Adaptive treatment allocation and selection in multi-arm clinical trials: a Bayesian perspective

Elja Arjas (elja.arjas@helsinki.fi)
University of Helsinki
Dario Gasbarra
University of Helsinki

Research Article

Keywords: Phase II, Phase III, adaptive design, likelihood principle, posterior inference, decision rule, frequentist performance, binary data, time-to-event data, vaccine efficacy trial

Posted Date: October 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-943138/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on February 20th, 2022. See the published version at https://doi.org/10.1186/s12874-022-01526-8.

Adaptive treatment allocation and selection in multi-arm clinical trials: a Bayesian perspective

Elja Arjas*[1] and Dario Gasbarra[2]

[1] University of Helsinki and University of Oslo
[2] University of Helsinki

October 6, 2021

Abstract

Background: Adaptive designs offer added flexibility in the execution of clinical trials, including the possibilities of allocating more patients to the treatments that turned out more successful, and early stopping due to either declared success or futility. Commonly applied adaptive designs, such as group sequential methods, are based on the frequentist paradigm and on ideas from statistical significance testing. Interim checks during the trial will have the effect of inflating the Type 1 error rate, or, if this rate is controlled and kept fixed, lowering the power.

Results: The purpose of the paper is to demonstrate the usefulness of the Bayesian approach in the design and in the actual running of randomized clinical trials during Phase II and III. This approach is based on comparing the performance of the different treatment arm in terms of the respective joint posterior probabilities evaluated sequentially from the accruing outcome data, and then taking a control action if such posterior probabilities fall below a pre-specified critical threshold value. Two types of actions are considered: treatment allocation, putting on hold at least temporarily further accrual of patients to a treatment arm (Rule 1), and treatment selection, removing an arm from the trial permanently (Rule 2). The main development in the paper is in terms of binary outcomes, but extensions for handling time-to-event data, including data from vaccine trials, are also discussed. The performance of the proposed methodology is tested in extensive simulation experiments, with numerical results and graphical illustrations documented in a Supplement to the main text. As a companion to this paper, an implementation of the methods is provided in the form of a freely available R package.

Conclusion: The proposed methods for trial design provide an attractive alternative to their frequentist counterparts.

Keywords: Phase II, Phase III, adaptive design, likelihood principle, posterior inference, decision rule, frequentist performance, binary data, time-to-event data, vaccine efficacy trial.

*corresponding author, elja.arjas@helsinki.fi

1 Introduction

From the earliest contributions to the present day, the statistical methodology for designing and executing clinical trials has been dominated by frequentist ideas, most notably, on testing a precise hypothesis of "no effect difference" against an alternative, using a fixed sample size, and applying a pre-specified significance level to control for Type 1 error, as a means to guard against false positives in long term. An important drawback of this basic form of the standard methodology is that the design does not include the possibility of interim analyses during the trial. Particularly in exploratory studies during Phase II aimed at finding effective treatments from among a number of experimental candidates it is natural look for extended designs that allow the execution of the trial to be modified based on the results from interim analyses. For example, such results could provide reasons for terminating the accrual of additional patients to some treatments for lack of efficacy or, if the opposite is true, for allocating more patients to the treatments that turned out more successful. Allowing for earlier dissemination of such findings may then also benefit the patient population at large.

These motivations have led to the development of a whole spectrum of adaptive trial designs, and of corresponding methods for the statistical analysis of such data. An authoritative presentation of group sequential methods is provided in the monograph Jennison and Turnbull (1999). More general reviews of adaptive clinical trial designs, from the perspective of classical inference, can be found in, e.g., Chow and Chang (2008), Mahajan and Gupta (2010), Chow (2014), Chang and Balser (2016), Pallmann et al. (2018) and Atkinson and Biswas (2019). While such adaptive designs allow for greater flexibility in the running of actual trials, their assessment is usually based on selected frequentist performance measures. In the standard version, interim analyses are planned before the trial is started, and need then to be accounted for, due to the consequent multiple testing, in computing the probability of Type 1 error. Although such rigid form of planning can be relaxed when employing the so-called alpha spending functions (e.g., Pocock (1977), O'Brien and Fleming (1979), Demets and Lan (1994)), looking into the data before reaching the pre-planned end of the trial carries a cost either in terms of an inflated probability of Type 1 error or, if that is fixed, in a reduced power of the test to detect meaningful differences between the considered treatments.

These classical approaches in the design and execution of clinical trials have been challenged from both foundational and practical perspectives. Important early contributions include, e.g., Thompson (1933), Flühler et al. (1983), Berry (1985), Spiegelhalter et al. (1986), Berger and Berry (1988), Spiegelhalter et al. (1994) and Thall and Simon (1994); for a brief historical account and a large number of references, see Grieve (2016). Comprehensive expositions of the topic are provided in the monographs Spiegelhalter et al. (2004), Berry et al. (2011) and Yuan et al. (2017).

The key argument here is the change of focus: instead of guarding against false positives in a series of trials in long term, the main aim is to utilize the full information potential in the observed data from the ongoing trial itself. Then, looking into the data in interim analyses is not viewed as something incurring a cost, but rather, as providing an opportunity to act more wisely. The foundational arguments enabling this change are provided by the adoption of the likelihood principle, e.g., Berger and Wolpert (1984).

In practice, this also implies a change of the inferential paradigm, from frequentist into Bayesian. In Bayesian inference, the conditional (posterior) distribution for unknown model parameters is being updated based on the available data, via updates of the corresponding likelihood.

In a clinical trial, it is even possible to continuously monitor the outcome data as they are observed, and thereby utilize such data in a fully adaptive fashion during the execution of the trial. The advantages of this approach are summarized neatly in the short review paper Berry (2006), in Berry (2011), Lee and Chu (2012), and more recently, in Yin et al. (2017), Ruberg et al. (2019) and Giovagnoli (2021). The paper Villar et al. (2015) contains a useful review of the theoretical background, connecting the theory of the optimal design of clinical trials with that of *multi-armed bandit* problems. Unfortunately, general results on optimal strategies are largely lacking and their application in practice often infeasible because of computational complexity; however, see Press (2009). Recently, simulation based approximations have been used for applying Bayesian decision theory in the clinical trials context (e.g., Müller et al. (2017), Yuan et al. (2017), Alban et al. (2018)).

Importantly, the posterior probabilities provide intuitively meaningful and directly interpretable answers to questions concerning the mutual comparison of different treatments, given the available evidence, and do so without needing reference to concepts such as sampling distribution of a test statistic under given hypothetical circumstances.

Here we consider adaptive designs mainly from the perspective of multi-arm Phase II clinical trials, in which one or more experimental treatments are compared to a control. However, the same ideas can be applied, essentially without change, in confirmatory Phase III trials, where only a single experimental treatment is compared to a control, but the planned size of the trial is larger. In both situations, treatment allocation of individual trial participants is assumed to take place according to a fixed block randomization, albeit with an important twist: The performance of each treatment arm is assessed after every measured outcome in terms of the posterior distribution of a corresponding model parameter. Different treatments arms are then compared to each other according to pre-defined criteria. If a treatment arm is found to be inferior in such a comparison to the others, it can be closed off either temporarily or permanently from further accrual.

Of the recent clinical trials literature, the papers by Villar et al. (2015) and Jacob et al. (2016) seem most closely related to our approach, although in different ways. In the latter part of Villar et al. (2015), the authors discuss and compare several adaptive strategies according to which patients can be allocated to different treatments in a multi-arm trial. Although the paper uses Bayesian inferential methods in parameter estimation, the final comparison between alternative methods is based on frequentist ideas and measures: testing of hypotheses, using fixed sample size and given significance level. In contrast to this, Jacob et al. (2016) introduces three dynamic rules for dropping inferior treatment arms during the trial; these rules are closely similar to our Rules 1 and 2 below. On the other hand, and unlike Villar et al. (2015), Jacob et al. (2016) does not explicitly consider the possibility of adaptive treatment allocation.

We consider first, in Section 2, the simple situation in which the outcomes are binary, and they can be observed soon after the treatment has been delivered. In Section 3, the approach is extended to cover situations in which either binary outcomes are measured after a fixed time lag from the treatment, or the data consist of time-to-event measurements, with the possibility of right censoring. This section includes also some notes on vaccine efficacy trials. The paper concludes with a discussion in Section 4. A Supplement accompanied with the main text reports results from extensive simulation experiments, which follow closely the settings of two examples in Villar et al. (2015) but apply the adaptive methods introduced in Section 2. Its presentation is to a large extent comparative and expository. As a companion to this paper, we provide an implementation of the proposed method in the form of a freely available R package Marttila

et al. (2021) that facilitates the simulation of clinical trials with adaptive treatment allocation.

2 The case of Bernoulli outcomes

2.1 An adaptive method for treatment allocation: Rule 1

As in Villar et al. (2015) and Jacob et al. (2016) and numerous other papers, we consider first the 'prototype' example of a trial with binary outcomes and two types of treatments, one type representing a *control* or *reference* treatment indexed by 0, and K *experimental* treatments indexed by $k, 1 \leq k \leq K$. Motivated by a *conditional exchangeability* postulate between trial participants (with conditioning corresponding to their assignment to the different treatment arms), independent Bernoulli outcomes can in this case be assumed for all treatments, with respective response rates θ_0 and $\theta_1, \theta_2, \ldots, \theta_K$ considered as model parameters.

We index the participants in their order of recruitment to the trial by $i, 1 \leq i \leq N_{\max}$, where $N_{\max}$ is an assumed maximal size of the trial. If no such maximal size is specified, we choose $N_{\max}$ to be infinite. In this prototype version it is assumed that, for each i, the outcome Y_i from the treatment of patient i is observed soon after the treatment has been delivered. This assumption simplifies the consideration of adaptive designs, as the rule applied for deciding the treatment given to each participant can then directly account for information on such earlier outcomes. The meaning of 'soon' here should be understood in a relative sense to the accrual of participants to the trial. If the considered medical condition is rare in the background population, accrual will usually be slow with relatively long times between the arrivals. Then this requirement of outcome information being available when the next participant arrives may apply even if 'soon' is not literally true in chronological time. Extensions of this simple situation are considered in Section 3.

We assume that, before starting the trial, a sequential block randomization to the treatment arms $0, 1, ..., K$ has been performed. We index by $n \geq 1$ the positions on that list, calling n *list index*, and denote by $r(n)$ the corresponding treatment arm. Thus, we have a fixed sequence $((r(1), r(2), ...r(K+1)), (r(K+2), r(K+3), ...r(2(K+1)), ...)$ of randomized blocks of length $K+1$, where the blocks are independent random permutations of the treatment arm indexes $0, 1, ..., K$.

Assignment of the participants to the different treatment arms is now assumed to follow this list, but with the possibility of skipping a treatment arm in case it has been determined to be in the *dormant* state for the considered value of n. This leads to a balanced design in the sense that, as long as no treatment arms have been skipped by the time of considering list index n, the numbers of participants assigned to different treatments can differ from each other by at most 1, and they are equal when n is a multiple of $K+1$.

Denote by $I_{k,n}$ the binary indicator variable of arm k being in *active* state at list index value n, $n \geq 0$, $0 \leq k \leq K$, and let $I_n = (I_{0,n}, I_{1,n}, ..., I_{K,n})$ be the corresponding activity state vector. The values of these vectors are determined in an inductive manner to be specified later.

By inspection we find that, at the time a value $n \geq 1$ of the list index is considered, altogether

$$N(n) = \sum_{m=1}^{n} I_{r(m),m-1} \tag{2.1}$$

trial participants have so far arrived and been assigned to some treatment. Clearly $N(n) \leq n$. Let now the sequence $\{N^{-1}(i); i \geq 1\}$ be defined recursively by

$$N^{-1}(1) = 1;\ N^{-1}(i) = \inf\left\{n > N^{-1}(i-1) : I_{r(n),n-1} = 1\right\}, i > 1. \tag{2.2}$$

Then $N^{-1}(i)$ is the value of the list index n at which participant i is assigned to a treatment, while $A_i = r(N^{-1}(i))$ is the index of the corresponding treatment arm. Having postulated independent Bernoulli outcomes with treatment arm specific parameters $\theta_k, 0 \leq k \leq K$, we then get that Y_i is distributed according to $Bernoulli(\theta_{r(N^{-1}(i))})$.

The distinction between active and dormant states is that no trial participants are assigned, at a value n of the list index, to a treatment arm $r(n)$ if it is in the dormant state. Generally speaking, treatments whose performance in the trial has been poor, in a relative sense to the others, are more likely to be transferred into the dormant sate. However, with more data, there may later turn out to be sufficient evidence for such a trial arm to be returned back to the active state.

The data D_n that have accrued from the trial when it has proceeded up to list index value n consist of the values of the state indicators $I_{k,m-1},\ 0 \leq k \leq K, 1 \leq m \leq n$, and of treatments A_i and outcomes Y_i for $i \leq N(n)$.

Next, we outline the inductive rule by which the values of state vectors $I_n = (I_{0,n}, I_{1,n}, ..., I_{K,n})$ in a data sequence $\{D_n; n \geq 1\}$ are updated when the value of n is increased by 1. We write $\theta = (\theta_0, \theta_1, \ldots \theta_K)$ and use, for clarity, boldface notation $\boldsymbol{\theta}_k$ when the parameters are unknown and considered as random variables. Denote also $\boldsymbol{\theta}_\vee = \max\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

According to this rule, called Rule 1, for $n \geq 1$ and if $r(n) = k$ is an experimental treatment arm, we let $I_{k,n} = 0$ if $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \big| D_n) < \varepsilon$, and otherwise $I_{k,n} = 1$. Similarly, for the control arm $r(n) = 0$ we let $I_{0,n} = 0$ if $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee \big| D_n) < \varepsilon$, and otherwise $I_{0,n} = 1$. Here the threshold values $\varepsilon > 0$ and $\delta \geq 0$ are selected *operating characteristics* of the algorithm. A smaller value of ε reflects then a more conservative attitude towards moving a treatment into the dormant state. The value of δ can be viewed as specifying the *minimal important difference* (MID) or *minimal clinically important difference* (MCID) in the trial; if positive, it provides some extra protection to the control arm from being moved into the dormant state.

At the beginning, for $n = 0$, the coordinates of $I_0 = (I_{0,0}, I_{1,0}, ..., I_{K,0})$ are determined in a similar fashion directly from the prior. In practice, the prior is never so strong that we would not have $I_0 = (1, 1, ..., 1)$.

Rule 1 *Adaptive method for treatment allocation.*

---

if $\pi\left(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee\right) < \varepsilon$ **then**
  $I_{0,0} \leftarrow 0$;
else
  $I_{0,0} \leftarrow 1$;
end
for $k \leftarrow 1$ **to** K *(experimental treatment arms)* **do**
  if $\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee\right) < \varepsilon$ **then**
    $I_{k,0} \leftarrow 0$;
  else
    $I_{k,0} \leftarrow 1$;
  end
end
$I_0 \leftarrow (I_{0,0}, I_{1,0}, ..., I_{K,0})$;
$\text{Lik}_0(\theta) \leftarrow 1$;
$n \leftarrow 0$;
$N(0) \leftarrow 0$;
while $N(n) < N_{\max}$ **do**
  $n \leftarrow n + 1$;
  if $I_{r(n),n-1} = 0$ **then**
    $N(n) \leftarrow N(n-1)$;
    $I_n \leftarrow I_{n-1}$;
    $\text{Lik}_n(\theta) \leftarrow \text{Lik}_{n-1}(\theta)$;
  else
    (in this case $I_{r(n),n-1} = 1$*)*;
    $N(n) \leftarrow N(n-1) + 1$;
    $A_{N(n)} \leftarrow r(n)$;
    $Lik_n(\theta) \leftarrow Lik_{n-1}(\theta) \times \theta_{r(n)}^{Y_{N(n)}} \left(1 - \theta_{r(n)}\right)^{1-Y_{N(n)}}$;
    for $k \leftarrow 1$ **to** K *(experimental treatment arms)* **do**
      if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \middle| D_n\right) < \varepsilon$ **then**
        $I_{k,n} \leftarrow 0$;
      else
        $I_{k,n} \leftarrow 1$;
      end
    end
    if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee \middle| D_n\right) < \varepsilon$ **then**
      $I_{0,n} \leftarrow 0$;
    else
      $I_{0,n} \leftarrow 1$;
    end
  end
end

As a byproduct, successive applications of Rule 1 give us an explicit expression for the likelihood $L\left(\theta|D_n\right) = \text{Lik}_n, n \geq 1$, arising from observing data D_n as specified above. According to this rule, the likelihood expression $L\left(\theta|D_n\right)$ is updated only at values of n at which

$I_{r(n),n} = 1$, and then this is done by multiplying the previous value $L(\theta|D_{n-1})$ by the factor $\theta_{r(n)}^{Y_{N(n)}} \left(1-\theta_{r(n)}\right)^{1-Y_{N(n)}}$. By repeatedly applying the chain multiplication rule for conditional probabilities, we get that

$$L(\theta|D_n) = \prod_{m=1}^{n} \theta_{r(m)}^{I_{r(m),m}Y_{N(m)}} \left(1-\theta_{r(m)}\right)^{I_{r(m),m}(1-Y_{N(m)})} = \prod_{k=0}^{K} \theta_k^{N_{k,1}(n)} (1-\theta_k)^{N_{k,0}(n)}. \tag{2.3}$$

The right hand side expression is obtained by re-arranging the terms and denoting by

$$N_{k,1}(n) = \sum_{m=1}^{n} I_{k,m} 1_{\{Y_{N(m)}=1\}},\ N_{k0}(n) = \sum_{m=1}^{n} I_{k,m} 1_{\{Y_{N(m)}=0\}},\ 0 \le k \le K,\ n \ge 1, \tag{2.4}$$

respectively, the number of successful and failed outcomes from treatment k when considering list index values up to n. Of intrinsic importance in this derivation is that, when conditioning sequentially at n on the data D_n, the criteria according to which the values of the indicators $I_{k,n}$ are updated to $I_{k,n+1}$ do not depend on the parameter θ. As a consequence, these updates do not contribute to the likelihood terms that would depend on θ. Different formulations of this result can be found in many places, e.g., Villar et al. (2015).

As a consequence we can change the focus from the full data $\{D_n, n \ge 1\}$, indexed according to the original list indexes used for randomization, to "condensed" data $\{D_i^*, i \ge 1\}$ indexed according to the order in which the participants were treated. We denote by

$$S_k(i) = \sup\{N_{k,1}(n) : N(n) \le i\},\ F_k(i) = \sup\{N_{k,0}(n) : N(n) \le i\},\ 0 \le k \le K, \tag{2.5}$$

respectively, the number of successful and failed outcomes from treatment k when considering the first i participants. Let

$$S(i) = \sum_{k=0}^{K} S_k(i),\ F(i) = \sum_{k=0}^{K} F_k(i) \tag{2.6}$$

be the corresponding total number of successes and of failures, across all treatment arms.

Following the usual practice in similar contexts, we assume that the unknown parameter values $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ have been assigned independent *Beta*-priors, with $Beta(\theta_k|\alpha_k, \beta_k)$ for treatment arm k, where α_k and β_k are separately chosen hyperparameters. The choice of appropriate values of these hyperparameters (e.g., Thall and Simon (1994)) is always context specific, and is not discussed here further. Then, due to the well-known conjugacy property of the *Beta*-priors and the Bernoulli-type likelihood (2.3), the posterior $p\left(\theta_k|D_{k,i}^*\right)$ for $\boldsymbol{\theta}_k$, corresponding to data D_i^*, has the form of Beta-distribution with its parameters updated directly from the data:

$$p\left(\theta_k|D_{i,k}^*\right) = Beta\left(\theta_k|\alpha_k + S_k(i), \beta_k + F_k(i)\right),\ i \ge 1, k = 0, 1, \ldots, K. \tag{2.7}$$

This, together with the product form of the likelihood (2.3) and the assumed independence of the priors π, allows then for an easy computation of the joint posterior distribution for $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ for any i. The density $p_\pi\left(\theta_0, \theta_1, \ldots, \theta_K|D_{i,k}^*\right)$ becomes the product of $K+1$ *Beta*-densities. For example, posterior probabilities of the form $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \middle| D_n\right)$, or posterior distributions for pairwise differences of the type $\boldsymbol{\theta}_k - \boldsymbol{\theta}_0$ or $\boldsymbol{\theta}_k - \boldsymbol{\theta}_l$, can be computed numerically, in practice either by numerical integration as in Jacob et al. (2016), or by performing Monte

Carlo sampling from this distribution; see also Zaslavsky (2012). In our numerical examples in Section 3 we have applied this latter possibility.

While Rule 1 may at least temporarily inactivate some less successful treatment arms and thereby close them off from further accrual, this closure need not be final. As long as a treatment arm is in the dormant state, the posterior for the corresponding parameter $\boldsymbol{\theta}_k$ remains fixed. In contrast, with the accrual of participants to active treatment arms still continuing, the posteriors for their parameters can be expected to become less and less dispersed. As a consequence, returns from dormant to active state tend to become increasingly rare.

Thompson's rule. Rule 1 has much similarity with Thompson's rule (Thompson (1933), see also, e.g., Thall et al. (2015), Villar et al. (2015)), and both can be viewed as particular versions of *response-adaptive randomization* (RAR) designs (Chow and Chang (2008)). In its standard version, this Thompson's rule randomizes new patients to different treatment arms $k, 0 \leq k \leq K$, directly according to the posterior probabilities $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \middle| D_n\right)$, updating the values of these probabilities as described above. Fractional versions of Thompson's rule use probability weights for this purpose, based on powers $\left(\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \middle| D_n\right)\right)^\kappa$, with $0 \leq \kappa \leq 1$, normalized into probabilities by dividing such terms by their sum over different values of k. Thus, for $\kappa = 0$, the randomization is symmetric to all $K+1$ treatments, and its adaptive control mechanism becomes stronger with increasing κ. We consider Thompson's rule for comparative purposes in the context of the simulation experiments described in the Supplement.

2.2 An adaptive method for treatment selection: Rule 2

While an open end recipe such as Rule 1 or Thompson's algorithm may seem attractive, for example, from the perspective of drawing increasingly accurate inferences on the response parameters, practical considerations will often justify incorporation of rules for more definitive selection of some treatments and elimination of others. This is the case if the continued availability of more than one experimental treatment alternative at a later point in time is judged to be impracticable, as when entering the study into Phase III. Another reason is that incorporation of such decision rules enables us to make more direct comparisons to trial designs utilizing classical hypothesis testing ideas.

With this in mind, we complement Rule 1 with an optional possibility to conclusively terminate the accrual of additional participants to the less successful treatment arms. Rule 2 below is an adaptation and extension of the corresponding definitions in, e.g., Thall and Wathen (2007), Berry et al. (2011), Xie et al. (2012) and Jacob et al. (2016). In the commonly adopted terminology of adaptive designs, Rule 2 can be said to be a combination of versions of *response-adaptive randomization* (RAR) and *drop-the-losers* designs (Chow and Chang (2008)).

In the definition of the algorithm, the letter $\mathbb{T}$ is used as a generic notation for the set of treatment arms still left in the trial at the considered value of n. Each elimination of a treatment reduces its size by one. Rule 2 contains, as part, Rule 1 for moving treatments to the dormant state. It then involves, in addition to the operating characteristics ε and δ for Rule 1, three new parameters, viz. θ_{low}, ε_1 and ε_2. Specifying a value for θ_{low} means setting up a level of *minimum required treatment response rate* (MRT), e.g., Xie et al. (2012). A treatment $k \in \mathbb{T}$ is eliminated from the trial if the posterior probability for $\{\boldsymbol{\theta}_k > \theta_{low}\}$ falls below ε_1. The criteria for eliminating treatments are formally identical to those for moving them into the dormant state except that the bounds for the posterior probabilities then need to be tighter, $\varepsilon_1 \leq \varepsilon$.

Rule 2 *Adaptive rule for treatment allocation and selection.*

---

if $\pi\left(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee\right) < \varepsilon$ **then**
  $I_{0,0} \leftarrow 0$;
else
  $I_{0,0} \leftarrow 1$;
end
for $k \leftarrow 1$ **to** K *(experimental treatment arms)* **do**
  if $\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee\right) < \varepsilon$ **then**
    $I_{k,0} \leftarrow 0$;
  else
    $I_{k,0} \leftarrow 1$;
  end
end
$I_0 \leftarrow (I_{0,0}, I_{1,0}, ..., I_{K,0})$;
$\mathbb{T} \leftarrow \{0, 1, ..., K\}$;
$N(0) \leftarrow 0$;
$\text{Lik}_0(\theta) \leftarrow 1$;
$n \leftarrow 0$;
while $N(n) < N_{\max}$ **do**
  $n \leftarrow n + 1$;
  if $I_{r(n),n-1} = 0$ **then**
    $N(n) \leftarrow N(n-1)$;
    $I_n \leftarrow I_{n-1}$;
    $\text{Lik}_n(\theta) \leftarrow \text{Lik}_{n-1}(\theta)$;
  else
    in this case $(r(n) \in \mathbb{T})$ *and* $(I_{r(n),n-1} = 1)$;
    $N(n) \leftarrow N(n-1) + 1$;
    $A_{N(n)} \leftarrow r(n)$;
    $\text{Lik}_n(\theta) \leftarrow \text{Lik}_{n-1}(\theta) \times \theta_{r(n)}^{Y_{N(n)}} \left(1 - \theta_{r(n)}\right)^{1 - Y_{N(n)}}$;
    for $k \in \mathbb{T} \setminus \{0\}$ *(experimental treatment arms)* **do**
      if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \middle| D_n\right) < \varepsilon_1$ **or** $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon_2$ **then**
        $\mathbb{T} \leftarrow \mathbb{T} \setminus \{k\}$;
        $I_{k,n} \leftarrow 0$;
        $n_{k,last} \leftarrow n$;
      else if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon$ **then**
        $I_{k,n} \leftarrow 0$;
      end
      else
        $I_{k,n} \leftarrow 1$;
      end
    end
    if $0 \in \mathbb{T}$ **then**
      if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \theta_{low} \middle| D_n\right) < \varepsilon_1$ **or** $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon_2$ **then**
        $\mathbb{T} \leftarrow \mathbb{T} \setminus \{0\}$;
        $I_{0,n} \leftarrow 0$;
        $n_{0,last} \leftarrow n$;
      else if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell | D_n\right) < \varepsilon$ **then**
        $I_{0,n} \leftarrow 0$;
      end
      else
        $I_{0,n} \leftarrow 1$;
      end
    end
  end
end

Notes. The state indicator $I_{r(n),n}$ at list index value n depends on the recorded past trial history $\{D_m; 1 \leq m \leq n-1\}$. However, given this history, it is conditionally independent of the model parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_K)$. As was the case in Rule 1, for a given original block randomization, the likelihood expression arising from applying Rule 2 depends only on the outcome data.

This property is crucially important from the perspective of being able to draw correct statistical inferences from the trial. But it is also important from the perspective of practical implementation. Having assumed the initial randomization $\{r(n) : n \geq 1\}$ to be fixed, no further randomization is needed when the trial is run since, at any point in time, the next move to be made will be fully determined by the observed past data.

After every new observed outcome, the algorithm of Rule 2 determines the current state of each treatment arm, choosing between the three possible options: active, dormant, or dropped. All moves between these states are possible except that the dropped state is absorbing: once a treatment arm has been dropped, it will stay. If an arm is in dormant state, it is at least momentarily closed from further patient accrual.

Consider then the different actions based on Rule 2 in more detail. The posterior probabilities $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \middle| D_n\right)$ for the experimental arms, and $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \theta_{low} \middle| D_n\right)$ for the control arm, express how likely it is, given the currently available data, that their response rate exceeds the pre-specified MRT θ_{low}. The first criterion in Rule 2 then says that if this probability is below a selected threshold value ε_1, the treatment arm is dropped from the trial. The value of ε_1 can then be said to represent an acceptable risk level of error when concluding that $\{\boldsymbol{\theta}_k \geq \theta_{low}\}$, or $\{\boldsymbol{\theta}_0 + \delta \geq \theta_{low}\}$, would not be true. This part of Rule 2 will obviously not be active if either $\theta_{low} = 0$ or $\varepsilon_1 = 0$.

The second criterion in Rule 2 makes a comparison of the response rate of a treatment and that of the best treatment in the trial. Both values are unknown, and the comparison is made in terms of the posterior probabilities $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \big| D_n)$ for the experimental arms and $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \big| D_n)$ for the control. Here $\mathbb{T} \subset \{0, 1, ..., K\}$ is the set of treatment arms left in the trial at time n. The composition of $\mathbb{T}$ is determined in an inductive manner, starting from $\mathbb{T} = \{0, 1, ..., K\}$ at $n = 1$. A treatment is dropped from the trial if the corresponding posterior probability falls below the selected threshold level ε_2. Thus, for small ε_2, the decision to drop an experimental treatment k is made if, in view of the currently available data D_n, the event $\{\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell\}$ is true only with probability close to 0, with ε_2 representing the selected risk level. The control arm is protected even more strongly from inadvertent removal from the trial if a positive safety margin δ is employed; the comparison to experimental arms becomes symmetric if $\delta = 0$. This entire mechanism of eliminating treatments based on mutual comparisons is inactivated by letting $\varepsilon_2 = 0$.

One should note that, while Rule 1 is compatible with the likelihood principle, Rule 2 has an element which violates it. This is because, in multi-arm trials with $K > 1$, when considered at times n at which some treatment arms have already been dropped, the definition of the maximal response parameter value $\theta_V = \max_{\ell \in \mathbb{T}} \theta_\ell$ ignores those indexed in $\{0, 1, ..., K\} \setminus \mathbb{T}$. Sequential elimination of treatments, as embodied in Rule 2, while it has an obvious practical appeal in running a clinical trial, also renders properties such as standard Bayesian consistency inapplicable.

In the third criterion of Rule 2 copies Rule 1: A n experimental treatment arm $k \in \mathbb{T}$ is made

dormant if $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \max_{\ell\in\mathbb{T}} \boldsymbol{\theta}_\ell \big| D_n) < \varepsilon$, and the control arm if $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell\in\mathbb{T}} \boldsymbol{\theta}_\ell \big| D_n) < \varepsilon$, where ε is a selected threshold. For this part of Rule 2 to function in a nontrivial way, we need to choose $\varepsilon > \varepsilon_1$ and $\varepsilon > \varepsilon_2$. If either $\varepsilon = \varepsilon_1$ or $\varepsilon = \varepsilon_2$, then the possibility of a treatment arm being moved into the dormant state is ruled out, and if $\varepsilon_1 = \varepsilon_2 = 0$, then Rule 2 is easily seen to collapse into the simpler Rule 1. Finally, if also $\varepsilon = 0$, then treatment allocation will follow directly the original block randomization, which was assumed to be symmetric between all treatment arms, and no treatments are dropped before reaching N_{max}.

The selection of appropriate threshold values δ and θ_{low} in Rule 1 and Rule 2 should be based on substantive contextual arguments in the trial. If a positive value for δ is specified, then, as already mentioned in the context of Rule 1, this is commonly viewed as the *minimal clinically important difference* (MCID) in the trial. Employing such a positive threshold value when comparing the response rate of the control arm to that of an experimental arm, and not doing so when comparing two experimental arms to each other, reflects the idea that the design should be more conservative towards moving the control arm to the dormant state, let alone dropping it conclusively from the trial, than when contemplating about a similar move for an experimental treatment.

Once selected, the design parameters $\varepsilon, \varepsilon_1$ and ε_2 in applying Rule 2, and then deciding to either drop the treatment or putting it into the dormant state, can be interpreted directly as upper bounds for the risk that this decision was in fact unwarranted. By *risk* is here meant the posterior probability of error, each time conditioned on the current data actually observed. Suppose, for example, that a finite value for $n_{k,last}$ has been established due to $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \big| D_{n_{k,last}}\right) < \varepsilon_1$. Further accrual of trial participants to treatment arm k is then stopped after the patient indexed by $N_{k,last}$ because the response rate θ_k from that arm is judged, with only a small probability $\leq \varepsilon_1$, given the data, to be above the MRT level θ_{low}.

If all experimental treatments have been dropped as a result of applying Rule 2, the trial ends with a negative result, *futility*, e.g. Thall and Wathen (2007). On the other hand, if the control arm has been dropped, at least one of the experimental arms was deemed better than the control, which is a positive finding. In case more than two experimental arms were left at that time, the trial design may allow for a continued application of Rule 2, with the goal of ultimately identifying the one with the highest response rate.

As remarked earlier, the application of Rule 2 is optional. If it is not enforced, Rule 1 is open ended and will only control the assignment of new participants to the different treatments. Then, if the trial size $N_{\max}$ has been specified and fixed in advance, and regardless of whether Rule 1 was previously employed or not, the posterior probabilities $\mathbb{P}_\pi(\boldsymbol{\theta}_k \geq \theta_{low} | D^*_{N_{\max}})$, $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D^*_{N_{\max}})$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee | D^*_{N_{\max}})$ can be computed routinely after all outcome data $D^*_{N_{\max}}$ have been observed, to provide the final assessment of the results from the trial.

A frequentist perspective. A different perspective to the application of Rule 2 is offered by the classical frequentist theory of statistical hypothesis testing. While the main point of this paper is to argue in favor of reasoning directly based on posterior inferences, this may not be sufficient to satisfy stake holders external to the study itself, including the relevant regulatory authorities in question, which may be concerned about frequentist measures such as the overall Type 1 error rate at a pre-specified significance level (Chow and Chang (2008)).

From a frequentist point of view, the posterior probabilities $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_n)$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee | D_n)$, via their dependence on the data D_n, can be viewed as test statistics in respective sequential testing problems, with Rule 2 defining the stopping boundaries. In the case $K = 1$,

they correspond to considering two overlapping hypotheses (e.g., Lewis and Berry (1994)), null hypothesis $H_0 : \theta_1 \leq \theta_0 + \delta$ and its alternative $H_1 : \theta_1 \geq \theta_0$. For $K \geq 1$, the null hypothesis becomes $H_0 : \theta_\vee \leq \theta_0 + \delta$, and the alternative $H_1 : \theta_\vee \geq \theta_0$. The posterior probabilities $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_n)$ can then be used as test statistics in testing H_0, and $\mathbb{P}_\pi(\boldsymbol{\theta}_\vee \geq \boldsymbol{\theta}_0 | D_n)$ for testing H_1.

The size of the test depends on the hypothesized "true" values of the response parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_K)$, on the selected threshold values $\delta, \theta_{low}, \varepsilon, \varepsilon_1, \varepsilon_2$ and, if specified in advance, on the maximal size $N_{\max}$ of the trial. For clarity, we denote such a hypothesized distribution generating the data by $\mathbb{Q}$, distinct from the mixture distribution $\mathbb{P}_\pi$ used, after being conditioned on current data, in applying Rule 1 and Rule 2.

Frequentist measures such as true and false positive and negative rates, characterizing the performance of a test, can be computed numerically to a good approximation by performing a sufficiently large number of forward simulations from the selected $\mathbb{Q}$ and then averaging the sampled values. Explicit consideration of such frequentist measures is here deferred to a Supplement, which contains a large number of figures and tables from simulations run under different parameter settings. Two types of experiments are considered, one concerned with a 2-arm and the other a 4-arm trial. One may note that such frequentist considerations are of interest essentially only at the design stage when no outcome data are yet available and a trial design needs to be selected and approved.

When the trial is then run, it is natural to utilize, at each time n, the currently available data D_n and the consequent posterior probabilities such as $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_n)$, $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee | D_n)$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_k \geq \theta_{low} | D_n)$. Illustrations of this can be found in Figures S1 and S7 in the Supplement. In this context it may be useful to recall the well known result from general decision theory: For any prior, the smallest Bayes risk is achieved by minimizing "pointwise" the expected loss with respect to the posterior. In other words, a decision rule which is optimal locally, for each observed sample path, will be optimal also globally, on average.

3 Extensions for handling delayed outcome data

Data of the kind considered in Section 2, where binary outcomes are determined and observed soon after the treatment is delivered, may be rare in practical applications such as drug development. More likely, it takes some time until a response to a treatment can measured in a useful manner. For example, the status of a cancer patient could be determined one month after the treatment was given. Incorporation of such a delay into the model is not technically very difficult, but it necessitates explicit introduction of the recruitment or arrival process, in continuous time, of the patients to the trial. A somewhat different problem arises if the outcome itself is a measurement of time, such as time from treatment to relapse or to death in a cancer trial, or to infection in vaccine development. When such information would be needed for adaptive treatment allocation, part of the data are typically right censored. Both types of extensions of the basic Bernoulli model in Section 2 are considered briefly below.

3.1 Fixed delay from treatment to binary outcome

We now consider a model, where a binary outcome is systematically measured after a fixed time period has elapsed from the time at which the patient in question received the treatment. Modelling such a situation, rather obviously, requires that the model is based on a continuous time parameter.

Let, therefore, $t > 0$ be a continuous time parameter, and denote by $U_1 < U_2 < \ldots < U_i < \ldots$ the arrival times of the patients to the trial, again using $i = 1, 2, \ldots$ to index the participants. We then assume that the treatment is always given immediately upon arrival, and that the outcome Y_i is measured at time $V_i = U_i + d$, where $d > 0$ is fixed as part of the design. Let $N(t) = \sum_{i \geq 1} 1_{\{U_i \leq t\}}, t > 0$, be the counting process of arrivals. At time t, outcome measurements are available from only those patients who arrived and were treated before time $t - d$. Therefore, the adaptive rule for assigning a treatment to a participant arriving at time t can utilize only the data

$$D_t = \{U_i, A_i, C_i(t), C_i(t) Y_i : i \leq N(t)\},$$

where the indicator $C_i(t) = 1_{\{U_i < t-d\}}$ signals that Y_i has been measured by time t.

With a minor change from (2.4), let

$$N_{k,1}(t) = \sum_{i=1}^{N(t)} C_i(t) 1_{\{A_i=k, Y_i=1\}}, \; N_{k0}(t) = \sum_{i=1}^{N(t)} C_i(t) 1_{\{A_i=k, Y_i=0\}}, \; 0 \leq k \leq K, 0 < t \leq T_{\max}. \tag{3.1}$$

As before, we assume that the arrival process is not informative about the model parameters, that the participants are conditionally exchangeable given their respective treatment assignments, and that the assignment rule is the same as in Section 2. The main distinction between the model with instantaneous response times and the present one with delayed measured outcomes is that, in the former case, once the outcome on an arriving patient becomes known, there is no additional information in the data until the next patient arrives and is treated. In the present situation, however, during such a time period some other patients, who had arrived earlier, may complete the required duration d from treatment to measured outcome and thereby provide new information to the data that are available. That information can then be utilized when deciding on the treatment for the next arriving patient.

By inspection we find that the basic product form of the likelihood expression (2.3) can be retained in this case. More concretely, the only change needed in the algorithms of Rule 1 and Rule 2 is that, instead of $\text{Lik}_n(\theta) \leftarrow \text{Lik}_{n-1}(\theta) \times \theta_{r(n)}^{Y_{N(n)}} \left(1 - \theta_{r(n)}\right)^{1-Y_{N(n)}}$, the inductive step for updating the likelihood becomes

$$\text{Lik}_n(\theta) \leftarrow \text{Lik}_{n-1}(\theta) \prod_{k=0}^{K} \theta_k^{N_{k,1}\left(U_{N(n)}\right) - N_{k,1}\left(U_{N(n)-1}\right)} (1 - \theta_k)^{N_{k,0}\left(U_{N(n)}\right) - N_{k,0}\left(U_{N(n)-1}\right)}. \tag{3.2}$$

3.2 The case of time-to-event data

Time-to-event data can arise in several different ways. For example, the times from treatment to relapse or death are often used as primary endpoints in cancer trials. Below we show how Rule 1 and Rule 2 need to be modified to apply for such data.

Let U_i be the time of treatment and V_i the time of response for patient i, and let $X_i = V_i - U_i$. Changing the notation slightly, we now denote by $N(t) = \sum_{i\geq 1} 1_{\{U_i \leq t\}},\ t > 0$, the process counting the arrivals to the trial. If the data are collected at time t, and $U_i \leq t$ and $V_i > t$ hold for patient i, the response time X_i will be right censored. Observed in the data are then the times $Y_i(t) = [(V_i \wedge t) - U_i]^+$ and the indicators $C_i(t) = 1_{\{V_i \leq t\}} = 1_{\{X_i = Y_i(t)\}}$.

Suppose now that the original response times X_i arising from treatment k, i.e., those for which $A_i = k$, are independent and distributed according to some distribution $F(x|\theta_k)$ with respective parameter value $\theta_k > 0,\ k = 0, 1, \ldots, K$. Denote the corresponding densities by $f(x|\theta_k)$. As above, we assume that the arrival process is not informative about the model parameters, and that the participants are conditionally exchangeable given their respective treatment assignments. Then the likelihood expression corresponding to data

$$D_{k,t} = \{U_i, A_i, Y_i(t), C_i(t) : i \leq N(t), A_i = k\},$$

collected from treatment arm k up to time t, has the familiar form

$$L(\theta_k|D_{k,t}) = \prod_{i=1}^{N(t)} f(X_i|\theta_k)^{C_i(t)1_{\{A_i=k\}}} (1 - F(Y_i(t)|\theta_k))^{(1-C_i(t))1_{\{A_i=k\}}}. \tag{3.3}$$

Such data are in the survival analysis literature commonly referred to as data with *staggered entry*. Due to the assumed conditional independence of the response times across the different treatment arms, given the respective parameters θ_k, the combined data

$$D_t = \bigcup_{k=0}^{K} D_{k,t} = \{U_i, A_i, Y_i(t), C_i(t) : i \leq N(t)\}$$

give rise to the product form likelihood

$$L(\theta|D_t) = \prod_{k=0}^{K} L(\theta_k|D_{k,t}), \tag{3.4}$$

where $\theta = (\theta_0, \theta_1, \ldots, \theta_K)$. Upon specifying a prior for θ, the posterior probabilities corresponding to the data D_t can then be computed and utilized in Rule 1 or Rule 2.

Remarks. It is well known that, in Bayesian inference, *Gamma*-distributions are conjugate priors to the likelihood arising from exponentially distributed survival or duration data, with θ_k representing the corresponding intensity parameters. This holds also when such data are right censored, in which case the likelihood (3.3) corresponding to $D_{k,t}$ has the Poisson form, with $\sum_{i=1}^{N(t)} C_i(t) 1_{\{A_i=k\}}$ being the number of measured positive outcomes and $\sum_{i=1}^{N(t)} Y_i(t) 1_{\{A_i=k\}}$ the corresponding *Total Time on Test* (TTT) statistic. Assuming independent $Gamma(\theta_k \mid \alpha_k, \beta_k)$-priors for the respective treatment arms $k = 0, 1, \ldots, K$, the posterior for θ_k corresponding to data $D_{k,t}$ becomes

$$p(\theta_k|D_{k,t}) = Gamma\left(\theta_k \mid \alpha_k + \sum_{i=1}^{N(t)} C_i(t) 1_{\{A_i=k\}}, \beta_k + \sum_{i=1}^{N(t)} Y_i(t) 1_{\{A_i=k\}}\right), \tag{3.5}$$

and the joint posterior $p(\theta \mid D_t)$ is the product distribution of these independent marginals.

When considering the application of Rule 1 or Rule 2 in this exponential response time model, the natural target would often be to decrease, rather than increase, the value of the intensity

parameter corresponding to an experimental treatment in the trial. Moreover, for measuring the degree of such potential improvements, use of hazard ratios, or relative risks, seems often more appropriate than of absolute differences. Criteria such as $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \middle| D_n\right) < \varepsilon_1$ and $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell\in\mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon_2$ applied previously in Rule 2 should then be replaced by corresponding requirements of the form $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \leq \theta_{high} \middle| D_t\right) < \varepsilon_1$ and $\mathbb{P}_\pi\left(\rho\boldsymbol{\theta}_0 \leq \min_{\ell\in\mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_t\right) < \varepsilon_2$, where $\rho < 1$ is a given safety margin protecting the control arm from inadvertent dropping. Writing $\rho = \exp\{-\delta\}$ and using $\eta_k = -\log\theta_k$ as model parameters brings us back to the absolute scale, with the last inequality becoming the requirement $\boldsymbol{\eta}_0 + \delta \geq \max_{\ell\in\mathbb{T}} \boldsymbol{\eta}_\ell$.

3.3 Notes on application to vaccine trials

An important and timely special case of time-to-event data are data coming from large scale Phase III vaccine trials. When a newly developed vaccine candidate has reached the stage when it is tested in humans for efficacy, the trial participants are usually healthy individuals and the control treatment is either placebo or some existing vaccine that has been already approved for wider use. In such trials adaptive treatment allocation is less likely to be an issue, whereas it would be important to arrive at some reasonably definitive conclusion about efficacy already before reaching the planned study endpoint N_{max}. For this reason, in the recent trials for testing COVID-19 candidate vaccines in humans, the design has allowed for from two to five 'looks' into the data before trial completion, usually defined as times at which some pre-specified number of infections have been observed. To our knowledge, most of these trials have applied frequentist group sequential methods for testing, adjusting the targeted significance level by suitably defined spending functions. This standard practice is followed in spite of that, arguably, in trials for experimental vaccines such as the COVID-19 candidates, for which Phase II has been already successfully completed, Type 1 errors could be considered less worrisome than Type 2 errors.

Entertaining the idea that such vaccine trials had been designed by using the Bayesian framework as presented above in 3.2, this task could have been accomplished by applying Rule 2 and thereby selecting suitable values for its design parameters $\rho, \theta_{high}, \varepsilon_1, \varepsilon_2$ and N_{max}, letting finally $\varepsilon = \varepsilon_2$ to inactivate the separately defined adaptive mechanism for treatment allocation. For example, considering the case of a single experimental vaccine, the value $\rho = 0.4$ would signify the target of sixty percent decrease in the value of the intensity parameter θ_1 compared to the placebo control θ_0, and thereby a corresponding reduction in the expected number of infected individuals among those vaccinated.

The trial could then be run, and it would stop with declared *success* if a posterior probability $\mathbb{P}_\pi\big(\rho\boldsymbol{\theta}_0 < \boldsymbol{\theta}_1 \big| D_i^*\big) < \varepsilon_2$ were obtained for some $i \leq N_{max}$. On the other hand, *futility* would be declared if either $\mathbb{P}_\pi\big(\boldsymbol{\theta}_1 \leq \theta_{high} \big| D_i^*\big) < \varepsilon_1$ or $\mathbb{P}_\pi\big(\rho\boldsymbol{\theta}_0 \geq \boldsymbol{\theta}_1 \big| D_i^*\big) < \varepsilon_2$ were established for such i. In either case, the monitoring of these probabilities could in principle be done in an open book form, and not just in a few 'looks' made at pre-planned check points.

A somewhat different approach to modeling and analyzing vaccine trial data can be outlined as follows. Suppose that the design is fixed by allocating, at time $t = 0$, n_1 individuals to the vaccination group and n_0 individuals to the placebo group. Denote by $0 < T_{1,1} < T_{1,2} < ...$ the times at which the individuals in the former group become infected and by $0 < T_{0,1} < T_{0,2} < ...$

the corresponding times in the latter group. Expressed in terms of counting processes, $N_1(t) = \sum_{m\geq 1} 1_{\{T_{1,m}\leq t\}}$ and $N_1(t) = \sum_{m\geq 1} 1_{\{T_{1,m}\leq t\}}$ count the number of infections up to time t in these two groups. We then assume that infections occur at respective rates $(n_1 - N_1(t-))\lambda_1(t)$ and $(n_0 - N_0(t-))\lambda_0(t)$, where $\lambda_1(t)$ and $\lambda_0(t)$ are unknown functions of the follow-up time t. In practice, n_1 and n_0 are large, of the order 10.000 or more, while $N_1(t)$ and $N_0(t)$ can during the observation interval be at most a few hundred. Therefore, $\{N_1(t); t \geq 0\}$ and $\{N_0(t); t \geq 0\}$ can be approximated quite well by Poisson processes with respective intensities $n_1\lambda_1(t)$ and $n_0\lambda_0(t)$.

Suppose that these processes are (conditionally) independent given their intensities. Then the likelihood corresponding to the data $D_t = \{N_0(s), N_1(s); s \leq t\}$, combined from both groups and up to time t, gets the familiar Poisson-form expression

$$L(\lambda_0, \lambda_1|D_t) = \prod_{k=0}^{1} \exp\Big\{-\int_0^t n_k\lambda_k(s)ds\Big\} \prod_{m\leq N_k(t)} n_k\lambda_k(T_{k,m}). \tag{3.6}$$

Assuming that the processes $\{T_{0,m}; t \geq 1\}$ and $\{T_{1,m}; t \geq 1\}$ do not have exact ties, we now consider their superposition $\{0 < T_1 < T_2 < ...\}$ and the corresponding counting process $N(t) = N_0(t) + N_1(t) = \sum_{m\geq 1} 1_{\{T_m\leq t\}}$, which then has intensity $n_0\lambda_0(t) + n_1\lambda_1(t)$. In what follows, for the purposes of statistical inference, this superposition is decomposed back into its components. For this, we define a sequence $\{\delta(T_m); m \geq 1\}$ of indicators, letting $\{\delta(T_m) = 1\}$ if $\{N_0(T_m) - N_0(T_m-) = 1\}$. Expressed in concrete terms, the event $\{\delta(T_m) = 1\}$ occurs if the m^{th} individual in the trial who was recorded as being infected happens to belong to the placebo group, and $\{\delta(T_m) = 0\}$ if to the vaccination group. It is well known that the conditional probability of these events, given $\lambda_0(.)$, $\lambda_1(.)$ and $\{N(T_m)-N(T_m-) = 1\}$, are equal respectively to $n_0\lambda_0(T_m)(n_0\lambda_0(T_m) + n_1\lambda_1(T_m))^{-1}$ and $n_1\lambda_1(T_m)(n_0\lambda_0(T_m) + n_1\lambda_1(T_m))^{-1}$.

Estimation of the function $\lambda_0(.)$, describing the infection pressure in the non-vaccinated population, may be possible by utilizing data sources that are external to the trial, but estimation of $\lambda_1(.)$ would be hard. This problem can be circumvented if we are ready to impose a proportionality assumption, according to which, although the rates at which infections occur in the vaccination and placebo groups generally vary in time, their ratio is a constant $\rho > 0$. Expressed in symbols, we assume then that $\lambda_1(t) = \rho\lambda_0(t), t \geq 0$. The smaller the value of ρ, the better protected, according to this model, the vaccinated individuals are. The value $1 - \rho$ is what is commonly called *vaccine efficacy at reducing infection susceptibility*, abbreviated as VE_S (e.g., Halloran et al. (2010)).

The postulated proportionality property appears to be reasonable if all trial participants are vaccinated approximately at the same time, in which case t refers to time from vaccination, and if both groups, due to randomization, can be assumed to be exposed to approximately the same infection pressure. If the trial participants have been recruited from different geographical regions with highly varying levels of infection pressure, a stratified analysis based on a common vaccine efficacy value might still be possible. However, if vaccination takes place over a longer time period, it becomes difficult to differentiate from each other the effects of infection pressure, varying in the population with calendar time, and that of individual level susceptiblity, which is likely to depend on the build-up of the immune response and thereby on the time from vaccination.

A different matter, which has received much attention recently in connection of COVID-19 vaccine trials, is the dependence of ρ on age, due to the immune response in the older age

groups generally developing more slowly. Stratification of the analyses by using some age threshold has been applied, but the selected thresholds have varied. This is a problem for statistical analysis as long as the numbers of infected individuals in some age groups remain low.

Supposing now a common value for ρ, there are two alternative approaches to be selected from: Either (i) considering joint inferences on the pair $(\lambda_0(.), \rho)$, using the "full" likelihood (3.6) for this purpose and introducing a separate model for a description of $\lambda_0(.)$, or (ii) following the path well known from the context of the Cox proportional hazards model and employing a corresponding *partial likelihood* expression (e.g., Yip and Chen (2000)). In a stratified analysis, the (partial) likelihood expressions would become products across the considered strata. Here we consider briefly the approach based on partial likelihood. A comparative assessment of these approaches is beyond the scope of this presentation.

By inserting the assumed form $\lambda_1(.) = \rho\lambda_0(.)$ of the intensity $\lambda_1(.)$ into (3.6), it can be written, after some re-arrangement and cancellation of terms, in the form

$$L(\lambda_0, \rho|D_t) = \exp\left\{-(n_0 + n_1\rho)\int_0^t \lambda_0(s)ds\right\}(n_0 + n_1\rho)^{N(t)} \prod_{m\leq N(t)} \lambda_0(T_m)$$
$$\times \prod_{m\leq N(t)} \left(\frac{n_0}{n_0 + n_1\rho}\right)^{\delta(T_m)} \left(\frac{n_1\rho}{n_0 + n_1\rho}\right)^{1-\delta(T_m)}.$$

The latter product in this expression simplifies further into

$$L_{part}(\rho|D_t) = \left(\frac{n_0}{n_0 + n_1\rho}\right)^{\sum_{m\leq N(t)} \delta(T_m)} \left(\frac{n_1\rho}{n_0 + n_1\rho}\right)^{\sum_{m\leq N(t)} (1-\delta(T_m))} = \theta^{N_0(t)}(1-\theta)^{N(t)-N_0(t)}, \quad (3.7)$$

where we have denoted $\theta = n_0(n_0 + n_1\rho)^{-1}$. This is the sought-after partial likelihood and, parameterized in this way, it has the familiar Binomial form. The word *partial* signifies the fact that the parts in the "full" likelihood that were omitted in the derivation of (3.7) also contain the unknown model parameter ρ. We now proceed by employing the approximation where the partial likelihood is treated as if it were the "full". On specifying a $Beta(\ .\ |\ \alpha, \beta)$-prior for θ, and using the conjugacy property of the *Beta-Binomial* distribution family, we would get the posterior $p(\theta\ |\ D_t) = Beta(\theta\ |\ \alpha + N_0(t), \beta + N(t) - N_0(t))$, and further the posterior for $\boldsymbol{\rho}$ by noting that $\rho = n_0(1-\theta)/n_1\theta$.

However, a *Beta*-prior may not be fully appropriate for this particular application. More naturally we could postulate, for example, the $Uniform(0, 1)$ prior for ρ. It would correspond to the assumption that infectivity in the vaccine group cannot be larger than in the placebo group, but all values of vaccine efficacy between 0 and 100 percent are a priori equally likely. This would entail for θ a prior density, which is no longer of *Beta*-form. With the conjugacy property lacking in this case, the posterior can nevertheless be computed easily by applying Markov Chain Monte Carlo sampling.

While adaptive treatment allocation appears to be less of an issue in vaccine trials, there will be more interest in how, and when, results from such trials could be appropriately reported. At times such as the current SARS-CoV-2 pandemic, there is much pressure towards making the results from vaccine trials available as soon as a pre-specified level of certainty can be assured. Again, consistent with the likelihood principle, all monitoring of posterior probabilities could be done in an open book form, and not just in a few 'looks' at pre-planned check points. For

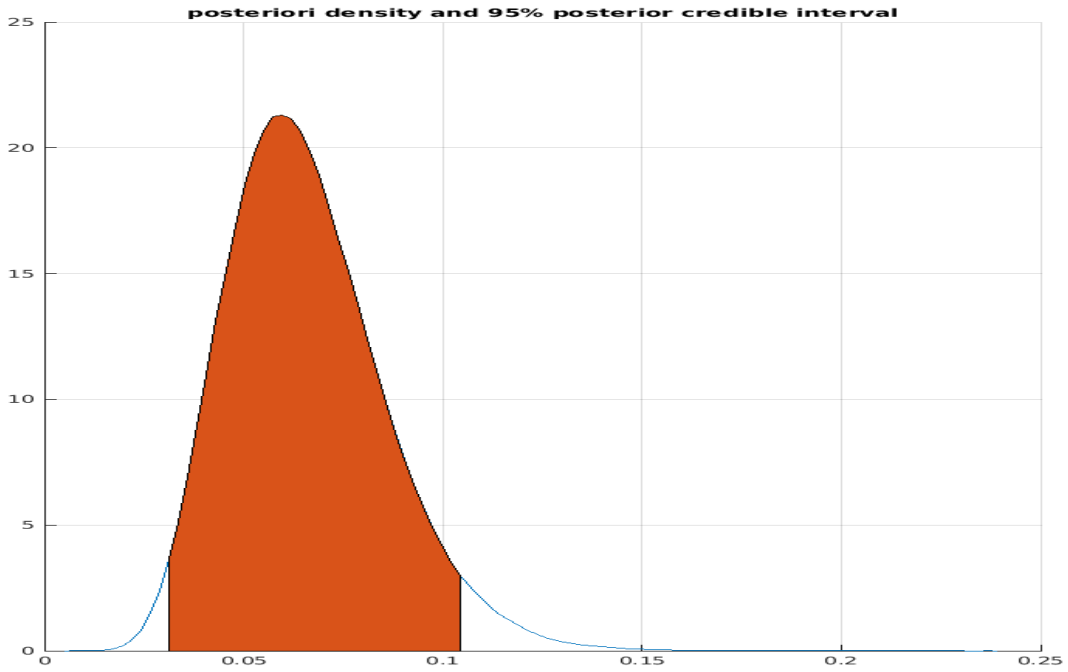

Figure 1: Posterior density of $\boldsymbol{\rho}$ based on Moderna, Inc. COVID-19 primary efficacy data, with posterior mode at 0.0595 and 95% HPD interval $(0.030, 0.105)$.

example, the trial could be run, and it could stop with declared success at time t if the posterior probability $\mathbb{P}_\pi(VE_S \geq ve^* | D_t) > 1 - \varepsilon_1$ were obtained, with ve^* a pre-specified minimal target value and ε_1 having a small value such as 0.05 or 0.01. (To compare, according to the WHO guidelines for evaluation of COVID-19 vaccines (World Health Organization (2020)), for a candidate vaccine the primary efficacy endpoint point estimate in a placebo-controlled efficacy trial should be at least 50 percent, and the lower bound of the appropriately alpha-adjusted confidence interval around the primary efficacy endpoint point estimate should be larger than 30 percent. Note that, while such a criterion defines a stopping time with respect to the internal history of the trial, it violates the likelihood principle.) A similar criterion could be set up for declaring futility.

To give an example from a recent real study, Moderna, Inc. announced on November 30, 2020 (Moderna Inc. (2020)) a primary efficacy analysis of their Phase III COVID-19 Vaccine Candidate. The announcement, based on a randomized, 1:1 placebo-controlled study of 30.000 participants, reported 185 infections in the placebo group and 11 in the vaccine group, leading to the point estimate $11/185 = 0.059$ of ρ and thereby efficacy estimate 0.941. We computed the posterior density $p(\rho \mid D_t)$ of $\boldsymbol{\rho}$, using these data $N_0(t) = 185$ and $N_1(t) = 11$ and assuming the uniform prior for $\boldsymbol{\rho}$ as described above. The result, together with the 95 percent HPDI $(0.030, 0.105)$, is shown in Figure 1. The corresponding HPDI for $VE_S = 1 - \rho$ is then $(0.895, 0.970)$.

Remarks. A practical advantage of the Poisson process approximation entertained above is that only the numbers $N_0(t)$ and $N_1(t)$ are needed for computing the posterior of $\boldsymbol{\rho}$ at time t. If n_0 and n_1 are not large enough to justify such an approximation, statistical inference based on partial likelihood is still possible, but it then necessitates monitoring of the sizes of the two

risk sets. The exact times of infection are not required, but the ordering in which members of either the placebo or of the vaccine groups become infected needs to be known. As in the case of the Cox proportional hazards model, the partial likelihood expression is then somewhat more involved and the computations more slow.

In the above approach and analysis we have assumed that the risk set sizes are reduced only due to the trial participants becoming infected. This may not be so, as there may be various other reasons why they may be lost from follow-up. If the resulting right censoring concerns a large proportion of the participants, this has to be accounted for in the analysis. It does not create a conceptually difficult problem, but it requires that the sizes of the risk sets, both in the vaccine and the placebo groups, are known at the times at which new infections are registered. The simple power form expression (3.7) for partial likelihood is then not valid any more, and needs to be replaced by the product

$$L_{part}(\rho|D_t) = \prod_{T_m \le t} \left(\frac{R_{0,T_m}}{R_{0,T_m} + R_{1,T_m}\rho} \right)^{\delta(T_m)} \left(\frac{R_{1,T_m}\rho}{R_{0,T_m} + R_{1,T_m}\rho} \right)^{1-\delta(T_m)}, \tag{3.8}$$

where R_{0,T_m} and R_{1,T_m} are the sizes of the two risk sets at time T_m. It is, in fact, a simple form of the familiar expression used for the Cox proportional hazards model, connected to the latter by the transformation $\rho = \exp\{-\beta\}$.

Currently, several vaccines against COVID-19 have been successfully tested in placebo controlled Phase III trials and, somewhat depending on the country, have then been approved by the relevant regulatory authorities for wider use in their respective population. In addition to the original efficacy trials, there are now several studies on the population level effectiveness of COVID-19 vaccines (e.g., Dagan et al. 2021, Vasileiou et al. 2021). On the other hand, in the present situation in which several vaccines that are demonstrably efficacious against both infection and the more serious forms of COVID-19 disease are available, it is difficult to find support, for a number of different reasons, to additional large-scale placebo controlled trials for testing new candidate vaccines, cf. Krause et al. 2020.

A possible alternative to such testing would be to use one or more of these existing vaccines as controls, and then make a comparative study. Such a design presents two major challenges, however. The first difficulty is demonstrated clearly by the Moderna study described briefly above: Of the approximately 15.000 individuals in the vaccine group only 11 were infected during the trial. If the candidate vaccine has at all comparable efficacy, as would naturally be desirable, the number of infected individuals in the vaccine group of a similar size, and assuming a comparable infection pressure in the study population, could not be expected to be much larger. With such small frequencies from both treatment arms in the trial, it would not be possible to arrive at a sufficiently firm conclusion concerning the desired target of *superiority* or *non-inferiority*, and this would be the case regardless of the statistical paradigm that were applied for such purpose.

To overcome this problem, it would therefore be almost mandatory to seek regulatory approval to a design in which healthy volunteers, some vaccinated by the candidate and some by an already approved vaccine, say *Vaccine**, used as a control treatment, are exposed to the virus under a carefully specified protocol. The possibility of a *human challenge* design, albeit with placebo controls, was already discussed at the time when no efficacious vaccine was available (World Health Organization 2020, Eyal et al. 2020, Richards 2020), and it is still considered relevant now (Eyal and Lipsitch 2021). One could anticipate that in a challenge trial, naturally depending on the level of viral exposure that would be applied, a much smaller number of

participants would be needed for reaching a statistically valid conclusion on comparability. If desired, such a design could be extended to involve more than a single candidate and/or control vaccine. Note that adaptive sequential recruitment and Bayesian decision making, as exemplified by Rule 2, would find here their natural place: It would not be necessary to fix the group sizes in advance; the trial could be run with newly recruited individuals until the desired level of certainty, as specified in the design, has been reached.

A second issue arising in the context of such a design concerns statistical modeling and inference in a situation in which information comes from different data sources: While the design may lead to an efficacy estimate where the candidate vaccine is compared to another in routine use, this estimate cannot be readily converted to a corresponding VE_S-estimate, where the candidate vaccine is compared to placebo. For practical consideration, this latter estimate could be the one of most interest. An approximate solution to this problem could be provided by assuming that the relative VE_S-efficacy measures obtained from different trials, viz. an 'old' trial for testing *Vaccine** vs. placebo, and the 'new' trial for testing the candidate vaccine vs. *Vaccine**, act multiplicatively on each other, which would correspond to the structure of the Cox proportional hazards model. This would then yield a synthetic VE_S-estimate for comparing the candidate vaccine to placebo, with a corresponding posterior derived by applying Bayesian inferential tools providing an uncertainty quantification. The relevance of this idea of combining estimates from different trials needs to be given careful scrutiny, however, and in particular since the dominant virus variant may have changed in between. This approach will be studied in more detail elsewhere.

4 Discussion

Clinical trials are an instrument for making informed decisions. In Phase II trials, the usual goal is to make a comparative evaluation on the success rates of one or more experimental treatments to a standard or control, and in multi-arm trials, also to each other. More successful treatments among the considered alternatives, if found, can then be selected for further study, possibly in Phase III.

With this as the stated goal for a trial, the conclusions should obviously be drawn as fast as possible, but not jumping ahead of the evidence provided by the acquired data. Both aspects can be accounted for by applying a suitable adaptive design, allowing for a continuous monitoring of the outcome data, and then utilizing in the execution of the trial the information that the data contain. Still, there is always the antagonism *Exploration* versus *Exploitation*: From the perspective of an individual patient in the trial, under postulated exchangeability, the optimal choice of treatment would be to receive the one with the largest current posterior mean of the success rate, as this would correspond to the highest predictive probability of treatment success. However, as demonstrated in Villar et al. (2015), this *Current Belief* (CB) strategy leads to a very low probability of ultimately detecting the best treatment arm among the considered alternatives and would therefore be a poor choice when considering the overall aims of the trial.

Finding an appropriate balance between these two competing interests is a core issue in the design and execution of clinical trials, and can realistically be made only in each concrete context. For example, in trials involving medical conditions such as uncomplicated urinary infections, or acute ear infections in children, use of balanced non-adaptive 1:1 randomization to both symp-

tomatic treatment and antibiotics groups appears fully reasonable. A very different example is provided by the famous ECMO trial on the use of the potentially life-saving technique of extracorporeal membrane oxygenation in treating newborn infants with severe respiratory failure (e.g., Bartlett et al. (1985), Wolfson (2003)). While statisticians advising clinical researchers have the responsibility of making available the best methods in their tool kit, there may well be overriding logistic, medical or ethical arguments which determine the final choice of the trial design. It has been even suggested that randomized clinical trials as such can present a scientific/ethical dilemma for clinical investigators, see Royall (1991).

Bayesian inferential methods are naturally suited to sequential decision making over time. In the present context, this involves deciding at each time point whether to continue accrual of more participants to the trial or to stop, either temporarily or permanently, and if such accrual is continued, selecting the treatment arm to which the next arriving participant is assigned. The current joint posterior distribution of the success parameters captures then the essential information in the data that is needed for such decisions.

The posterior probabilities used for formulating Rule 2, when considered as functions of the accumulated data D_n, can be viewed as test statistics in sequential tests of null hypotheses against corresponding alternatives. This link between the Bayesian and the frequentist inferential approaches makes it possible to compute, for the selected design parameters, the values of traditional performance criteria such as false positive rate and power. In the present approach, specifying a particular value for the trial size has no real theoretical bearing, and would serve mainly as an instrument for resource planning. Instead, the emphasis in the design is on making an appropriate choice of the operating characteristics, the ε's and δ, which control the execution of the trial, and on the direct consideration of posterior probabilities of events of the form $\{\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee\}$ and $\{\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee\}$ when monitoring outcome data from the trial.

An important difference to the methods based on classical hypothesis testing is that posterior probabilities, being conditioned on the observed data, are directly interpretable and meaningful concepts as such, without reference to their quantile value in a sampling distribution conditioned on the null. This is true regardless of whether the trial design applies adaptive treatment allocation and selection while the trial is in progress, or whether only a final posterior analysis is performed when an initially prescribed number of trial participants have been treated and their outcomes observed.

Large differences between the success parameters, if present, will often be detected early without need to wait until reaching a planned maximal trial size. On the other hand, if the joint posterior stems from an interim analysis, it forms a principled basis for predicting, in the form the consequent posterior predictive distribution, what may happen in the future if the trial is continued (e.g., Spiegelhalter et al. (1986), Yin et al. (2012)). Note, however, that future outcomes are uncertain even in the fictitious situation in which the true values of the success parameters were known. Therefore, from the perspective of decision making, the predictive distribution involves only "more uncertainty" than the posterior, not less.

Another advantage of the direct consideration of posterior probabilities is that the joint posterior of the success parameters may contain useful empirical evidence for further study even when no firm final conclusion from the trial has been made. This is in contrast to classical hypothesis testing, where, unless the observed significance level is below the selected α-level so that the stated null hypothesis is rejected, the conclusion from the trial remains hanging in mid-air, without providing much guidance on whether some parts of the study would perhaps

deserve further experimentation and consequent closer assessment.

The standard paradigm of null hypothesis significance testing (NHST), and particularly the version where the observed p-value is compared mechanistically to a selected α-level such as 0.05, have been criticised increasingly sharply in the recent statistical literature (e.g., Wasserstein and Lazar (2016), Greenland et al. (2016)). In spite of this, the corresponding strong emphasis on controlling the frequentist Type 1 error rate at a pre-specified fixed level has been largely adopted in the Bayesian clinical trials literature as well (e.g., Shi and Yin (2019), Stallard et al. (2020)). These error rates are conditional probabilities, evaluated from a sampling distribution under an assumed null hypothesis $\mathbb{Q}_{null}$ and in practice computed during the design stage when no actual outcome data from the trial are yet available. In contrast, in the Bayesian clinical trials methodology as outlined here, error control against false positives is performed continuously while the trial is run by applying bounds of the form $\mathbb{P}_\pi\big(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee \big| D_i^*\big) < \varepsilon_2$, where the considered posterior probabilities are conditioned on the currently available trial data D_i^*. For this reason, in our view, calibration of Bayesian trial designs on a selected fixed frequentist Type 1 error rate (e.g., Thall et al. (2015)) does not form a natural basis for comparing such designs. More generally, the role of testing a null hypothesis and the consequent emphasis on Type 1 error rate should not enjoy primacy over other relevant criteria in drawing concrete conclusions from a clinical trial (Greenland (2020)). Even posterior inferences alone are not sufficient for rational decision making in such a context, and should therefore optimally be combined with appropriately selected utility functions (e.g., D.V. Lindley in Grieve et al. (1994)).

If the trial is continued into Phase III, this can be done in a seamless fashion by using the joint posterior of the selected treatments from Phase II as the prior for Phase III. In particular, if some treatment arms have been dropped during Phase II, the trial can be continued into Phase III as if the selected remaining treatments had been the only ones present from the very beginning. Recall, however, from the remarks made in Section 2 that such treatment elimination, as encoded into Rule 2, contains a violation of the likelihood principle.

If Rule 2 is employed in Phase III, and considering that Phase III trials are commonly targeted at providing confirmatory evidence on the safety and efficacy of the new experimental treatment against the current standard treatment used as a control, it may be a reasonable idea to lower the threshold values ε_1 and ε_2 from their levels used in Phase II, and thereby apply stricter criteria for final approval.

No statistical method is uniformly superior to others on all accounts. Important criticisms against the use of adaptive randomization in clinical trials have been presented, e.g., in Thall et al. (2015). There, computer simulations were used to compare adaptive patient allocation based on Thompson's rule (Thompson (1933), Villar et al. (2015)) in its original and fractional forms, in a two-arm 200-patient clinical trial, to an equally randomized group sequential design. The main argument against using methods applying adaptive randomization was their potential instability, that is, there was, in the authors' view, unacceptably large (frequentist) $\mathbb{Q}$-probability of allocating more patients to the inferior treatment arm, the opposite of the intended effect. Although these simulations were restricted to Thompson's rule, the criticism in Thall et al. (2015) was directed more generally towards applying adaptive randomization and would therefore in principle apply to our Rules 1 and 2 as well. The results from our simulation experiments, shown in graphical form in Figures S3, S4, S10 and S11 in the Supplement, do not support such a firm negative conclusion, however. This holds provided that the deviations from balance in the opposite directions are not weighted completely differently, and particularly so if the possibility of actually dropping a treatment arm is deferred to a somewhat later time

from the beginning of the trial. A precautionary approach to the design, from a frequentist perspective, could apply a sandwich structure, starting with a symmetric burn-in, followed by an adaptive treatment allocation realized by Rule 1 or Thompson's rule, and finally coupling in Rule 2 for actual treatment selection.

Another criticism presented in Thall et al. (2015) was that, for trial data collected from an adaptive design, the considered tests had lower power than in a corresponding equally randomized design, and particularly so if the tests were calibrated to have the same Type 1 error rate. This question is discussed in subsections A.1.3 and C of the Supplement. In these experiments, adaptive treatment allocation methods based on Rule 1 (a) and (b), and on Thompson's rule with fractional power $\kappa = 0.25$, demonstrated frequentist performance quite comparable to what was observed when applying the fully symmetric block randomization design (d).

All adaptive methods favoring treatment arms with relatively more successes in the past will inevitably introduce some degree of bias in the estimation of the respective success parameters, see Bauer and Köhne (1994) and Villar et al. (2015). A comprehensive review of the topic is provided in Robertson et al. (2021). We have only considered this matter briefly in the simulation experiments described in the Supplement, and instead emphasized the, in our view, more important aspect of the mutual comparison of the performance of different treatment arms in the trial. All biases in these experiments were relatively small and in the same direction, downward, and therefore unlikely to have had a strong influence on the conclusions that were drawn.

Our main focus has been on trials with binary outcome data, where individual outcomes could be measured soon after the treatment was delivered. More complicated data situations were outlined in Section 4. The important case of normally distributed outcome data was by-passed here; there is a large body of literature relating to it, e.g., Spiegelhalter et al. (1994) and Gsponer et al. (2014). A complication with the normal distribution is that, unless the variance is known to a good approximation already from before, there are two free parameters to be estimated for each treatment. If a suitable yardstick at the start is missing, many observations are needed before it becomes possible to separate the statistical variability of the outcome measures from a true difference between treatment effects.

In principle, the logic of Rules 1 and 2 remains valid and these rules can be applied for different types of outcome data, requiring only the ability to update the posterior distributions of the model parameters of interest when more data become available. The computation of the posteriors is naturally much less involved if the prior and the likelihood are conjugate to each other. Vague priors, or models containing more than a single parameter to be updated, will necessarily require more outcome data before adaptive actions based on Rule 1 or Rule 2 can kick in.

If such updating is not done systematically after each individual outcome is measured, for example, for logistic reasons, but less frequently in batches, Rule 1 and Rule 2 can still be used at the times at which the batches are completed. The same holds if updating is done at regularly spaced points in time. Such thinning of the data sequence has the effect that some of the actions that would have been otherwise implied by Rule 1 and Rule 2 are then postponed to a later time or even omitted. In designing a concrete trial, one then needs to find an appropriate balance between, on one hand, the costs saved in logistics and computation, and on the other, the resulting loss of information and the effect this may have to the quality of the inferences that can be drawn.

5 Declarations

Ethics approval and consent to participate: Not applicable (only simulated data were used).

Consent for publication: Not applicable (no permission form the University is needed for submission or publication).

Availability of data and materials: The R package *barts* written by Mikko Marttila generating simulated datasets and implementing the methods is freely available at https://github.com/Orion-Corporation/barts

Competing interests: The authors declare that they have no competing interests.

Funding: No separate funding for the work was provided.

Authors' contributions: EA conceived the idea for the paper, to which DG contributed some key methodological insights. EA wrote the text of the paper. DG wrote the computer code for the simulation experiments reported in the Supplement, carried out the numerical computations and drew the Figures. Both authors approved the contents of the manuscript.

Acknowledgements:

We are grateful to Jukka Ollgren for comments and encouragement, and to Mikko Marttila for useful suggestions on the text. E.A. thanks Arnoldo Frigessi and David Swanson for support and useful discussions during an early stage of this work.

References

Thompson, William R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4, pp. 285–294.

Pocock, Stuart J. (1977). "Group sequential methods in the design and analysis of clinical trials". In: *Biometrika* 64.2, pp. 191–199.

O'Brien, Peter C. and Thomas R. Fleming (1979). "A multiple testing procedure for clinical trials". In: *Biometrics*, pp. 549–556.

Flühler, Hannes et al. (1983). "Bayesian approach to bioequivalence assessment: an example." In: *Journal of pharmaceutical sciences* 72 10, pp. 1178–81.

Berger, James O. and Robert L. Wolpert (1984). *The likelihood principle*. Vol. 6. Institute of Mathematical Statistics Lecture Notes—Monograph Series. Institute of Mathematical Statistics, Hayward, CA, pp. xi+206.

Bartlett, Robert H. et al. (1985). "Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study". In: *Pediatrics* 76.4, pp. 479–487. eprint: https://pediatrics.aappublications.org/content/76/4/479.full.pdf.

Berry, Donald A. (1985). "Interim analyses in clinical trials: classical vs. Bayesian approaches". In: *Statistics in medicine* 4.4, pp. 521–526.

Spiegelhalter, David J., Laurence S. Freedman, and Patrick R. Blackburn (1986). "Monitoring clinical trials: Conditional or predictive power?" In: *Controlled Clinical Trials* 7.1, pp. 8 –17.

Berger, James O. and Donald A. Berry (1988). "Statistical analysis and the illusion of objectivity". In: *American Scientist* 76.2, pp. 159–165.

Royall, Richard M. (1991). "Ethics and statistics in randomized clinical trials". In: *Statistical Science*, pp. 52–62.

Bauer, P. and K. Köhne (1994). "Evaluation of experiments with adaptive interim analyses." In: *Biometrics* 50 4, pp. 1029–41.

Demets, David L. and K. K. Gordon Lan (1994). "Interim analysis: the alpha spending function approach". In: *Statistics in medicine* 13.13-14, pp. 1341–1352.

Grieve, A. P. et al. (1994). "Bayesian approaches to randomized trials, Discussion". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 157.3, pp. 387–416.

Lewis, Roger J. and Donald A. Berry (1994). "Group sequential clinical trials: a classical evaluation of Bayesian decision-theoretic designs". In: *Journal of the American Statistical Association* 89.428, pp. 1528–1534.

Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K. B. Parmar (1994). "Bayesian approaches to randomized trials". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 157.3, pp. 357–387.

Thall, Peter F. and Richard Simon (June 1994). "Practical Bayesian Guidelines for Phase IIB Clinical Trials". In: *Biometrics* 50.2, p. 337.

Jennison, Christopher and Bruce W. Turnbull (Sept. 1999). *Group sequential tests with applications to clinical trials*. English. Chapman & Hall/CRC Interdisciplinary Statistics. UK United Kingdom: Chapman & Hall.

Yip, P. and Q. Chen (2000). "A Partial Likelihood Estimator of Vaccine Efficacy". In: *Australian and New Zealand Journal of Statistics* 42, pp. 367–374.

Wolfson, Philip J. (2003). "The development and use of extracorporeal membrane oxygenation in neonates". In: *The Annals of thoracic surgery* 76.6, S2224–S2229.

Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Vol. 13. John Wiley & Sons.

Berry, Donald A. (Jan. 2006). "Bayesian clinical trials". In: *Nature Reviews Drug Discovery* 5.1, pp. 27–36.

Thall, Peter and J. Wathen (Apr. 2007). "Practical Bayesian Adaptive Randomization in Clinical Trials". In: *European journal of cancer (Oxford, England : 1990)* 43, pp. 859–66.

Chow, Shein-Chung and Mark Chang (2008). "Adaptive design methods in clinical trials–a review". In: *Orphanet journal of rare diseases* 3.1, pp. 1–13.

Press, William H. (2009). "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research". In: *Proceedings of the National Academy of Sciences* 106.52, pp. 22387–22392. eprint: https://www.pnas.org/content/106/52/22387.full.pdf.

Halloran, M. Elizabeth et al. (2010). *Design and analysis of vaccine studies*. Vol. 18. Springer.

Mahajan, Rajiv and Kapil Gupta (2010). "Adaptive design clinical trials: Methodology, challenges and prospect". In: *Indian journal of pharmacology* 42.4, p. 201.

Berry, Donald A. (2011). "Adaptive Clinical Trials: The Promise and the Caution". In: *Journal of Clinical Oncology* 29.6. PMID: 21172875, pp. 606–609. eprint: https://doi.org/10.1200/JCO.2010.32.2685.

Berry, Scott M. et al. (2011). *Bayesian adaptive methods for clinical trials*. Vol. 38. Chapman & Hall/CRC Biostatistics Series. With a foreword by David J. Spiegelhalter. CRC Press, Boca Raton, FL, pp. xviii+305.

Lee, J. Jack and Caleb T. Chu (2012). "Bayesian clinical trials in action." In: *Statistics in medicine* 31 25, pp. 2955–72.

Xie, Fang, Yuan Ji, and Lothar Tremmel (2012). "A Bayesian adaptive design for multi-dose, randomized, placebo-controlled phase I/II trials". In: *Contemporary clinical trials* 33.4, pp. 739–748.

Yin, Guosheng, Nan Chen, and J. Jack Lee (2012). "Phase II trial design with Bayesian adaptive randomization and predictive probability". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.2, pp. 219–235.

Zaslavsky, Boris G. (Sept. 2012). "Bayesian Hypothesis Testing in Two-Arm Trials with Dichotomous Outcomes". In: *Biometrics* 69.1, pp. 157–163.

Chow, Shein-Chung (2014). "Adaptive clinical trial design". In: *Annual review of medicine* 65, pp. 405–415.

Gsponer, Thomas et al. (2014). "A practical guide to Bayesian group sequential designs". In: *Pharmaceutical statistics* 13.1, pp. 71–80.

Thall, Peter F., Patricia S. Fox, and J. Kyle Wathen (2015). "Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials." In: *Annals of oncology : official journal of the European Society for Medical Oncology* 26 8, pp. 1621–8.

Villar, Sofía S., Jack Bowden, and James Wason (May 2015). "Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges". In: *Statistical Science* 30.2, pp. 199–215.

Chang, M. and J. Balser (2016). "Adaptive Design-Recent Advancement in Clinical Trials". In: *J Bioanal Biostat* 1.1, p. 14.

Greenland, Sander et al. (2016). "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". In: *European journal of epidemiology* 31.4, pp. 337–350.

Grieve, Andrew P. (2016). "Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development." In: *Pharmaceutical statistics* 15 2, pp. 96–108.

Jacob, Louis et al. (June 2016). "Evaluation of a multi-arm multi-stage Bayesian design for phase II drug selection trials – an example in hemato-oncology". In: *BMC Medical Research Methodology* 16.1.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). "The ASA Statement on p-Values: Context, Process, and Purpose". In: *The American Statistician* 70.2, pp. 129–133. eprint: https://doi.org/10.1080/00031305.2016.1154108.

Müller, Peter, Yanxun Xu, and Peter F. Thall (2017). "Clinical Trial Design as a Decision Problem." In: *Applied stochastic models in business and industry* 33 3, pp. 296–301.

Yin, Guosheng, Chi Kin Lam, and Haolun Shi (2017). "Bayesian randomized clinical trials: From fixed to adaptive design". In: *Contemporary clinical trials* 59, pp. 77–86.

Yuan, Ying, Hoang Q. Nguyen, and Peter F. Thall (2017). *Bayesian designs for phase I-II clinical trials*. CRC Press.

Alban, Andres, Stephen E. Chick, and Martin Forster (2018). "Extending a Bayesian decision-theoretic approach to a value-based sequential clinical trial design". In: *2018 Winter Simulation Conference (WSC)*, pp. 2459–2470.

Pallmann, Philip et al. (Feb. 2018). "Adaptive designs in clinical trials: why use them, and how to run and report them". In: *BMC Medicine* 16.1.

Atkinson, Anthony C and Atanu Biswas (2019). *Randomised response-adaptive designs in clinical trials*. Chapman and Hall/CRC.

Ruberg, Stephen J. et al. (2019). "Inference and Decision Making for 21st-Century Drug Development and Approval". In: *The American Statistician* 73.sup1, pp. 319–327. eprint: https://doi.org/10.1080/00031305.2019.1566091.

Shi, Haolun, Guosheng Yin, et al. (2019). "Control of type I error rates in Bayesian sequential designs". In: *Bayesian Analysis* 14.2, pp. 399–425.

Eyal, Nir, Marc Lipsitch, and Peter G Smith (2020). "Human challenge studies to accelerate coronavirus vaccine licensure". In: *The Journal of infectious diseases* 221.11, pp. 1752–1756.

Greenland, Sander (2020). "Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons". In: *Paediatric and Perinatal Epidemiology*.

Krause, Philip et al. (2020). "COVID-19 vaccine trials should seek worthwhile efficacy". In: *The Lancet* 396.10253, pp. 741–743.

Moderna Inc. (2020). *Moderna announces Primary Efficacy analysis in Phase 3 COVE study for Its Covid-19 Vaccine candidate and Filing today with U.S. FDA for emergency use authorization*. https://investors.modernatx.com/news-releases/news-release-details/moderna-announces-primary-efficacy-analysis-phase-3-cove-study.

Richards, Adair D (2020). "Ethical guidelines for deliberately infecting volunteers with COVID-19". In: *Journal of Medical Ethics* 46.8, pp. 502–504. eprint: https://jme.bmj.com/content/46/8/502.full.pdf.

Stallard, Nigel et al. (2020). "Comparison of Bayesian and frequentist group-sequential clinical trial designs". In: *BMC medical research methodology* 20.1, pp. 1–14.

World Health Organization (2020). *Key criteria for the Ethical acceptability of Covid-19 human challenge studies*. https://www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_criteria-2020.1.

Dagan, Noa et al. (2021). "BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting". In: *New England Journal of Medicine* 384.15, pp. 1412–1423.

Eyal, Nir and Marc Lipsitch (2021). "How to test SARS-CoV-2 vaccines ethically even after one is available". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*.

Giovagnoli, Alessandra (2021). "The Bayesian Design of Adaptive Clinical Trials". In: *International Journal of Environmental Research and Public Health* 18.2.

Marttila, Mikko, Elja Arjas, and Dario Gasbarra (2021). *barts: Bayesian adaptive rules for treatment selection. R package version 0.0.1*. https://github.com/Orion-Corporation/barts.

Robertson, David S. et al. (2021). *Point estimation for adaptive trial designs*. arXiv: 2105.08836.

Vasileiou, Eleftheria et al. (2021). "Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study". In: *The Lancet* 397.10285, pp. 1646–1657.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- mainsupplement1.pdf