BMC Medical Research Methodology

RESEARCH Open Access

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

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

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 arms 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, and treatment selection, removing an arm from the trial permanently. 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 *'barts'*.

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

Keywords: Superiority trial, Phase II, Phase III, Adaptive design, Likelihood principle, Posterior inference, Decision rule, Frequentist performance, Binary data, Time-to-event data, Vaccine efficacy trial

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

*Correspondence: elja.arjas@helsinki.fi
[1] University of Helsinki, Helsinki, Finland
[2] University of Oslo, Oslo, Norway

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 [1]. More general reviews of adaptive clinical trial designs, from the perspective of classical inference, can be found in, e.g., Chow and Chang [2], Mahajan and Gupta [3], Chow [4], Chang and Balser [5], Pallmann et al. [6] and Atkinson and Biswas [7]. 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 [8], O'Brien and Fleming [9], Demets and Lan [10]), 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 [11], Flühler et al. [12], Berry [13], Spiegelhalter et al. [14], Berger and Berry [15], Spiegelhalter et al. [16] and Thall and Simon [17]; for a brief historical account and a large number of references, see Grieve [18]. Comprehensive expositions of the topic are provided in the monographs Spiegelhalter et al. [19], Berry et al. [20] and Yuan et al. [21].

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 [22].

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 [23], in Berry [24], Lee and Chu [25], and more recently, in Yin et al. [26], Ruberg et al. [27] and Giovagnoli [28]. 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.

Much of the recent literature on adaptive methods in clinical trials falls into two categories: adaptive randomization (AR) designs, also called response adaptive randomization (RAR), and multi-arm multi-stage (MAMS) designs. In AR, the patients are randomized to the different treatment arms sequentially, with probabilities updated from the preceding outcome data either continuously or at the times of pre-planned interim analyses. For reviews on AR designs, see Chow and Chang [2] and Robertson et al. [29]. Villar et al. [30] 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. Of particular interest to us are papers dealing with Bayesian versions of AR, where the randomization probabilities are updated directly by applying Bayes' rule, e.g., Trippa et al. [31], Wathen and Thall [32], Wathen and Thall [33], Viele et al. [34], Viele et al. [35] and Bassi et al. [36].

MAMS designs, on the other hand, aim at selecting the best treatments, or even the single best if there is one, of several that are tested in a multi-arm trial. This is often done indirectly by applying pre-specified stopping boundaries, to determine whether a considered treatment should be dropped. Recent contributions to such designs include Wason and Jaki [37], Wason and Trippa [38], Jacob et al. [39], Wathen and Thall [32], Yu et al. [40], Ryan et al. [33] and Ryan et al. [35].

Unfortunately, general results on optimal strategies are largely lacking and their application in practice often infeasible because of computational complexity; however, see Press [41] and Yu et al. [40]. Recently, simulation based approximations have been used for applying Bayesian decision theory in the clinical trials context, e.g., Müller et al. [42], Yuan et al. [21], Alban et al. [43] and Bassi et al. [36].

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 usually 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 predefined 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.

We consider first, in The case of Bernoulli outcomes section, the simple situation in which the outcomes are binary, and they can be observed soon after the treatment has been delivered. In Extensions for handling delayed outcome data section, 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 Discussion section. 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. [30] but apply the adaptive methods introduced in The case of Bernoulli outcomes section. The 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 called *barts*, Marttila et al. [44], that facilitates the simulation of clinical trials with adaptive treatment allocation and selection.

The case of Bernoulli outcomes

An adaptive method for treatment allocation: rule 1

As in Villar et al. [30] and Jacob et al. [39] 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 allocation 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 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, for later use, $\boldsymbol{\theta}_\vee = \max\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

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 Extensions for handling delayed outcome data section.

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 $r = ((r(1), r(2), ..., r(K+1)), (r(K+2), r(K+3), ..., r(2(K+1)), ...)$ of randomized blocks of length $K + 1$, where the blocks are independent uniformly distributed random permutations of the treatment arm indexes $0, 1, ..., K$.

Allocation 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 allocated 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

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

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

$$N^{-1}(1) = 1;$$

$$N^{-1}(i) = \min\left\{n > N^{-1}(i-1) : I_{r(n),n-1} = 1\right\}, i > 1. \tag{2}$$

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 allocated, 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.

For $n \geq 1$, the activity states I_n will be determined in an inductive manner during the trial, and will then depend, according to criteria specified below, on the earlier treatment allocations and on the corresponding observed outcomes. 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 $i \leq N(n)$.

To explain the algorithm, suppose that initially, for list index value $n = 0$, all treatment arms are active so that $I_0 = (I_{0,0}, I_{1,0}, ..., I_{K,0}) = (1, 1, ..., 1)$. More generally, their activity states can be determined from the prior of $\boldsymbol{\theta}$. The first participant recruited to the trial, indexed by $i = 1$, is allocated to arm $r(1)$ and then given the respective treatment $A(1) = r(1)$. The outcome Y_1 is then measured and included, as part, in the data D_1. After this, the value of the activity vector I_0 is updated into I_1 as follows: If $r(1) = k$ is an experimental treatment arm, we let $I_{k,1} = 0$ if $\mathbb{P}_\pi(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee \big| D_1) < \varepsilon$, and otherwise $I_{k,1} = 1$. Similarly, if $r(1) = 0$ is the control arm, we let $I_{0,1} = 0$ if $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee \big| D_1) < \varepsilon$, and otherwise $I_{0,1} = 1$. In a 2-arm trial obviously $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_1) = \mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_1 | D_1)$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_1 = \boldsymbol{\theta}_\vee | D_1) = \mathbb{P}_\pi(\boldsymbol{\theta}_1 \geq \boldsymbol{\theta}_0 | D_1)$.

Here the threshold values $\varepsilon > 0$ and $\delta \geq 0$ are selected design parameters of the algorithm. A smaller value of ε reflects then a more conservative attitude towards moving a treatment arm 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.

The general step of the induction follows the same pattern: Consider a list index $n \geq 1$. 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, if $r(n) = 0$ is the control arm, we let $I_{0,n} = 0$ if $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee \big| D_n) < \varepsilon$, and otherwise $I_{0,n} = 1$. The earlier activity state vector I_{n-1} has thereby been updated to a new value I_n. After this, the value of the list index is increased by 1, from n to $n+1$. Again, in a 2-arm trial, $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_n) = \mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_1 | D_n)$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_1 = \boldsymbol{\theta}_\vee | D_n) = \mathbb{P}_\pi(\boldsymbol{\theta}_1 \geq \boldsymbol{\theta}_0 | D_n)$.

A pseudocode of this algorithm, called *BARTA* (for *Bayesian adaptive rule for treatment allocation*), is provided in Section A of the Supplement.

As a byproduct, successive applications of *BARTA* give us an explicit expression for the likelihood $L(\theta | D_n) = L_n(\theta), n \geq 1$, arising from observing data D_n. According to this rule, the likelihood expression $L(\theta | D_n)$ is updated only at values of n at which $I_{r(n),n} = 1$, and is then performed 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

$$
\begin{aligned}
L(\theta | D_n) &= \prod_{m=1}^{n} \theta_{r(m)}^{I_{r(m),m} Y_{N(m)}} \left(1 - \theta_{r(m)}\right)^{I_{r(m),m}(1 - Y_{N(m)})} \\
&= \prod_{k=0}^{K} \theta_k^{N_{k,1}(n)} (1 - \theta_k)^{N_{k,0}(n)} .
\end{aligned} \tag{3}
$$

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

$$
\begin{aligned}
N_{k,1}(n) &= \sum_{m=1}^{n} I_{k,m} 1_{\{Y_{N(m)}=1\}}, \\
N_{k,0}(n) &= \sum_{m=1}^{n} I_{k,m} 1_{\{Y_{N(m)}=0\}}, \ 0 \leq k \leq K, \ n \geq 1,
\end{aligned} \tag{4}
$$

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. [30].

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

$$
\begin{aligned}
S_k(i) &= \max \left\{ N_{k,1}(n) : N(n) \leq i \right\}, \\
F_k(i) &= \max \left\{ N_{k,0}(n) : N(n) \leq i \right\}, \ 0 \leq k \leq K,
\end{aligned} \tag{5}
$$

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

$$
S(i) = \sum_{k=0}^{K} S_k(i), \ F(i) = \sum_{k=0}^{K} F_k(i) \tag{6}
$$

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 [17]) 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 (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:

$$p\left(\theta_k|D^*_{i,k}\right) = Beta\left(\theta_k|\alpha_k + S_k(i), \beta_k + F_k(i)\right),\ i \geq 1, \\ k = 0, 1, \ldots, K. \tag{7}$$

This, together with the product form of the likelihood (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. [39], or by performing Monte Carlo sampling from this distribution; see also Zaslavsky [45]. In our numerical examples in Extensions for handling delayed outcome data section we have applied this latter possibility.

While application of *BARTA* 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, and given that the priors for different treatments have been assumed to be independent, 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.

BARTA has much in common with the adaptive randomization (AR) methods considered in the literature, briefly reviewed in the Introduction. The similarity to *BARTA* is particularly close to the Bayesian versions of AR, where such adaptive updating is performed by a direct application of Bayes' rule. The idea of employing a single initial block randomization $r = ((r(1), r(2), ..., r(K+1)), (r(K+2), r(K+3), ..., r(2(K+1)), ...)$, together with considering the treatment arms to be momentarily either active or dormant, appears to be novel, however. As a result, once r and the operating characteristics ε and δ have been fixed, no further 'coin tossing' is performed during the trial since each treatment allocation of a new patient is fully determined by r and the posterior probabilities computed from the preceding outcome data. In principle, the list r could be even made available in advance to all parties concerned; if this is not done, it will nevertheless be easy to check afterwards that the selected allocations were consistent with the design.

Note also that, in *BARTA*, all currently active treatment arms in a block are considered symmetrically, with exactly one patient allocated to each active treatment; after this has been done, the algorithm proceeds to considering the next permutation of the $K+1$ treatments, etc. Unless this is not regulated differently by the prior, fully balanced block randomization of all $K+1$ treatments, reminiscent to a burn-in, is applied during the early part of the trial, until there is one arm that is made dormant.

To compare, most AR-designs suggested in the literature update the randomization probabilities only at the times of a few pre-planned interim analyses, whereas in the prototype version of *BARTA*, the posterior probabilities for determining the activity states are computed after every new measured outcome. If such a continuous monitoring scheme is difficult to employ in practice, for example, for logistic reasons, it can in principle be replaced by any more appropriate non-informative stopping rule. However, the results in Viele et al. [35] suggest that, in AR designs, more frequent checks and updates are advantageous from the perspective of several different performance measures, and the same is likely to hold for *BARTA* as well.

Thompson's rule. We restrict the numerical comparison of *BARTA* to a single AR design, by considering the historically oldest, classical Thompson's rule ([11], see also, e.g., Thall et al. [46], Villar et al. [30]). In its standard version, 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. The results from these comparative simulation experiments are given in Sections B, C and D of the Supplement.

An adaptive method for treatment selection: BARTS

While an open end recipe such as *BARTA* 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 *BARTA* with an optional possibility to conclusively terminate the accrual of additional participants to the less successful treatment arms. The consequent algorithm *BARTS* (for *Bayesian adaptive rule for treatment selection*), is provided in the form of a pseudocode in Section A of the Supplement. The treatment allocation procedure is identical to that in *BARTA*, and makes a treatment arm dormant if its performance, according to pre-specified criteria, is assessed to be poor when compared to the current best. *BARTS* does the same, but will actually drop a treatment arm permanently if such judgement holds with respect to an even stricter criterion. *BARTS* can therefore be said to be an adaptation of corresponding ideas and definitions in, e.g., Thall and Wathen [47], Berry et al. [20], Xie et al. [48], Jacob et al. [39] and Wathen and Thall [32]. In the commonly adopted terminology of adaptive designs, it can be said to combine elements from different versions of AR and MAMS designs.

After every new observed outcome, the algorithm of *BARTS* 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.

Next, we explain how *BARTS* works; for the exact definition of this algorithm, see the pseudocode in the Supplement.

The posterior probability $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \middle| D_n\right)$ for an experimental arm k expresses how likely it is, given the currently available data, that its response rate $\boldsymbol{\theta}_k$ exceeds a pre-specified level θ_{low} of *minimum required treatment response rate* (MRT), e.g., Xie et al. [48]. The first criterion in *BARTS* then says that if this probability is below a selected threshold value ε_1, treatment k is dropped from the trial. For the control arm, the corresponding comparison is based on the posterior probability $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \theta_{low} \middle| D_n\right)$, thereby involving an extra safety margin δ against accidental removal. 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 *BARTS* will obviously not be active if either $\theta_{low} = 0$ or $\varepsilon_1 = 0$.

The second criterion in *BARTS* 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 $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right)$ for the experimental arms and $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right)$ 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 $\left\{\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell\right\}$ 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 $\delta = 0$, whereas a sufficiently large value for δ would make it impossible to drop the control arm. This entire mechanism of eliminating treatments based on mutual comparisons is inactivated by letting $\varepsilon_2 = 0$.

For later use, we denote by $n_{k,last}$ the largest value of the list index for which treatment arm k is still left in the trial, $0 \leq k \leq K$, and by $N_{k,last} = N(n_{k,last})$ the index of the last patient who got treatment k.

The third criterion of *BARTS* copies *BARTA*: An experimental treatment arm $k \in \mathbb{T}$ is made dormant if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon$, and the control arm if $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\theta}_\ell \middle| D_n\right) < \varepsilon$, where ε is a selected threshold. For this part of *BARTS* to function in a non-trivial 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 moving a treatment arm to the dormant state is ruled out, and if $\varepsilon_1 = \varepsilon_2 = 0$, then *BARTS* is easily seen to collapse into the simpler rule *BARTA*. 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 *BARTS* 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 *BARTA*, this is commonly viewed as the *minimal clinically important difference* (MCID) in the trial. Employing such a positive threshold value when comparing the control to the experimental treatments reflects the idea that the design should be more conservative towards moving the control arm to the dormant state, let alone dropping it from the trial for good.

Once selected, the design parameters $\varepsilon, \varepsilon_1$ and ε_2 in applying *BARTS*, 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 $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} \middle| D_{n_{k,last}}\right) < \varepsilon_1$. 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 *BARTS*, the trial ends with a negative result, *futility*, e.g. Thall and Wathen [47]. 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 *BARTS*, with the goal of ultimately identifying the one with the highest response rate.

As remarked earlier, the application of *BARTS* is optional. If it is not enforced, *BARTA* is open ended and will only control the allocation 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 *BARTA* was previously employed or not, the posterior probabilities $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k \geq \theta_{low} | D^*_{N_{\max}}\right)$, $\mathbb{P}_\pi\left(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D^*_{N_{\max}}\right)$ and $\mathbb{P}_\pi\left(\boldsymbol{\theta}_k = \boldsymbol{\theta}_\vee | D^*_{N_{\max}}\right)$ 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.

The above version of *BARTS* is intended to be used in *superiority trials*, where the goal is to select, if possible, the best treatment among those $K + 1$ considered in the trial. It is another matter whether the phrase 'dropping a treatment arm' should then be understood literally or not. For example, in a 2-arm trial, one is supposed to keep track on the posterior probabilities of the form $\mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_\vee | D_n) = \mathbb{P}_\pi(\boldsymbol{\theta}_0 + \delta \geq \boldsymbol{\theta}_1 | D_n)$ and $\mathbb{P}_\pi(\boldsymbol{\theta}_1 = \boldsymbol{\theta}_\vee | D_n) = \mathbb{P}_\pi(\boldsymbol{\theta}_1 \geq \boldsymbol{\theta}_0 | D_n)$, and then drop an arm if either of them falls below the selected threshold value ε_2. In reality, dropping the control may only mean that the experimental arm is selected for further study, perhaps in phase III. As a reviewer of this paper has pointed out, what is proposed in *BARTS* is not a *drop-the-losers* type approach, as the latter, often used in group-sequential selection, involves treatment ranking, e.g., Gerber, Gsponer, et al. [49].

This terminology is even less fitting if *BARTS* is modified to be appropriate for *non-inferiority* or *equivalence* trials, e.g. Lesaffre [50]. For example, in a 2-arm trial for the former purpose, one would be led to considering, for some $\delta > 0$, posterior probabilities of the form $\mathbb{P}_\pi(\boldsymbol{\theta}_0 - \delta \geq \boldsymbol{\theta}_1 | D_n)$, and then conclude non-inferiority if such probabilities would fall below a selected ε_2.

Finally, one should note that, while *BARTA* is compatible with the likelihood principle, *BARTS* 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 *BARTS*, although it has an obvious practical appeal in running a clinical trial, it also renders properties such as standard Bayesian consistency inapplicable.

A frequentist perspective. A different perspective to the application of *BARTS* 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 [2]).

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 *BARTS* defining the stopping boundaries. In the case $K = 1$, they correspond to considering two overlapping hypotheses (e.g., Richards [51]), 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. Similar remarks would hold if, as remarked above, *BARTS* were modified to be used in a non-inferiority of equivalence trial.

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 *BARTA* and *BARTS*.

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. A more extensive 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 with a 4-arm trial.

To give just one example of the many frequentist measures considered in the Supplement, we reproduce here Fig. S3, in Fig. 1. It shows, in the top part, the cumulative distribution functions (CDFs) of $N_1(200)$, the number of patients, of the first 200 allocated by *BARTA* to the experimental treatment, in a 2-arm trial when varying the values of the design parameters ε and δ and considering two different data generating models, $\mathbb{Q}_{null}$ and $\mathbb{Q}_{alt}$. The bottom part makes similar comparisons for $S(200)$, the total number of successes from both treatments combined. Also shown are the CDFs of these variables when adaptive treatment allocation of patients was applied by using Thompson's rule with different values of the fractional exponent κ.

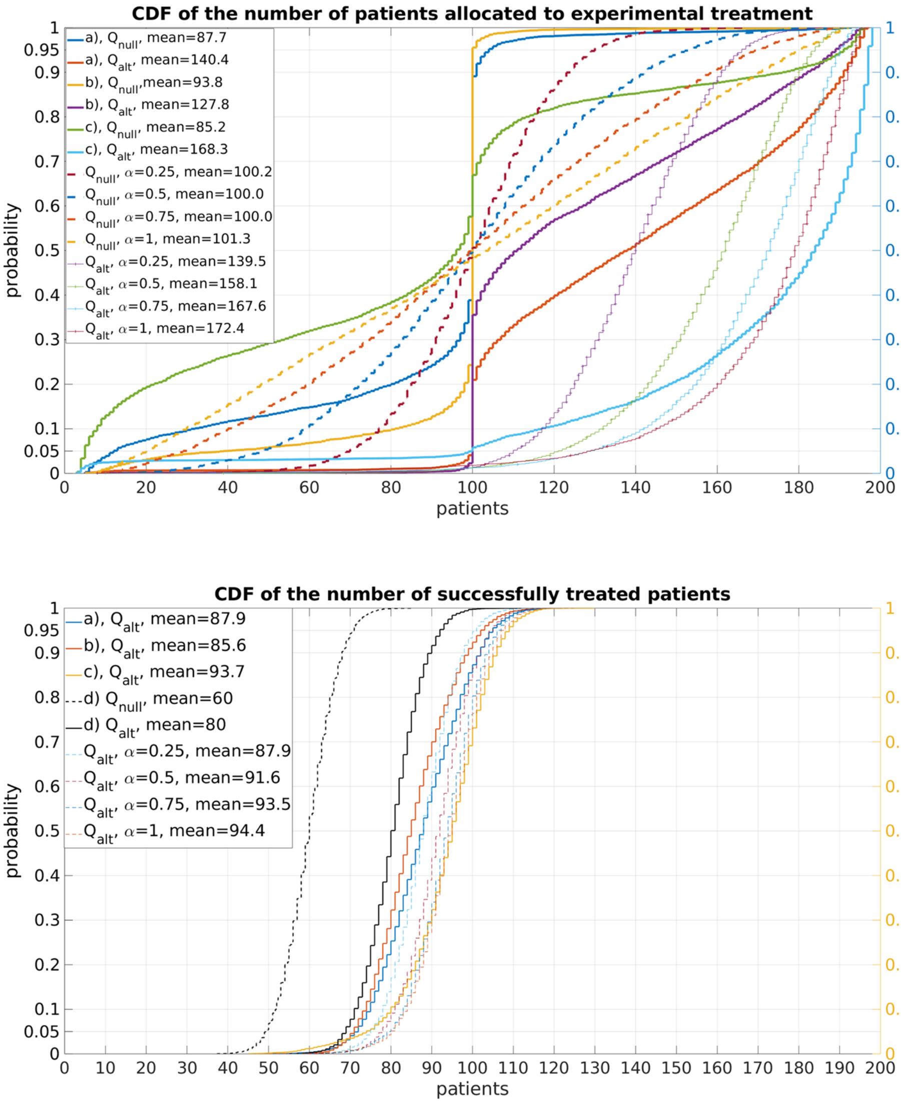

Fig. 1 Effect of the choice of the design parameters ε and δ in *BARTA* on the number of patients allocated to the experimental treatment and on the total number of treatment successes. Cumulative distribution functions of $N_1(200)$ (top) and $S(200)$ (bottom) are shown, based on 5000 simulated data sets, under $\mathbb{Q}_{null}$ with true parameter values $\theta_0 = \theta_1 = 0.3$ and $\mathbb{Q}_{alt}$ with values $\theta_0 = 0.3, \theta_1 = 0.5$. Three combinations of the design parameters were used: (a) $\varepsilon = 0.1, \delta = 0.1$, (b) $\varepsilon = 0.05, \delta = 0.1$, (c) $\varepsilon = 0.2, \delta = 0.05$. In addition, (d) represents a completely symmetric treatment allocation. For comparison we also plot the corresponding CDF under the alternative hypothesis obtained by using fractional Thompson's rule with respective parameters $\kappa = 0.25, 0.5, 0.75$ and 1

Perhaps of most interest here is to note, from the top part of Fig. 1, how the application of *BARTA*, and particularly under $\mathbb{Q}_{null}$ in which case the treatment arms have the same true response rate, leads to often allocating exactly half of the patients to both treatment arms; this happens in trial runs during which the dormant state had not been entered even once. For Thompson's rule, although the distribution of $N_1(200)$ under $\mathbb{Q}_{null}$ is symmetric, it has a large variance, signalling corresponding instability in treatment allocation. For additional comments on Fig. 1, see the Supplement.

Finally, 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 Figs. S1 and S7 in the Supplement. The same holds also when declaring the final results from a trial that was carried out. 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.

Extensions for handling delayed outcome data

Data of the kind considered in The case of Bernoulli outcomes section, 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 a vaccine efficacy study. 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 The case of Bernoulli outcomes section are considered briefly below.

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 (4), let

$$N_{k,1}\,(t) = \sum_{i=1}^{N(t)} C_i\,(t)\,1_{\{A_i=k, Y_i=1\}},$$
$$N_{k0}\,(t) = \sum_{i=1}^{N(t)} C_i\,(t)\,1_{\{A_i=k, Y_i=0\}},\;\; 0 \leq k \leq K, 0 < t \leq T_{\max}. \tag{8}$$

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 allocations, and that the allocation rule is the same as in The case of Bernoulli outcomes section. 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 of the next arriving patient.

By inspection we find that the basic product form of the likelihood expression (3) can be retained in this case. More concretely, the only change needed in the algorithms of *BARTA* and *BARTS* is that, instead of $L_n(\theta) \leftarrow L_{n-1}(\theta) \times \theta_{r(n)}^{Y_{N(n)}} \left(1 - \theta_{r(n)}\right)^{1-Y_{N(n)}}$, the inductive step for updating the likelihood becomes

$$L_n(\theta) \leftarrow L_{n-1}(\theta) \prod_{k=0}^{K} \theta_k^{N_{k,1}(U_{N(n)})-N_{k,1}(U_{N(n)-1})} \\ (1-\theta_k)^{N_{k,0}(U_{N(n)})-N_{k,0}(U_{N(n)-1})}. \tag{9}$$

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 *BARTA* and *BARTS* 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 allocations. 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

$$L\left(\theta_k | D_{k,t}\right) = \prod_{i=1}^{N(t)} f(X_i|\theta_k)^{C_i(t) 1_{\{A_i=k\}}} \left(1 - F(Y_i\,(t)\,|\theta_k)\right)^{(1-C_i(t)) 1_{\{A_i=k\}}}. \tag{10}$$

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

$$L\left(\theta | D_t\right) = \prod_{k=0}^{K} L\left(\theta_k | D_{k,t}\right), \tag{11}$$

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 *BARTA* or *BARTS*.

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 (10) 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

$$p\left(\theta_k | D_{k,t}\right) \\ = Gamma\left(\theta_k \mid \alpha_k + \sum_{i=1}^{N(t)} C_i\,(t)\, 1_{\{A_i=k\}}, \beta_k + \sum_{i=1}^{N(t)} Y_i\,(t)\, 1_{\{A_i=k\}}\right), \tag{12}$$

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

When considering the application of *BARTA* or *BARTS* 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 *BARTS* 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 $\boldsymbol{\eta}_0 + \delta \geq \max_{\ell \in \mathbb{T}} \boldsymbol{\eta}_\ell$.

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 prespecified 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 was followed in spite of that, arguably, in trials for experimental vaccines such as the COVID-19 candidates, for which phase II had 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 in The case of time-to-event data, this task could have been accomplished by applying *BARTS* 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, with 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 $\mathbb{P}_\pi\left(\rho\boldsymbol{\theta}_0 < \boldsymbol{\theta}_1 \middle| D_i^*\right) < \varepsilon_2$ were obtained for some $i \leq N_{max}$. On the other hand, *futility* would be declared if either $\mathbb{P}_\pi\left(\boldsymbol{\theta}_1 \leq \theta_{high} \middle| D_i^*\right) < \varepsilon_1$ or $\mathbb{P}_\pi\left(\rho\boldsymbol{\theta}_0 \geq \boldsymbol{\theta}_1 \middle| D_i^*\right) < \varepsilon_2$ 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 vaccine 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 $N_0(t) = \sum_{m \geq 1} 1_{\{T_{1,m} \leq t\}}$ 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

$$L(\lambda_0, \lambda_1 | D_t) = \prod_{k=0}^{1} \exp\left\{-\int_0^t n_k\lambda_k(s)ds\right\} \prod_{m \leq N_k(t)} n_k\lambda_k(T_{k,m}). \tag{13}$$

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 vaccine group. It is well known that the conditional probabilities 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 vaccine 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. [52]).

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 (13) for this purpose and introducing a separate model for a description of $\lambda_0(.)$, or (ii) following a path well known from the context of the Cox proportional hazards model and employing a corresponding *partial likelihood* expression (e.g., Yip and Chen [53]). 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 (13), 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

$$L_{part}(\rho|D_t) = \left(\frac{n_0}{n_0 + n_1\rho}\right)^{\sum_{m\leq N(t)} \delta(T_m)} \left(\frac{n_1\rho}{n_0 + n_1\rho}\right)^{\sum_{m\leq N(t)} (1-\delta(T_m))}$$
$$= \theta^{N_0(t)}(1-\theta)^{N(t)-N_0(t)}, \tag{14}$$

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 (14) 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 $Beta(\ . \mid \alpha, \beta)$-prior for θ, and using the conjugacy property of the *Beta-Binomial* distribution family, we would get the posterior $p(\theta \mid D_t) = Beta(\theta \mid \alpha + N_0(t), \beta + N(t) - N_0(t))$, 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 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. A similar criterion could be set up for declaring futility.

To give an example from a real study, Moderna, Inc. announced on November 30, 2020 (Moderna Inc. [54]) 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 $p(\rho|D_t)$ 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 Fig. 2. 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

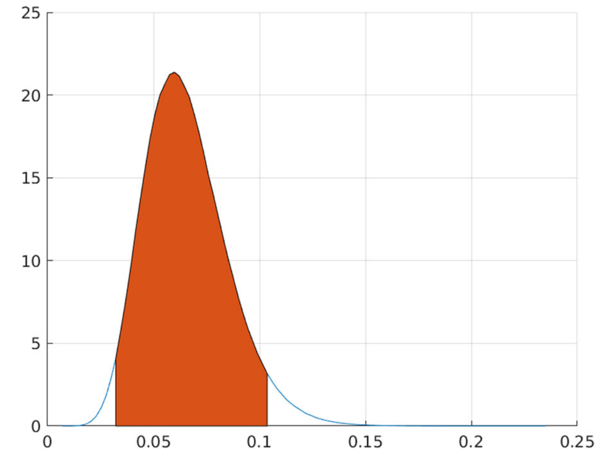

Fig. 2 Posterior density of $\boldsymbol{\rho}$ based on Moderna, Inc. COVID-19 primary efficacy data, with posterior mode at 0.0595 and 95% HPD interval (0.030, 0.105)

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 (14) for partial likelihood is then not valid any more, and needs to be replaced by the product

$$L_{part}(\rho|D_t) = \prod_{T_m \le t} \left(\frac{R_{0,T_m}}{R_{0,T_m} + R_{1,T_m}\rho} \right)^{\delta(T_m)} \left(\frac{R_{1,T_m}\rho}{R_{0,T_m} + R_{1,T_m}\rho} \right)^{1-\delta(T_m)}, \tag{15}$$

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. [55], Vasileiou et al. [56]). 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 may be 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. [57].

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 (WHO [58], Eyal et al. [59], Richards [60]), and it is still considered relevant now (Eyal and Lipsitch [61]). 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 *BARTS*, 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 $1-\varepsilon_1$ of certainty, according to the updated posterior probabilities, 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.

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. [30], 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 symptomatic 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. [62], Wolfson [63]). 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 [64].

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 allocated. 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 the *BARTS* algorithm, 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 its parameters, 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. [14], Yin et al. [65], Hobbs et al. [66]). 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 [67], Greenland et al. [68]). 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, Yin, et al. [69], Stallard et al. [70]). 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 $\mathbb{P}_{\pi}\left(\boldsymbol{\theta}_0+\delta \geq \boldsymbol{\theta}_{\vee} \middle| D_i^*\right) < \varepsilon_2$, 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. [46]) 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 [71]). 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. [72]).

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 The case of Bernoulli outcomes section that such treatment elimination, as encoded into *BARTS*, contains a violation of the likelihood principle.

If *BARTS* 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. [46] and Wathen and Thall [32]. In Thall et al. [46], computer simulations were used to compare adaptive patient allocation based on Thompson's rule (Thompson [11], Villar et al. [30]) 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. [46] was directed more generally towards applying adaptive randomization and would therefore in principle apply to our rules *BARTA* and *BARTS* as well. The results from our simulation experiments, shown in graphical form in Figs. 1, 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 *BARTA* or Thompson's rule, and finally coupling in *BARTS* for actual treatment selection.

Another criticism presented in Thall et al. [46] was that, for trial data collected from a trial applying adaptive randomization, the considered tests had lower power than in the case of equal randomization, provided that the tests were calibrated to have the same Type 1 error rate. This question is discussed in subsections B.1.3 and D of the Supplement. In these experiments, adaptive treatment allocation methods based on *BARTA* designs (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 [73] and Villar et al. [30]. A comprehensive review of the topic is provided in Robertson et al. [74]. 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 are 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 Extensions for handling delayed outcome data. 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. [16], Gsponer et al. [75] and Gerber, Gsponer, et al. [49]. 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 true differences in the treatment effects.

In principle, the logic behind *BARTA* and *BARTS* 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 *BARTA* or *BARTS* 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, *BARTA* and *BARTS* can still be used in interim analyses 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 *BARTA* or *BARTS* 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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01526-8.

Additional file 1: Supplementary Materials.

Acknowledgments

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 during an early stage of this work. Finally, we thank two Reviewers of this paper for their good comments and suggestions for preparing a revision.

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.

Funding

Open access funded by Helsinki University Library.

Availability of data and materials

Only publicly available or simulated data were used. The R package *barts* written by Mikko Marttila generating simulated data sets and implementing the methods is freely available at https://github.com/Orion-Corporation/barts

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

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

Competing interests

The authors declare that they have no competing interests.

Received: 27 September 2021 Accepted: 19 January 2022
Published online: 20 February 2022

References

1. Jennison C, Turnbull B. Group Sequential Tests with Applications to Clinical Trials (Chapman & Hall/CRC Interdisciplinary Statistics). UK: Chapman & Hall; 1999.
2. Chow S-C, Chang M. Adaptive design methods in clinical trials–a review. Orphanet J Rare Dis. 2008;3(1):1–13.
3. Mahajan R, Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect. Indian J Pharmacol. 2010;42(4):201.
4. Chow S-C. Adaptive clinical trial design. Annu Rev Med. 2014;65:405–15.
5. Chang M, Balser J. Adaptive design-recent advancement in clinical trials. J Bioanal Biostat. 2016;1(1):14.

6. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1). https://doi.org/10.1186/s12916-018-1017-7.
7. Atkinson AC, Biswas A. Randomised Response-adaptive Designs in Clinical Trials. Boca Raton: Chapman and Hall/CRC; 2019.
8. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191–9.
9. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35:549–56.
10. Demets DL, Lan KKG. Interim analysis: the alpha spending function approach. Stat Med. 1994;13(13-14):1341–52.
11. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 1933;25(3/4): 285–94.
12. Flühler H, Grieve AP, Mandallaz D, Mau J, Moser HA. Bayesian approach to bioequivalence assessment: an example. J Pharm Sci. 1983;72(10): 1178–81.
13. Berry DA. Interim analyses in clinical trials: classical vs. bayesian approaches. Stat Med. 1985;4(4):521–6.
14. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power?. Control Clin Trials. 1986;7(1):8–17. https://doi.org/10.1016/0197-2456(86)90003-6.
15. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. Am Sci. 1988;76(2):159–65.
16. Spiegelhalter DJ, Freedman LS, Parmar MKB. J R Stat Soc Ser A (Stat Soc). 1994;157(3):357–87.
17. Thall PF, Simon R. Practical bayesian guidelines for phase IIB clinical trials. Biometrics. 1994;50(2):337. https://doi.org/10.2307/2533377.
18. Grieve AP. Idle thoughts of a 'well-calibrated' bayesian in clinical drug development. Pharm Stat. 2016;15(2):96–108.
19. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-care Evaluation, vol. 13. Chichester: John Wiley & Sons; 2004.
20. Berry SM, Carlin BP, Lee JJ, Müller P, Vol. 38. Bayesian Adaptive Methods for Clinical Trials (Chapman & Hall/CRC Biostatistics Series). Boca Raton: CRC Press; 2011, p. 305. With a foreword by David J. Spiegelhalter.
21. Yuan Y, Nguyen HQ, Thall PF. Bayesian Designs for Phase I-II Clinical Trials. Boca Raton: CRC Press; 2017.
22. Berger JO, Wolpert RL, Vol. 6. The Likelihood Principle (Institute of Mathematical Statistics Lecture Notes—Monograph Series). Hayward: Institute of Mathematical Statistics; 1984, p. 206.
23. Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5(1):27–36. https://doi.org/10.1038/nrd1927.
24. Berry DA. Adaptive clinical trials: The promise and the caution. J Clin Oncol. 2011;29(6):606–9. https://doi.org/10.1200/JCO.2010.32.2685.
25. Lee JJ, Chu CT. Bayesian clinical trials in action. Stat Med. 2012;31(25): 2955–72.
26. Yin G, Lam CK, Shi H. Bayesian randomized clinical trials: From fixed to adaptive design. Contemp Clin Trials. 2017;59:77–86.
27. Ruberg SJ, Jr. FEH, Gamalo-Siebers M, LaVange L, Lee JJ, Price K, Peck C. Inference and decision making for 21st-century drug development and approval. Am Stat. 2019;73(sup1):319–27. https://doi.org/10.1080/00031305.2019.1566091.
28. Giovagnoli A. The bayesian design of adaptive clinical trials. Int J Environ Res Publ Health. 2021;18(2). https://doi.org/10.3390/ijerph18020530.
29. Robertson DS, Lee KM, Lopez-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. arXiv preprint arXiv:2005.00564. 2021. http://arxiv.org/abs/2005.00564 Accessed 15 Nov 2021.
30. Villar SS, Bowden J, Wason J. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. Stat Sci. 2015;30(2): 199–215. https://doi.org/10.1214/14-sts504.
31. Trippa L, Lee EQ, Wen PY, Batchelor TT, Cloughesy T, Parmigiani G, Alexander BM. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. J Clin Oncol. 2012;30(26):3258.
32. Wathen JK, Thall PF. A simulation study of outcome adaptive randomization in multi-arm clinical trials. Clin Trials. 2017;14(5):432–40.
33. Ryan EG, Lamb SE, Williamson E, Gates S. Bayesian adaptive designs for multi-arm trials: an orthopaedic case study. Trials. 2020;21(1):1–16.
34. Viele K, Broglio K, McGlothlin A, Saville BR. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. Clin Trials. 2020;17(1):52–60.
35. Viele K, Saville BR, McGlothlin A, Broglio K. Comparison of response adaptive randomization features in multiarm clinical trials with control. Pharm Stat. 2020;19(5):602–12.
36. Bassi A, Berkhof J, de Jong D, van de Ven PM. Bayesian adaptive decision-theoretic designs for multi-arm multi-stage clinical trials. Stat Methods Med Res. 2021;30(3):717–30.
37. Wason JM, Jaki T. Optimal design of multi-arm multi-stage trials. Stat Med. 2012;31(30):4269–79.
38. Wason JM, Trippa L. A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. Stat Med. 2014;33(13): 2206–21.
39. Jacob L, Uvarova M, Boulet S, Begaj I, Chevret S. Evaluation of a multi-arm multi-stage bayesian design for phase II drug selection trials – an example in hemato-oncology. BMC Med Res Methodol. 2016;16(1). https://doi.org/10.1186/s12874-016-0166-7.
40. Yu Z, Ramakrishnan V, Meinzer C. Simulation optimization for bayesian multi-arm multi-stage clinical trial with binary endpoints. J Biopharm Stat. 2019;29(2):306–17.
41. Press WH. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. Proc Natl Acad Sci. 2009;106(52):22387–92. https://doi.org/10.1073/pnas.0912378106.
42. Müller P, Xu Y, Thall PF. Clinical trial design as a decision problem. Appl Stoch Model Bus Ind. 2017;33(3):296–301.
43. Alban A, Chick SE, Forster M. Extending a bayesian decision-theoretic approach to a value-based sequential clinical trial design. 2018 Winter Simul Conf (WSC). 2018:2459–70.
44. Marttila M, Arjas E, Gasbarra D. barts: Bayesian adaptive rules for treatment selection. R package version 0.0.1. 2021. https://github.com/Orion-Corporation/barts Accessed 25 Aug 2021.
45. Zaslavsky BG. Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. Biometrics. 2012;69(1):157–63. https://doi.org/10.1111/j.1541-0420.2012.01806.x.
46. Thall PF, Fox PS, Wathen JK. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann Oncol Off J Eur Soc Med Oncol. 2015;26(8): 1621–8.
47. Thall P, Wathen J. Practical bayesian adaptive randomization in clinical trials. Eur J cancer (Oxford, England : 1990). 2007;43:859–66. https://doi.org/10.1016/j.ejca.2007.01.006.
48. Xie F, Ji Y, Tremmel L. A bayesian adaptive design for multi-dose, randomized, placebo-controlled phase i/ii trials. Contemp Clin Trials. 2012;33(4):739–48.
49. Gerber F, Gsponer T, et al. gsbdesign: an r package for evaluating the operating characteristics of a group sequential bayesian design. J Stat Softw. 2016;69(11):1–27.
50. Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bull NYU Hosp Joint Dis. 2008;66(2):150–4.
51. Richards AD. Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. J Am Stat Assoc. 1994;89(428): 1528–34.
52. Halloran ME, Longini IM, Struchiner CJ, Longini IM. Design and Analysis of Vaccine Studies, vol. 18. New York: Springer; 2010.
53. Yip P, Chen Q. A partial likelihood estimator of vaccine efficacy. Aust New Zealand J Stat. 2000;42:367–74.
54. Moderna Inc. 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. Cambridge, Mass: Moderna Inc.; 2020.
55. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, Hernán MA, Lipsitch M, Reis B, Balicer RD. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. N Engl J Med. 2021;384(15):1412–23.
56. Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, Bedston S, Beggs J, Bradley D, Chuter A, et al. Interim findings from first-dose mass covid-19 vaccination roll-out and covid-19 hospital admissions in scotland: a national prospective cohort study. The Lancet. 2021;397(10285):1646–57.
57. Krause P, Fleming TR, Longini I, Henao-Restrepo AM, Peto R, Dean N, Halloran M, Huang Y, Fleming T, Gilbert P, et al. Covid-19 vaccine trials should seek worthwhile efficacy. The Lancet. 2020;396(10253):741–3.

58. World Health Organization. Key criteria for the Ethical acceptability of Covid-19 human challenge studies: World Health Organization; 2020. https://www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_criteria-2020.1 Accessed 28 May 2021.
59. Eyal N, Lipsitch M, Smith PG. Human challenge studies to accelerate coronavirus vaccine licensure. J Infect Dis. 2020;221(11):1752–6.
60. Richards AD. Ethical guidelines for deliberately infecting volunteers with covid-19. J Med Ethics. 2020;46(8):502–4. https://doi.org/10.1136/medethics-2020-106322.
61. Eyal N, Lipsitch M. How to test sars-cov-2 vaccines ethically even after one is available. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021;73:2332–34.
62. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. Pediatrics. 1985;76(4):479–87.
63. Wolfson PJ. The development and use of extracorporeal membrane oxygenation in neonates. Ann Thorac Surg. 2003;76(6):2224–9.
64. Royall RM. Ethics and statistics in randomized clinical trials. Stat Sci. 1991;6:52–62.
65. Yin G, Chen N, Jack Lee J. Phase ii trial design with bayesian adaptive randomization and predictive probability. J R Stat Soc Ser C (Appl Stat). 2012;61(2):219–35.
66. Hobbs BP, Chen N, Lee JJ. Controlled multi-arm platform design using predictive probability. Stat Methods Med Res. 2018;27(1):65–78.
67. Wasserstein RL, Lazar NA. The asa statement on p-values: Context, process, and purpose. Am Stat. 2016;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108.
68. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
69. Shi H, Yin G, et al. Control of type i error rates in bayesian sequential designs. Bayesian Anal. 2019;14(2):399–425.
70. Stallard N, Todd S, Ryan EG, Gates S. Comparison of bayesian and frequentist group-sequential clinical trial designs. BMC Med Res Methodol. 2020;20(1):1–14.
71. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. Paediatr Perinat Epidemiol. 2020;35(1):8–23.
72. Grieve AP, Pocock SJ, ABRAMS K, Ashby D, Healy M, Jennison C, Lewis J, Lindley D, Machin D, Newman G, et al. J R Stat Soc Ser A (Stat Soc). 1994;157(3):387–416.
73. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994;50(4):1029–41.
74. Robertson DS, Choodari-Oskooei B, Dimairo M, Flight L, Pallmann P, Jaki T. Point estimation for adaptive trial designs. 2021. http://arxiv.org/abs/2105.08836.
75. Gsponer T, Gerber F, Bornkamp B, Ohlssen D, Vandemeulebroecke M, Schmidli H. A practical guide to bayesian group sequential designs. Pharm Stat. 2014;13(1):71–80.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.