{(2)} & \otimes & I^{(2)}
\end{pmatrix}
\]
and variance
\[
\begin{pmatrix}
\Sigma^{(GS)}(I^{(1)}) & \rho \Sigma^{(GS)}(I^{(12)}) \\
\rho \Sigma^{(GS)}(I^{(12)}) & \Sigma^{(GS)}(I^{(2)})
\end{pmatrix}
\otimes
\Sigma^{(GS)}
\begin{pmatrix}
\tau \\
1
\end{pmatrix}
\]
where \(I^{(i)} = (I^{(1)}, \ldots, I^{(i)}_j)', I^{(12)} = (I^{(12)}, \ldots, I^{(12)}_j)'\) with \(I^{(12)}_j = \sqrt{I^{(1)}_j I^{(2)}_j}\), \(\theta^{(i)} = (\theta^{(i)}_1, \theta^{(i)}_2)\), and \(\Sigma^{(GS)}(I)\) and \(\Sigma^{(CS)}(r)\) are as defined above.

If more than two subgroups are considered, or if these are nested in some other way than the second being a subset of the first, the last matrix can be modified accordingly to reflect this.