Bandit Procedures for Designing Patient-Centric Clinical Trials

Sofia S. Villar and Peter Jacko

Abstract Multi-armed bandit problems (MABPs) are a special type of optimal control problem that has been studied in the fields of operations research, statistics, machine learning, economics and others. It is a framework well suited to model resource allocation under uncertainty in a wide variety of contexts. Across the existing theoretical literature, the use of bandit models to optimally design clinical trials is one of the most typical motivating application, where the word "optimally" refers to designing the so-called patient-centric trials, which would take into account the benefit of the in-trial patients and thus are by some researchers considered more ethical. Nevertheless, the resulting theory has had little influence on the actual design of clinical trials. Contrary to similar learning problems arising for instance in digital marketing where interventions can be tested on millions of users at negligible cost, clinical trials are about "small data", as recruiting patients is remarkably expensive and (in many cases) ethically challenging. In this book chapter, we review a variety of operations research and machine learning approaches that lead to algorithms to "solve" the finite-horizon MABP and then interpret them in the context of designing clinical trials. Due to the focus on small sizes, we do not resort to the use of the normal distribution to approximate a binomial distribution which is a common practice for large samples either "for simplicity" or "for ease of computation". Solving a MABP essentially means to derive a response-adaptive procedure for allocating patients to arms in a finite sample experiment with no early stopping. We evaluate and compare the performance of these procedures, including the traditional and still dominant clinical trial design choice: equal fixed randomization. Our results illustrate how bandit approaches could offer significant advantages, mainly in terms of allocating more patients to better interventions, but still pose important inferential challenges,

Sofia S. Villar
MRC Biostatistics Unit, University of Cambridge, UK. e-mail: sofia.villar@mrc-bsu.cam.ac.uk

Peter Jacko
Department of Management Science, Lancaster University, UK and Berry Consultants, UK. e-mail: p.jacko@lancaster.ac.uk

particularly in terms of their resulting lower statistical power, potential for bias in estimation and existence of closed-form tests distributions or asymptotic theory. We illustrate some promising modifications to bandit procedures to address power and bias issues and we reflect upon the open challenges that remain for an increased uptake of bandit models in clinical trials.

1 Introduction

Multi-armed bandit problems (MABPs) define a special class of an optimal control problem. The MABPs is a well studied and a well suited framework to model resource allocation under uncertainty in a wide variety of contexts. As Whittle (1980) put it: *The multi-armed bandit problem (as it has become known) is important as one of the simplest non-trivial problems in which one must face the conflict between taking actions which yield immediate reward and taking actions (such as acquiring information, or preparing the ground) whose benefit will come only later. It has proved difficult enough to become a classic, and has now a large literature.*

The MABP has developed over its history as a key example of a problem that has attracted considerable attention both from the Operations Research (OR) and Machine Learning (ML) literature, thus having an exceptional potential to act as a bridge between these two communities. As well, the MABP had its origins in the medical statistics literature, when Thompson (1933) published his work back in the 1930s, and one can easily argue today that its potential to improve health applications is high (Villar et al., 2015; Press, 2009).

However, despite the great theoretical attention from both OR and ML literature and the considerable potential of its application in practice, both the links between OR and ML as well as the uptake in practice remain relatively under-explored. The potential of the MABP to act as bridge between the OR and ML communities remains low because the perspective in tackling the problem has had a markedly different focus in the two fields. While in OR, formulations based on optimizing expected discounted (or average) rewards are the most common, in ML the dominant goal is that suggested by Robbins where the average regret is minimized. In both cases, the most common horizon considered is infinite and the focus usually is on asymptotic forms of optimality. Second, the uptake of so-called bandit methods in health care practice, and specially in clinical trials, is still virtually non-existing. This may very well be surprising to the reader as across all of this theoretical literature, the use of bandit models to optimally design clinical trials is posited as the typical motivating application. Yet, as it was explored in Villar et al. (2015) and we will further discuss in this chapter, little of the resulting theory has ever been used in the actual design and analysis of clinical trials. The focus on infinite horizon problems for OR and ML is one of the reasons for lack of practical impact but (as we will discuss later) not the only one.

At this point the reader may also wonder why could a MABP be a perfect fit to optimize the design of clinical trials. The development of a drug or medical

therapy follows a regulated and lengthy process which may take between 10 and 15 years (from discovery to being available for patients). Drugs are tested in humans only after laboratory testing and it is divided into a series of successive clinical trials traditionally known as phase I, II, III, and IV trials. These phases are usually separate clinical studies, and each has a unique objective. Typically, phase I trials establish safety and tolerability in healthy volunteers; phase II trials study the drugs' efficacy and adverse effects at different dosages in patients; phase III trials establish the effectiveness and safety of the drug compared with placebo or standard of care; and phase IV trials determine general risks and benefits after approval.

A clinical trial is an experiment designed to produce data in order to answer a specific question about a medical intervention (e.g., a drug's superiority versus a standard of care). A typical Phase III clinical trial would compare a single new intervention to a standard of care (which could be simply placebo) with the aim of establishing superiority (or non-inferiority) in terms of a certain efficacy metric. Many Phase II trials compare multiple variants of the same intervention (e.g., drug dosages, treatment durations, or treatment combinations), while some recent Phase II trials include and compare multiple (independent) interventions in one trial. Currently, there is a growing number of trials which might not be easily categorized into these four phases, and even the regulators seem to tend to move away from such strict definitions, and instead, talk about *exploratory* and *confirmatory* trials. Some trials might even answer several questions and/or run across various phases, such as the so-called *seamless phase II/III trials* or *platform trials*.

Bandit problems formalize the tension between two goals when collecting data to aid decision making under uncertainty. Those goals are, the desire to *learn* (or *explore*) about the different alternatives (i.e. to learn about the new interventions) and that to *earn* (or, *exploit*) from that learning to achieve a certain overall objective (i.e. to treat more patients effectively). Therefore, one could argue that in a confirmatory clinical trial there could be an aim of balancing two separate goals: (i) to correctly identify the best intervention (learning) and (ii) to treat patients as effectively as possible during the trial (earning). These two goals may appear to some as naturally complementary, but for those familiar with the MAPB it should be clear that this is not the case. If one is considering the case of a finite population of patients, then correctly identifying the best intervention requires some patients to be allocated to all interventions, and therefore the former acts to limit the possibility of treating more patients with a superior intervention.

As we will describe in this chapter, designing a clinical trial using a MABP solution will entail defining a so-called *response-adaptive* allocation procedure, which (together with specification of other aspects, e.g. statistical analysis methods to be used at the end of the trial) would be part of an adaptive design of a clinical trial. Traditional clinical trials, which have been the dominant design paradigm until the very last decade, follow a linear schematic: design, conduct and analysis of data according to a pre-specified plan. This approach allows for no form of change to the experiment based on the accumulated data. In contrast to this, adaptive designs permit pre-planned changes (or adaptations) to occur after interim looks of the trials data. The key element is that while one can be flexible and adapt based on the

observed data this should be done without undermining its integrity or validity. This latter part and the difficulties it poses for new designs will play a key role explaining the lower uptake of bandit results in practice. The interested reader may read Pallmann et al. (2018); Burnett et al. (2020) for a non-technical introduction to adaptive designs.

While adaptive designs broadly defined have generated a lot of interest in the clinical trials community recently, particularly after the COVID-19 crisis (Stallard et al., 2020), bandit models, methods and algorithms as a class of procedures potentially very useful to deliver adaptive designs for patient-centric trials remain largely unused in practice. Recent work has discussed the reasons for this lower uptake in detail (Villar et al., 2015), discussing what the potential benefits of their use can be as well as the challenges to its application in clinical trial practice. In this chapter, we revisit the ideas presented in the work above and build on them to explain what has changed since and what still calls for further research.

The structure of this chapter is as follows. In the following section we introduce terminology, assumptions and notation. In this chapter we shall follow the convention (for simplicity of presentation only) that two-arm clinical trials represent typical Phase III (confirmatory) trials while multi-armed trials reflect Phase II (exploratory) trials. This is an oversimplification as one could imagine two-armed trials that are exploratory or multi-armed ones that are confirmatory but it would aid presentation of statistical and design concepts that are relevant in one case more than in another.

2 The Bayesian Beta-Bernoulli MABP

In this section we present a Bayesian formulation of a finite-horizon multi-armed problem with binary outcomes as a collection of Markov decision processes (MDPs), which provides a framework for finding the Bayes-optimal allocation procedure by dynamic programming. Our problem of interest has the following defining elements: time, arms (interventions), and each arm is modelled as an MDP with states (information), actions (allocations), transition probabilities and expected one-period rewards (patient outcomes).

Time.

Patients arrive (i.e., are recruited) sequentially (i.e., one by one) at random moments in continuous time. Since we do not discount the future, we can without loss of generality focus only on the moments of patients' arrivals, which we call discrete time epochs and see as regularly spaced. That is, equivalently, we can consider that patients arrive at time epochs $t \in \mathcal{T} := \{0, 1, 2, \ldots, T-1\}$, where $T < +\infty$ is the number of patients in the trial, i.e., the trial size. To clarify, the $(t+1)$-st patient arrives at time epoch t. Note that $t = T$ is the time epoch denoting the end of the trial, when the outcome of the last patient is observed and no patient arrives.

Arms.

We consider arms (or, interventions) labelled by $k \in \mathcal{K} := \{0, 1, \ldots, K\}$, where arm $k = 0$ refers to a *control intervention* (typically, a standard of care for the studied disease), and arms $k = 1, \ldots, K$ refer to novel (experimental) interventions. A patient must be allocated to exactly one intervention (although this intervention may well be defined as a combination therapy), and such allocation results in a binary type of outcome from that intervention: 0 (failure) or 1 (success). The outcome set is denoted by $\mathcal{O} = \{0, 1\}$. In a clinical trial context, the success outcome represents, e.g., response to intervention, remission of tumor, etc. Patient outcomes are uncertain, i.e., modelled as Bernoulli-distributed with parameter p_k (the success probability), independent across arms. Taking the Bayesian approach, the initial prior for the success probability of arm k is Beta distribution with parameters $(\widetilde{s}_k(0), \widetilde{f}_k(0))$, which can be interpreted as the number of pseudo-successes and pseudo-failures observed before making the first allocation in the experiment. The rewards are immediate, meaning that the outcome of an allocated patient is observed before the next decision needs to be made.

States.

The *state space* for arm k, $\mathcal{X}_k := \{\boldsymbol{x}_k := (s_k, f_k) \in (\mathcal{T} \cup \{T\})^2 : s_k + f_k \leq T\}$, represents all the possible two-dimensional vectors of available information on the unknown parameter p_k at any time during the trial. Note that we exclude the prior information (i.e., pseudo-observations) from the state definition because it does not change over time and because in this way the model is as small as possible, which is beneficial from the computational point of view. However, to simplify some expressions, we also define the pseudo-state $\widetilde{\boldsymbol{x}}_k := (\widetilde{s}_k, \widetilde{f}_k)$ with $\widetilde{s}_k := \widetilde{s}_k(0) + s_k$, $\widetilde{f}_k := \widetilde{f}_k(0) + f_k$.

Actions.

The *action set* $\mathcal{A}_k$ for arm k is a binary set representing the action of drawing a sample observation from arm k ($a_k = 1$) or not ($a_k = 0$). In a clinical context, the action variable stands for the choice of allocating next patient to arm k or not.

Transition Probabilities.

The Markovian *transition law* $\mathcal{P}_k(\boldsymbol{x}'_k | \boldsymbol{x}_k, a_k)$ describing the evolution of the information state variable on arm k in state $\boldsymbol{x}_k$ under action a_k from one time epoch to the next, is given by:

$$\boldsymbol{x}'_k = \begin{cases} (s_k+1, f_k)\,, & \text{if } a_k = 1 \text{ w.p. } \frac{\widetilde{s}_k}{\widetilde{s}_k+\widetilde{f}_k}, \\ (s_k, f_k+1)\,, & \text{if } a_k = 1 \text{ w.p. } \frac{\widetilde{f}_k}{\widetilde{s}_k+\widetilde{f}_k}, \\ \boldsymbol{x}_k, & \text{if } a_k = 0 \text{ w.p. } 1, \end{cases} \tag{1}$$

where w.p. stands for 'with probability'. Note that under action 1, the transition probabilities are defined by the mean of the current posterior distribution, which, due to conjugacy, is a Beta distribution with parameters $\left(\widetilde{s}_k, \widetilde{f}_k\right)$.

Expected One-period Reward.

The expected reward on arm k in state $\boldsymbol{x}_k$ under action a_k is:

$$\mathcal{R}^{a_k}_{k,\boldsymbol{x}_k} = \frac{\widetilde{s}_k}{\widetilde{s}_k+\widetilde{f}_k} a_k, \tag{2}$$

where in accordance to the above specified dynamics, expected reward is the Bayes-expected number of successes from the current patient, computed using the current posterior Beta distribution.

Note that both the transition law and the rewards depend on the prior distributions, although we do not indicate it in the notation. The system dynamics is captured by the joint state process $(\boldsymbol{x}_k(t))_{k\in\mathcal{K}}$ for all $t \in \mathcal{T}\cup\{T\}$ and by the joint action process $(a_k(t))_{k\in\mathcal{K}}$ for all $t \in \mathcal{T}$. The actions are restricted by the fact that every patient in the trial must be allocated to one and only one arm, i.e., $\sum_{k\in\mathcal{K}} a_k(t) = 1$ for all $t \in \mathcal{T}$. This restriction implies a restriction on the joint state process so that $\sum_{k\in\mathcal{K}}(s_k(t)+f_k(t)) = t$ for all $t \in \mathcal{T}\cup\{T\}$.

A rule is required to operate the resulting (sometimes called *weakly-coupled*) MDP, which indicates which action to take for each arm $k \in \mathcal{K}$ for every possible combination of states of the arms at every time $t \in \mathcal{T}$. Such a rule forms a sequence of actions resulting in a joint action process $(a_k(t))_{k\in\mathcal{K}}$ and it is known as a *policy*, denoted by $\pi \in \Pi$, where Π is the set of all the feasible policies satisfying the above action constraint.

To complete the specification of the multi-armed bandit model as an *optimal control model*, the problem's *objective function* must be selected. The typical performance objective in the Bayesian Beta-Bernoulli MABP in a trial with T patients is to maximize the *Bayes-expected number of successes*. For a feasible policy $\pi \in \Pi$, the Bayes-expected number of successes is, i.e., the total value function conditional on the initial joint prior parameters $\widetilde{\boldsymbol{x}}(0)$,

$$\mathrm{ENS}^{\pi}_{\widetilde{\boldsymbol{x}}(0)} = E^{\pi}_{\widetilde{\boldsymbol{x}}(0)}\left[\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}} \mathcal{R}^{a_k(t)}_{k,\boldsymbol{x}_k(t)}\right] = E^{\pi}_{\widetilde{\boldsymbol{x}}(0)}\left[\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}} \frac{\widetilde{s}_k(t)}{\widetilde{s}_k(t)+\widetilde{f}_k(t)} a_k(t)\right], \tag{3}$$

where $E^{\pi}_{\widetilde{\boldsymbol{x}}(0)}[\cdot]$ denotes Bayesian expectation with joint Beta prior parameters $\widetilde{\boldsymbol{x}}(0) := (\widetilde{\boldsymbol{x}}_k(0))_{k\in\mathcal{K}}$ under policy π. The multi-armed bandit optimal control problem is

mathematically summarized as the problem of finding an optimal policy π^*, i.e., a feasible policy ($\pi^* \in \Pi$) that optimizes the performance objective. Formally, the optimal policy is

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} \, \mathrm{ENS}^{\pi}_{\widetilde{\boldsymbol{x}}(0)}, \tag{4}$$

and the optimal Bayes-expected number of successes is

$$\mathrm{ENS}^{*}_{\widetilde{\boldsymbol{x}}(0)} = \max_{\pi \in \Pi} \mathrm{ENS}^{\pi}_{\widetilde{\boldsymbol{x}}(0)}. \tag{5}$$

Note that the right-hand side of (4) suggests that π^* should depend on the prior $\widetilde{\boldsymbol{x}}(0)$, but the MDP theory implies that there is an optimal policy which is stationary (i.e., it prescribes joint action $(a_k(t))_{k \in \mathcal{K}}$ only as a function of the posterior joint state $\widetilde{\boldsymbol{x}}(t) := \left(\widetilde{s}_k(t), \widetilde{f}_k(t)\right)_{k \in \mathcal{K}}$ without a direct dependence on t), and thus we assume π^* is such and drop its dependence on the prior parameters.

The optimal policy π^* is, nevertheless, in general different for different trial sizes T, because larger T tends, for a given state, to lead to an allocation that provides a larger amount of learning about the arms' unknown success probability parameters in order to increase the expected number of successes from the remaining patients.

2.1 Discussion of the Model

The above model is probably the simplest model for the multi-armed bandit problem cast as an optimization problem. Analogous modelling approach can in theory be employed for other distribution of outcomes (discrete, continuous, etc.), although the state would need to be redefined as an appropriate sufficient statistic, and the transition law and reward would need to be adjusted correspondingly (see, e.g., Williamson and Villar, 2020). However, in practice, these often quickly become computationally unfeasible to be solved by dynamic programming, and approximate approaches need to be employed.

The action set can be generalized by making the actions randomized and/or by specifying an action to take when the original two actions are equivalent. In some states, one could modify the action set to either have a single action (for instance, allowing only allocation to a pre-specified arm or allowing only equal fixed randomization in the initial stage of the trial), or have more actions (for instance, allowing for stopping of the trial if the treatment difference seems to be large).

The model can be extended to include discounting of the future patients' outcomes and/or to be optimized over an infinite horizon using standard approaches from the theory of Markov decision processes, but we believe that the undiscounted finite-horizon formulation is the most relevant for healthcare applications.

The rewards can be generalized, for instance, by including penalties in some undesirable states in order to improve a particular statistical operating characteristic, such as in those that would lead to an extremely unbalanced allocation in order to

improve power and estimation (as in Williamson et al., 2017, 2021) or by using any other utility function of interest.

The above-defined model thus requires only the horizon T and the prior parameters to be set by the trial designer. The standard choice in the bandit literature is to set the horizon equal to the size of the trial, but in clinical trials it may sometimes be more reasonable to optimize over the size of the patient population, assuming that one of the arms is chosen at the end of the trial and is applied to the after-trial patients. The standard choice for the prior parameters is the so-called Bayes' prior $\left(\widetilde{s}_k(0), \widetilde{f}_k(0)\right) = (1,1)$, which is considered non-informative, although other priors with mean 0.5 are also considered uninformative, e.g. Jeffrey's prior $(0.5, 0.5)$ or Haldane's prior $(0,0)$. Note that Haldane's prior essentially reduces the optimization problem to a frequentist objective, where the posterior mean equals the sample mean, which is the maximum likelihood estimator of the mean, as shown in Bowden and Trippa (2017).

3 Metrics for Two-armed Problem (Confirmatory Trials)

The two-armed bandit problem with binary outcomes is probably the most studied version of all the bandit problems, intriguing researchers from several disciplines for almost a century (for a review, see, e.g., Jacko, 2019b). At the same time, clinical trials with two arms is probably the most common setup of clinical trials in practice, especially used for *confirmatory trials* which are typically defined with an objective of generating convincing evidence of efficacy (and safety) in order to seek regulatory approval. These are traditionally referred to as the *randomized controlled trials*, where "controlled" indicates that a novel intervention is being concurrently compared to another one (typically, the current standard of care), i.e., there are at least two arms, in order to control for seasonality effects, time trends, population changes and other shocks, and "randomized" indicates that patients are allocated to interventions using a procedure which ensures that patients and their doctors are not able to predict with certainty which intervention will be allocated next, in order to help avoiding the selection bias and other types of biases (see, e.g. Rosenberger and Lachin, 2015, for a discussion of importance of randomization). Throughout this section, we assume $K = 1$, having a control arm $k = 0$ and an experimental arm $k = 1$.

Traditionally, the *randomization ratio* is taken as 1:1, called the *equal fixed randomization* (EFR). This is done without any rigorous justification, often relying on a widespread myth that the 1:1 ratio maximizes statistical power, which is however true only under the assumption of equal variances of the efficacy of the two arms. That might be a somewhat reasonable assumption in some cases of continuous outcomes modelled using the normal distribution, but it is not appropriate for binary outcomes as the variance of the Bernoulli distribution is dependent on its mean (Robertson et al., 2021) and also for other types of outcomes such as time-to-event (Sverdlov et al., 2011). Clinical trials might also be too small to invoke the recommended conditions

for approximation of binomial samples by the normal distribution. Understanding of that and consideration of patient outcome (e.g. for deadly diseases that have no current treatment) leads some clinical trial designers to implement other fixed ratios in an ad hoc manner, e.g. 2:1, typically allocating higher probability to the novel intervention. Note that the 1:1 randomization ratio is often interpreted in the academic literature as that every patient's allocation is randomized with probability of 0.5 to either arm, but the ratio is in practice implemented essentially as a permutation of allocations within blocks of patients, e.g. in every block of 60 patients there are 30 patients allocated to each arm, i.e., in practice it is a *per-block allocation ratio* rather than a *per-patient randomization probability*.

Several stakeholders are involved in confirmatory trials and thus several metrics are of interest: the regulatory agencies would constrain the *Type I error* (typically at the one-sided level of 0.025), intervention sponsors would require high *statistical power* (typically at the level of 0.8 or 0.9) and small *trial size* (or, more generally, a good balance of expected trial costs and expected post-approval revenues), patient organizations would require high *patient benefit* (i.e., health benefit for in-trial patients), and health economics agencies and clinicians would require *accurate and precise estimation* of the interventions efficacy (or of their difference).

3.1 Accurate and Precise Estimation

Unequal fixed (i.e., not adaptive to observed successes and failures on each arm) randomization is well understood in the biostatistics literature, but the researchers in other disciplines and practitioners seem to be largely unaware of those results. For the two-armed setting there are closed formulae that give ratios that are optimal for different objectives. Any fixed procedure that allocates at least one patient to each intervention provides basis for an unbiased estimation of the efficacy of each arm using the *maximum likelihood estimator* (MLE) which equals to the mean of observed successes, and for statistical testing (Rosenberger et al., 2019).

While perfect accuracy (i.e., unbiased estimation) can be achieved by fixed randomization, using the MLE after an adaptive procedure always leads to a bias (Bowden and Trippa, 2017), which is typically negative, but can also be positive (Nie et al., 2018). In that case, improving accuracy by using other estimators that are unbiased can be done at a cost of decreased precision (e.g., the mean-squared error). To the best of our knowledge, maximization of the precision using adaptive procedures is not well understood, but there are some promising recent research lines (Hadad et al., 2021), although existing estimation methods typically do not apply to deterministic allocation procedures. Note also that it is linked to the maximization of statistical power. In practice, the block-randomization is sometimes implemented in a stratified way and/or using the so-called *minimization algorithms* that balance the covariates in order to increase the precision of estimators.

3.2 Statistical Errors

Statistical hypothesis testing is usually required by the regulators at the end of a confirmatory trial in order to apply for a marketing approval. This is usually done in a frequentist approach (but Bayesian approaches are also sometimes allowed after a discussion with the regulator). For one-sided test comparing two proportions, we specify the null hypothesis and the alternative hypothesis as follows:

$$H_0 : p_1 \leq p_0 \tag{6}$$

$$H_1 : p_1 > p_0 \tag{7}$$

One-sided testing is more appropriate than two-sided testing whenever the regulator is interested in limiting the probability of approving the novel intervention (arm 1) despite being worse or equal than the control intervention (arm 0), which is called the *Type I error*, formally defined as the probability of rejecting the null hypothesis if it is true. On the other hand, the sponsor of the novel intervention is interested in achieving a high probability of getting the novel intervention approved if it is indeed better than the control intervention, which is called the *statistical power*, formally defined as the probability of rejecting the null hypothesis if it is false.

A variety of tests have been proposed for a comparison of proportions of two binomial distributions, including z-tests (unpooled or pooled; with or without continuity correction), Fisher's exact test (and its modifications such as Boschloo's test), or simulation-based randomization tests. However, there is no consensus on which test is the most appropriate, because they all have certain disadvantages. The z-tests are based on approximation of binomial distribution by normal distribution, and therefore are suitable for large samples; typically it is suggested that there should be a minimum number of both successes and failures on each arm (5 or 10). The Fisher's exact test is considered too conservative, yielding the Type I error sometimes notably below the given significance level. Other tests, including randomization tests become computationally intractable for large samples.

For a given Type I error, the ratio that maximizes the statistical power if using the (unpooled) z-test is the Neyman's allocation ratio $\sqrt{\theta_C(1-\theta_C)} : \sqrt{\theta_D(1-\theta_D)}$ (Melfi and Page, 1998), which is the ratio of standard deviations of Bernoulli distributions with means θ_C and θ_D (we remark a connection with optimal designs of ranking and selection problems presented in Ryzhov (2021, equation (4))). We can see that the Neyman's allocation coincides with 1:1 when the efficacies of the two interventions are either equal (i.e., $\theta_C = \theta_D$) or equally distant from 0.5 (i.e., $\theta_C = 1 - \theta_D$). The monotonicity properties of the standard deviation formula imply that the intervention whose efficacy is closer to 0.5 is allocated more patients. So, the inferior intervention, which might be considered undesirable from the patient-benefit perspective, is allocated more patients if and only if $\theta_C > 1 - \theta_D$. For instance, if $\theta_C = 0.5$ and $\theta_D = 0.2$ (or 0.8), the ratio that maximizes the statistical power is 5:4, while $\theta_D = 0.1$ (or 0.9) gives the ratio 5:3; ratio 2:1 is optimal for instance if $\theta_C = 0.5$ and $\theta_D \approx 0.067$ (or 0.933) or if $\theta_C = 0.2$ (or 0.8) and $\theta_D \approx 0.042$ (or 0.958). However, as the Neyman's allocation ratio depends on the efficacies of the

two arms, which are unknown, it needs to be implemented adaptively in a "learning by doing" fashion, typically by adaptively estimating the efficacies using the accumulating observations (Rosenberger et al., 2001).

3.3 Patient Outcome

In order to measure the benefit for patients in the trial, we define the *expected number of successes* under procedure $\boldsymbol{\pi}$ if the probabilities of success are $\boldsymbol{p}$,

$$\mathrm{ENS}_{\boldsymbol{p}}^{\boldsymbol{\pi}} = E_{\boldsymbol{p}}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1}\sum_{k=0}^{K} p_k a_k(t)\right], \tag{8}$$

where $E_{\boldsymbol{p}}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under procedure $\boldsymbol{\pi} \in \Pi$ prescribing the vector $\boldsymbol{a}(t)$ of allocation processes $a_k(t) \in \{0, 1\}$ if the probabilities of success are $\boldsymbol{p}$. (Note the slight abuse of notation, with (3) being a Bayesian expectation depending on the prior parameters, while (8) being a frequentist expectation depending on the true success probabilities.) An alternative measure of patient benefit is the *expected proportion of allocations on the superior arm* under procedure $\boldsymbol{\pi}$ if the probabilities of success are $\boldsymbol{p}$,

$$\mathrm{EPASA}_{\boldsymbol{p}}^{\boldsymbol{\pi}} = \frac{1}{T} E_{\boldsymbol{p}}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} a_{k^*}(t)\right], \tag{9}$$

where $k^* := \min \operatorname{argmax}_{k\in\{0,\ldots,K\}} p_k$ is the lexicographically first of all the superior arms in the trial. The means of EPASA and ENS are linear transformations (so, produce an equivalent performance ordering of procedures) in the case of two arms, but their variability is not so easily linked (and they are not equivalent in the case of more than two arms because EPASA does not capture how the allocations are split among the non-superior arms).

Kelly (1981) derived an allocation procedure which is optimal to be used at the beginning of a trial (assuming an infinite trial size) with the objective of providing the maximum Bayes-expected patient benefit. It is known as the *least failures first* (LFF) rule, and it sequentially allocates patients to the intervention with fewer observed failures, breaking the ties in favour of the intervention with more observed successes (breaking the double ties arbitrarily). It is easy to see that this procedure continues allocating patients to the same intervention as long as observing successes and it switches to the other intervention after the first or after the second observed failure since the last switch. See also (Jacko, 2019b) for a discussion of its similarity to the "stay-with-a-winner&switch-on-a-loser" rule known in the biostatistics literature as the "Play-the-winner" rule (Zelen, 1969). This procedure in the long term converges to the ratio $(1 - \theta_D) : (1 - \theta_C)$, which is the same asymptotic ratio as of the "Play-the-winner" rule (Zelen, 1969) and of a specific configuration of the "Randomized play-the-winner" with its parameter $\alpha = 0$ (Wei and Durham, 1978; Rosenberger, 1999). For instance, if $\theta_C = 0.5$ and $\theta_D = 0.2$ (or 0.8), the ratio is 8:5 (or 2:5),

while $\theta_D = 0.1$ (or 0.9) gives the ratio 9:5 (or 1:5); ratio 2:1 is optimal for instance if $\theta_C = 0.5$ and $\theta_D = 0.0$ or if $\theta_C = 0.8$ and $\theta_D = 0.6$.

For a finite trial size, the maximum Bayes-expected patient benefit can be obtained only computationally, using dynamic programming (DP) methods such as the exact (optimal) method of backward recursion or approximate (near-optimal) methods such as the Whittle index rule and the Gittins index rule. All these methods result in allocation procedures which are not only adaptive (to observed successes and failures on each arm) but also non-myopic meaning that they depend on the trial size T. The backward recursion and the Whittle index rule have this dependence direct by defining the (remaining) time horizon of the optimization problem at every moment by the (remaining) trial size. The Gittins index rule has this dependence only indirectly by choosing the discount factor which should reflect the trial size. Jacko (2019b); Pilarski et al. (2021) illustrated that efficient coding in performance-oriented programming languages (such as Julia and C++) allows for using these computational methods for offline calculation of the allocation procedures (stored in lookup tables) for trials sizes of up to several thousand on standard computers. The backward recursion method is only practical when the number of arms is small, but the sub-optimality of some index rules is practically negligible (see Section 4).

Other allocation ratios that are patient-benefit optimal given a constraint on the variance of a function comparing the two interventions were developed in (Rosenberger et al., 2001).

3.4 Trial Size

While all the above approaches try to optimize a metric for a given trial size, a very common approach in practice is actually to minimize the trial size given (some of) the above metrics as constraints. This is because shorter trials are cheaper (recent studies report a cost of over $\$100,000$ per in-trial patient for some diseases) and, if approved, lead to a longer patent-protected marketing period.

3.5 Multiple Metrics

Besides the single-metric optimization, typically subject to a single constraint, researchers have developed procedures that come close to optimizing several metrics. These are usually tunable procedures, in which some parameters can be set to (directly or indirectly) give higher or lower weight to a particular metric. We will discuss two such families of procedures: the tunable Upper Confidence Bound (αUCB) procedures and the Constrained Randomized Dynamic Programming (CRDP) procedures.

Following Bubeck and Cesa-Bianchi (2012, Section 2), we consider the popular αUCB procedure which allocates each arm once in the initial two periods, and then deterministically allocates every patient to the arm with currently the largest index

	Crit.	$H_0 : p_0 = p_1 = 0.3$			$H_1 : p_0 = 0.3\ ,\ p_1 = 0.5$		
	Value	Type I E	EPASA (SD)	ENS (SD)	Power	EPASA (SD)	ENS (SD)
EFR	1.645	0.052	0.500 (0.04)	44.34 (5.62)	0.809	0.501 (0.04)	59.17 (6.03)
TS	1.645	0.066	0.499 (0.10)	44.39 (5.58)	0.795	0.685 (0.09)	64.85 (6.62)
2UCB	1.645	0.062	0.499 (0.10)	44.30 (5.60)	0.799	0.721 (0.07)	66.03 (6.57)
RBI	1.645	0.067	0.502 (0.14)	44.40 (5.57)	0.763	0.737 (0.07)	66.43 (6.54)
RGI	1.645	0.063	0.500 (0.11)	44.40 (5.61)	0.785	0.705 (0.07)	65.46 (6.40)
CB	F_a	0.046	0.528 (0.44)	44.34 (5.55)	0.228	0.782 (0.35)	67.75 (12.0)
WI	F_a	0.048	0.499 (0.35)	44.37 (5.59)	0.282	0.878 (0.18)	70.73 (8.16)
GI	F_a	0.053	0.501 (0.26)	44.41 (5.58)	0.364	0.862 (0.11)	70.21 (7.11)

Table 1 Comparison of procedures in a two-arm trial of size $T = 148$ by simulation. 1.645: the critical value used in z-test (two-sided; confidence level approximately 0.9); F_a: Fisher's adjusted test (two-sided). TS: Thompson sampling; RBI: Randomized belief index; RGI: Randomized Gittins index; CB: Current belief; WI: Whittle index; GI: Gittins index (with discount factor 0.99. Re-printed (adapted) from Villar et al. (2015, Table 5).

(breaking ties randomly) of the form

$$\frac{s_k(t)}{s_k(t) + f_k(t)} + \sqrt{\frac{\alpha \cdot \ln(t+1)}{s_k(t) + f_k(t)}} \tag{10}$$

where $\alpha \geq 0$. The original procedure introduced in Auer et al. (2002) used $\alpha = 2$. Theoretical upper bounds currently exist for $\alpha > 1$, but researchers have noticed empirically that lower values of α typically lead to better performance and some used $\alpha = 1$, see, e.g., Cserna et al. (2017). Numerical experiments of finite trials have revealed that approximately the best patient benefit is robustly achieved with $\alpha = 0.18$ (Jacko, 2019b) or $\alpha = 0.19$ (Pilarski et al., 2021). Note that setting $\alpha = 0$ recovers the (frequentist) myopic procedure which at every period selects the arm with highest sample mean.

Williamson et al. (2017) proposed an extension of the DP procedure called CRDP in which (i) the original identity between selected actions and arm allocations is disrupted by a random perturbation (i.e., adding randomization), and (ii) it is allowed to introduce penalties in undesirable end-of-trial states (i.e., adding constraining). They proposed that a good trade-off between patient benefit and statistical properties may be achieved by setting the randomization parameter to 0.9 and by penalizing the states with less than $0.15T$ observations on either arm. Note that DP is recovered by setting the randomization parameter to 1.0, while EFR is recovered by setting it to 0.5 (and not penalizing any states).

4 Illustrative Results for Two-armed Problem

We re-examine the experimental setting presented in Villar et al. (2015, Section 5.1), and we re-print results from the original Villar et al. (2015, Table 5) in Table 1

	$H_0 : p_0 = p_1 = 0.3$			$H_1 : p_0 = 0.3\ ,\ p_1 = 0.5$			
	z-test 0.95/0.98	F-test 0.91/0.95	EPASA (SD)	z-test 0.95/0.98	F-test 0.91/0.95	EPASA (SD)	ENS (SD)
EFR	0.051/0.021	0.058/0.024	0.500 (0.041)	0.805/0.676	0.755/0.589	0.500 (0.041)	59.200 (5.960)
LFF	0.054/0.023	0.057/0.024	0.500 (0.029)	0.804/0.672	0.746/0.567	0.586 (0.033)	61.735 (6.199)
2UCB	0.063/0.031	0.068/0.033	0.500 (0.101)	0.786/0.637	0.707/0.497	0.727 (0.077)	65.915 (6.543)
1UCB	0.073/0.038	0.079/0.040	0.500 (0.142)	0.751/0.581	0.652/0.432	0.785 (0.090)	67.638 (6.724)
0.5UCB	0.089/0.049	0.095/0.050	0.500 (0.199)	0.650/0.442	0.547/0.308	0.838 (0.103)	69.219 (6.894)
0.25UCB	0.097/0.051	0.105/0.051	0.500 (0.271)	0.462/0.243	0.379/0.173	0.872 (0.134)	70.221 (7.299)
0.18UCB	0.091/0.047	0.101/0.047	0.500 (0.308)	0.356/0.158	0.308/0.104	0.877 (0.163)	70.356 (7.740)
0UCB	0.001/0.000	0.001/0.000	0.500 (0.483)	0.012/0.007	0.011/0.004	0.692 (0.445)	64.883 (14.51)
37C+0.8RDP	0.063/0.030	0.068/0.031	0.500 (0.181)	0.746/0.600	0.663/0.478	0.714 (0.060)	65.527 (6.240)
22C+0.9RDP	0.077/0.040	0.085/0.040	0.500 (0.259)	0.650/0.492	0.565/0.371	0.801 (0.097)	68.116 (6.696)
15C+0.95RDP	0.091/0.048	0.101/0.049	0.500 (0.298)	0.580/0.412	0.504/0.314	0.840 (0.118)	69.270 (7.021)
0.95RDP	0.090/0.047	0.104/0.048	0.500 (0.313)	0.511/0.346	0.454/0.264	0.856 (0.144)	69.726 (7.455)
0.99RDP	0.077/0.031	0.097/0.034	0.500 (0.344)	0.323/0.170	0.308/0.123	0.882 (0.166)	70.504 (7.849)
37C+DP	0.063/0.030	0.068/0.031	0.500 (0.209)	0.715/0.575	0.634/0.461	0.734 (0.050)	66.128 (6.159)
30C+DP	0.068/0.032	0.073/0.036	0.500 (0.244)	0.675/0.523	0.586/0.407	0.776 (0.066)	67.371 (6.320)
22C+DP	0.076/0.040	0.086/0.039	0.500 (0.282)	0.604/0.453	0.522/0.344	0.820 (0.089)	68.682 (6.600)
15C+DP	0.092/0.047	0.105/0.047	0.500 (0.313)	0.536/0.376	0.467/0.288	0.854 (0.114)	69.666 (6.962)
7C+DP	0.089/0.029	0.116/0.032	0.500 (0.343)	0.411/0.250	0.369/0.219	0.880 (0.151)	70.441 (7.590)
DP	0.073/0.026	0.094/0.028	0.500 (0.352)	0.263/0.116	0.262/0.078	0.888 (0.172)	70.696 (7.964)
WI	0.065/0.022	0.090/0.024	0.500 (0.363)	0.233/0.102	0.240/0.069	0.887 (0.184)	70.667 (8.185)
ORACLE	0.000/0.000	0.000/0.000	0.500 (0.500)	0.000/0.000	0.000/0.000	1.000 (0.000)	74.000 (6.083)

Table 2 Comparison of different two-arm procedures for a trial of size $T = 148$ by exact calculation; all values are rounded to three digits. The first two columns report the Type I error under the null hypothesis and Power under the alternative hypothesis, respectively, of one-sided tests. F-test: Fisher's exact test; $\{0.91, 0.95, 0.98\}$: one-sided confidence levels; SD: uncorrected standard deviation. Note that ENS (SD) under the null hypothesis is 44.400 (5.575) for all procedures.

(adapting the notation and terminology to this paper) for easy reference. The table shows the results for a variety of two-arm procedures under both the null and alternative hypotheses. The size of the trial was set to be $T = 148$ to ensure that a traditional balanced design with EFR attains at least 80% power when rejecting H_0 with a (maximum) one-sided 5% Type-I error rate using the z-test.

In Table 2 we re-evaluate the measures of the trials designed using some of these procedures. The table presents results in the same scenarios as originally presented in Table 1, but it makes several extensions, improvements and corrections. First, we complement the results presented originally by including additional procedures, and provide a more robust picture due to the employment of several statistical tests and confidence levels. While the original table was obtained by simulations, the results presented here are for all the procedures obtained by *exact calculations* using the backward recursion (and are thus accurate up to a computer's numerical precision), as proposed in Jacko (2019b). A few measures are also calculated slightly differently. EPASA originally included the prior (i.e., 2 pseudo-allocations on each arm), so it slightly underestimated the value of EPASA reported here, which is based on observed (or realized) allocations only. Hypothesis testing is performed using both a

z-test and a Fisher's exact test for comparing two binomial distributions. The z-test was originally based on uncorrected variances (in order to allow for calculation of the variance even for arms with a single observation), while here we use it with corrected variances (using Bessel's correction to obtain an unbiased variance estimator); we moreover require that each arm has at least one observed success and at least one observed failure at the end of the trial in order for the z-test to be employed, otherwise the null hypothesis is not rejected; and we use the exact critical value instead of the rounded 1.645. The Fisher test was originally two-sided, while here we report a one-sided variant, which might not be fully equivalent due to the asymmetry of this test; moreover, the one-sided significance level was originally adjusted (increased) to achieve the one-sided Type I error of around 0.05, while here we present results for significance levels of 0.05 and 0.09 (i.e., confidence levels of 0.95 and 0.91). Both the original and our table report standard deviation, even though the original table in Villar et al. (2015) referred to it as "s.e.".

As discussed in Villar et al. (2015), if one compares a traditional EFR procedure to response-adaptive procedures (including bandit procedures) in the two-armed setting, the first realization is that power is always higher in EFR but its patient benefit metrics are always lower. Adaptive procedures have their power reduced because they induce correlation among intervention allocations; for the deterministic policies like the DP and UCB this effect is the most severe because they almost permanently skew intervention allocation towards an intervention as soon as one exhibits a certain advantage over the other arms. This table shows the tension between learning (high power) and earning (high EPASA and ENS) and how different procedures settle for a different balance between these two objectives.

Both tables show that even EFR leads to an inflated Type I error using the z-test because of not having at least a certain number of both successes and failures on each arm in order for the normal distribution to be an acceptable approximation of the binomial distribution. Academic literature typically recommends that number to be 5 or 10. In this scenario, we would need to require to have at least 11 successes and 11 failures on each arm in order to obtain a Type I error below the significance level of 0.05 (giving Type I error 0.0497 and power 0.8033). Looking at Table 2, LFF also leads to a slightly inflated Type I error under the z-test, but the power is almost the same as that of EFR, while bringing a notable patient benefit of 2.535 additional expected successes. Under the F-test, the Type I error of these two procedures is practically equal and notably below the significance level, while the power of LFF is slightly lower.

Table 2 also includes ORACLE, which is the procedure that assumes that the success probabilities are known, so it allocates all the patients to the superior arm; in case of a tie (i.e., under the null hypothesis), it randomly picks one of the arms at the beginning of the trial and sticks to it. Under the alternative hypothesis, this procedure provides an upper bound for EPASA and ENS, and a benchmark for SD of ENS (which is almost the same as that of EFR). Under the null hypothesis, it leads to the highest SD of EPASA of 0.500. Note that oUCB comes close to it (0.483), because this procedure is essentially a (frequentist) myopic procedure allocating the patients to the arm with the currently highest sample mean. A Bayesian version of the myopic

procedure is CB in Table 1, which allocates using the current belief (the mean of the posterior distribution). All the three procedures are extremely aggressive and they almost never end the trial with at least 1 success and 1 failure on each arm, and so their Type I error and Power are extremely low (unless the significance level is adjusted). We also see that under the alternative hypothesis, both 0UCB and CB are outperformed by many other procedures, and their SDs of ENS and of EPASA are notably larger than those of all the other procedures. It is thus clear that these two procedures are not good choices.

In terms of patient benefit, we look at both tables and focus on ENS under the alternative hypothesis (because EPASA was calculated slightly differently, as described above). The highest ENS is achieved by DP (70.696), closely followed by WI (70.667) in Table 2. We believe that WI (70.73) in Table 1 is better than DP only due to simulation error, but we do highlight that WI is an excellent approximation to the DP. There are several runners-up with less than 1% ENS suboptimality: 0.99RDP (70.504), 7C+DP (70.441), 0.18UCB (70.356) and GI (70.21). This patient benefit suboptimality comes with higher Type I error and higher power, but there are notable differences between these procedures, depending on the test and confidence level used, with no overall winner. For instance: in three out of the four tests, 7C+DP has lower Type I error than 0.18UCB, but notably higher power and higher ENS; in three out of the four tests, 0.99RDP has higher or equal power and higher ENS than 0.18UCB, but notably lower Type I error; in the two tests at higher confidence level, 7C+DP has lower Type I error and higher power than 0.99RDP, but lower ENS.

Table 2 illustrates the flexibility of each of the three families of procedures: UCB, CRDP, and CDP. For the CDP family, we increase the constraining parameter by approximately $0.05T$, penalizing if there are fewer than 7, 15, 22, 30, 37 observations on each arm. For the CRDP family, we include 0.99RDP and 0.95RDP to illustrate the performance of unconstrained procedures, and then we set the constraining parameter by approximately 0.05 above the complement of the randomization parameter (e.g., for 37C+0.8RDP, the complement of the randomization parameter 0.8 is 0.2, so we set the constraining parameter to $0.25T$). Note that varying the parameters of CRDP and CDP leads to a monotone change in ENS, but varying the α in the UCB leads to a concave change, as there is a maximum around $\alpha = 0.18$, and lower values quickly deteriorate the performance. For all three families, we can see that *the Type I error is concave, while Power is monotone.* These non-monotonicities give scope for parameter optimization if the designer knows the relative importance of the three metrics.

In order to compare among these three families, note that 2UCB, 37C+0.8RDP and 37C+DP are quite similar in the Type I error, under all four tests, and also quite similar in ENS, but there seems to be a mild difference in the power, with 2UCB dominating the other two. Another triple for comparison would be 0.5UCB, 15C+0.95RDP and 15C+DP, for which the conclusion would be similar, except for the F-test at 0.95 confidence level, at which 15C+0.95RDP becomes the best in power. Finally, comparing 0.18UCB, 0.99RDP and 7C+DP, 7C+DP is the best in power for all tests. Note however that 37C+0.8RDP 15C+0.9RDP and 0.99RDP are randomized procedures, while the other two families are deterministic.

We note that TS in Table 1 performs relatively poorly in ENS, outperforming only EFR and LFF, while losing only a bit of power and inflating the Type I error comparing to these two procedures. This may be surprising for the reader but we note that the table reports a finite sample performance of this asymptotically optimal procedure.

In terms of statistical testing (excluding 0UCB and ORACLE from this discussion due to their extremely low Type I errors), there are important differences between the z-test and F-test at confidence level 0.95. The Type I error (expected to be 0.05) of the z-test is inflated by all the procedures, from 0.051 (EFR) up to 0.097 (0.25UCB), while that of the F-test is controlled well (the only inflation is to 0.051 of 0.25UCB), showing its most extreme conservatism for EFR (0.024), LFF (0.024), and DP (0.028). In general, there is a strong correlation of Type I errors between these two tests, z-test achieving approximately twice the Type I error of the F-test. There are also notable differences in power, as the F-test achieves power of between 0.185 and 0.342 lower than the z-test. For the z-test at 0.98, the Type I error is also inflated by all the procedures, from 0.021 (EFR) up to 0.051 (0.25UCB). An attentive reader however might notice that the Type I errors reported for z-test at 0.98 and for F-test at 0.95 are very similar across all the procedures. In fact, except for 22C+DP, for which the relation is opposite by 0.001, the former always leads to a lower or equal Type I error. At the same time, it always leads to a higher power. Similarly, z-test at 0.95 is better than F-test at 0.91 is it always results in a lower Type I error and in a notably higher power. The F-test is often cited as conservative, however, Table 2 shows that at 0.91 confidence level that is not always true, especially for some of the more aggressive procedures, which can even inflate the Type I error. To the best of our knowledge, this is the first time that inflation of the Type I error by the Fisher's exact test has been reported in the literature. These observations suggest that in the null and alternative hypotheses scenarios we have presented, it might be preferable to use z-test over F-test. However, we emphasize that we have discussed only a single pair of scenarios of the null and alternative hypotheses, the performance of statistical tests for binomial samples is very sensitive to the specific scenario parameters and the appropriateness of using these tests is highly dependent on the specifics of each procedure, so we would refrain from any generalizations. In practice, the trial designer could replicate our analysis and study a variety of plausible scenarios. In theory, inference with data obtained by adaptive procedures remains an important open question and requires further research. Some recent examples of work in this area include Hadad et al. (2021); Zhang et al. (2020); Deliu et al. (2021) .

The tables do not include any measures related to estimation, because that on its own has trade-offs between precision and accuracy, which has been left out of this chapter.

5 Discussion

In this section we close the chapter by discussing how (and when) bandit models can be specified to design a clinical trial beyond the traditional assumptions considered in here. These include: the presence or possibility of delayed responses, other practicalities such as dropouts (or patients lost to follow up) and/or missing responses, safety concerns, early evidence of efficacy or futility, and unavailability of prior distributions. We also discuss how bandit models as those reviewed here, which are typically defined for binary outcomes, can be used in practice to accommodate for a primary endpoint that is non-binary through the use of an appropriate surrogate endpoint. Finally, we discuss how the computational limitations of optimal bandit approaches (i.e., those like CRDP for finite size trials) can be mitigated by using an efficient programming language and a more effective coding syntax to allow for designing and evaluating trials with several thousands of patients.

For many of the practicalities discussed below, we discuss how the MDP model of CRDP could be amended, as some of these have been recently explored in the literature. We are not aware how other procedures perform in the presence of them or how could they be adjusted to incorporate each practicality.

5.1 Safety Concerns

Many trials in practice are forced to stop recruitment due to safety concerns by observing secondary endpoints or adverse events, which have nothing to do with the observed (primary endpoint) outcomes on which a response-adaptive procedure is typically based. A designer using a response-adaptive procedure may need to incorporate the possibility of stopping for safety concerns to introduce more control over the number of observations from each arm. This can be done by incorporating the probability of such stopping in the MDP model of the DP and CRDP procedures (which we jointly refer to as (CR)DP), and by specifying constraints or by keeping the degree of randomization relatively balanced in early stages. We are not aware how that could be incorporated to procedures, which are agnostic to the trial size, apart from UCB in which we could perhaps adaptively change the parameter α as the trial evolves.

5.2 Prior Distributions

All the results presented in this paper assume for each arm the Bayes' prior $Beta(1,1)$, which is the uniform distribution, and is commonly considered non-informative. This is the standard choice for binary outcomes in methodological papers using Bayesian framework. Trial designers can however consider an informative one based on data from previous trials. The (CR)DP easily allows also for imple-

menting a decreasingly informative prior (Donahue and Sabo, 2021) by modifying the rewards and transition probabilities between states.

In some situations there is no previous reliable data or willingness to specify the prior distributions for each arm. In that case, the trial could have an initial phase in which a non-adaptive randomization procedure is used, and bandit approach is employed only after that phase accumulates sufficient amount of information, which will be taken as the prior distribution for the (CR)DP procedures. In Williamson and Villar (2020) some sensitivity analysis for different informative priors in a continuous endpoint case paired with an randomized index procedure is illustrated.

5.3 Delayed Responses

Williamson et al. (2021) evaluated how the (CR)DP procedure performs in two-armed trials with both fixed and random delays in responses (i.e., in observations of outcomes). This is an important question in practice which is natural to ask about any response-adaptive procedure. To summarize, they illustrated that one gains slightly in terms of power and bias through the delay, so in that sense delay could be viewed as a positive attribute from the statistical point of view (which seems somewhat counter-intuitive), but one loses in terms of patient benefit which is the main advantage of using such response-adaptive procedures over alternatives. However, this loss is not overly concerning and for a relatively large fixed delay length, for example, one third of the sample size 75, the percentage of patients on the superior arm when $p_0 = 0.5$ and $p_1 = 0.1$ is approximately 23% higher for CRDP and 30% higher for DP than the traditional approach of EFR. Further, when compared to the performance of the most commonly studied procedure for delayed responses scenarios (Hardwick et al., 2006), namely the Delayed Randomized Play-the-Winner Rule (DRPWR), there are still considerable improvements with respect to the patient benefit for (CR)DP. Therefore, this evaluation has shown that the (CR)DP procedures perform well in trials with delayed responses since they continue to dominate in terms of the patient benefit over other procedures for a range of (expected) delay lengths.

The investigation in Williamson et al. (2021) leads to a conclusion that it may not be necessary to adjust the CRDP optimization horizon (i.e., to decrease T by the delay length d) if the delay is large enough to satisfy the desired constraints already by the equal fixed-randomization of the first $d + 1$ patients, and essentially such constraints may not need to be included in the optimization model at all. For smaller delays, if the designer decides to adjust the horizon, it might be beneficial for fine tuning of the procedure to also appropriately adjust the constraining parameters taking into account the observations of the patients which will be revealed after the recruitment of the last patient. Another option the designer has is to reach the desired trial design objectives for statistical operating characteristics (high power, small bias) by modifying the randomization probabilities, either for the early patients that are fixed-randomized before the first observation or for the remaining patients that are allocated using the CRDP procedure, or both.

Special attention needs to be paid if there is a possibility of overly delayed responses so that these are not observed by the time of the final analysis. In that case, (CR)DP with non-adjusted horizon may not even reach the final stage in which the constraints are specified, so adjusting the horizon seems to be a preferred approach.

5.4 Dropouts and Missing Responses

When designing a randomized controlled trial, the designer needs to account for the possibility of dropouts and missing responses, i.e., patients who are recruited and get allocated to one of the arms, but we fail to observe their response, either because they leave the trial or their outcome is erroneous. A simple approach the designers can take is to estimate the probability of missing responses and inflate the trial size so that the expected number of observations excluding the missing responses is the desired one. With (CR)DP we can take this possibility into account by adjusting the procedure optimization horizon by a constant, e.g. for a trial size T, taking the procedure horizon $T - m$, where m is an estimate of the number of missing responses, and correspondingly specify the constraints for the final stage $T - m$. It is also possible to consider a random number of missing responses, which would keep the procedure horizon T but would include constraints not only in the final stage, but also in previous stages which we would like to avoid. In that case, the state-transition probabilities of the MDP model of the (CR)DP procedure could be modified to account for the probability of observed dropouts or erroneous outcomes.

5.5 Early Evidence of Efficacy or Futility

Although the trial size is usually planned based on existing data and/or expert opinion about the expected intervention effect (i.e. difference between the two intervention success probabilities), such estimates likely come with a large variance and bias. Both frequentist and Bayesian concepts have been developed to identify situations during the trial which would identify sufficient evidence of efficacy or futility of an intervention. In case of evidence of futility of a novel intervention, recruitment to this arm should be stopped to keep patient benefit for the remaining in-trial patients at least at the level of thecurrent standard of care. In case of evidence of efficacy of a novel intervention, there are two common design approaches: (1) a decision as a result of an interim analysis is made to stop the recruitment to the novel arm, and the intervention to "graduate" to another separate trial to confirm efficacy, or (2) the trial seemlessly transforms to such a confirmatory trial without an explicit interim analysis.

Both cases can be incorporated in the MDP model of the CRDP procedure. For instance, consider a state of the trial with 5 observations on each arm, with the most extreme data: 5 successes and 0 failures on one arm, and 0 successes and 5 failures on

the other arm. The Fisher's exact test would give a one-tailed p-value of 0.004 based on this data, showing evidence of difference between the two arms. In case of an interim analysis which would stop recruitment for futility of the novel arm, the MDP model of the CRDP procedure can be modified by assuming that all the remaining in-trial patients will be allocated to the control arm, i.e., by modifying the reward of that state and by modifying the state-transition probabilities to "jump" to the end of the trial. In case of an interim analysis which would stop recruitment for efficacy of the novel arm, the MDP model can be modified by assuming that all the remaining in-trial patients will be randomized in the new separate trial, i.e., by modifying the reward of that state and by modifying the state-transition probabilities to "jump" to the end of the trial. In case of a seamless transformation of the trial, the degree of randomization of the subsequent states can be defined differently from the degree of randomization of the subsequent states that do not show such a strong evidence, so, effectively, further generalizing the CRDP procedure to allow for randomization p to depend not only on arm j and time stage t as in Williamson et al. (2021) but also on the state (i.e., numbers of successes and failures) itself.

5.6 Non-binary Outcomes

Development of an analogous randomization procedure to (CR)DP when the primary endpoint is non-binary is theoretically possible, but computationally will become infeasible for much smaller trial sizes than the current variant for binary outcomes. The designer could still however employ the binary-outcomes (CR)DP by using a dichotomization of the primary endpoint or by using an auxiliary endpoint correlated with the primary endpoint. Although dichotomization may not lead to as high patient benefit as theoretically achievable using the original endpoint, if meaningfully defined it could lose only a negligible amount and thus still bring important patient benefit over alternative response-adaptive procedures. The degree of randomization could be adjusted in order to reflect the designer's confidence in the correlation between the primary and auxiliary endpoint. See, for instance, Williamson and Villar (2020) for such an investigation for normally distributed outcomes.

5.7 Exploratory Trials

In a two-armed setting, we discussed and illustrated the conflict between patient benefit (patient outcomes) and relevant statistical features (error levels and estimation metrics). In the two-arm setting, there is little scope for a bandit procedure to be superior to EFR in terms of the latter metrics. In a multi-armed setting (as for example large platform trials are), this is not necessarily the case, and depending on the main objective of the trial (e.g. the specific statistical power definition used) and the type of bandit procedure, one can find alternatives that may be superior to

EFR in both the statistical features and patient benefit. Exploratory trials, which are often multi-armed, are moreover not meant to directly lead to a regulatory approval, and thus may not need to perform in statistical operating characteristics as strictly as confirmatory trials would need to.

This was illustrated in Villar et al. (2015, Table 6) reproduced here as Table 3 for easy reference. The results in there show how some randomized and semi-randomized bandit procedures (i.e., TS, 2UCB, RBI, RGI) exhibit an advantage over EFR both in the achieved power and in ENS. These procedures continue to allocate patients to all arms during the trial while skewing allocation to the best performing arm, hence, ensuring that by the end of the trial the control arm will have a similar number of observations as with EFR while the best arm will (in expectation) have a larger number. Among these procedures, TS and 2UCB exhibit the best performance in power and ENS as they are both greater than those achieved by EFR, although they cause a slight inflation of the Type I error. Whie RBI and RGI were performing somewhat similarly to TS and 2UCB in the two-armed setting shown in Table 1, their performance in ENS terms is notably inferior in the multi-armed setting shown in Table 3.

The deterministic index-based procedures CB and GI increase the advantage in ENS over EFR even more, while the Type I error is controlled using an adjusted Fisher test. However, this conservative test causes a severe reduction in power of these procedures. A simple way to overcome the severe loss of statistical power of the deterministic procedures in the the multi-armed setting introduced in Villar et al. (2015) suggests to use a composite procedure in which the (random) allocation to the control arm is protected and the allocation to experimental arms is guided by a deterministic procedure. For example, in Table 3 results are shown for a procedure in which one in every K patients (note that K is the number of experimental arms) is allocated to the control group while the allocation of the remaining patients among the experimental treatments is done using the Gittins index procedure. This procedure was referred in there as the controlled Gittins index (CGI) procedure. Simulation results show that a simple procedure like CGI manages to solve the trade-off quite successfully, in the sense that it achieves the highest power, lowest Type I error and an ENS very close to that achieved by the myopic CB procedure but with a third of the variability that CB exhibits.

5.8 Large Trials

Williamson et al. (2017) developed the (CR)DP procedure in the context of rare diseases, and thus focused on relatively small trial sizes. They provided an "efficient algorithm" for (CR)DP implemented in the statistical software R and reported that the maximum time horizon that "can be computed on a standard laptop using R is $T = 215$" and that computations are "feasible on a standard performance workstation (1 TB of RAM) for $215 < T < 600$". Jacko (2019b,a) however showed that much larger horizons are possible to compute on standard computer (with 32 GB RAM)

	Crit.	$H_0 : p_0 = p_1 = p_2 = p_3 = 0.3$			$H_1 : p_0 = p_1 = p_2 = 0.3, p_3 = 0.5$		
	Value	Type I E	EPASA (SD)	ENS (SD)	Power	EPASA (SD)	ENS (SD)
EFR	2.128	0.047	0.250 (0.02)	126.86 (9.41)	0.814	0.250 (0.02)	148.03 (9.77)
TS	2.128	0.056	0.251 (0.07)	126.93 (9.47)	0.884	0.529 (0.09)	172.15 (13.0)
2UCB	2.128	0.055	0.251 (0.06)	126.97 (9.41)	0.877	0.526 (0.07)	171.70 (11.9)
RBI	2.128	0.049	0.250 (0.03)	126.77 (9.40)	0.846	0.368 (0.04)	158.34 (10.4)
RGI	2.128	0.046	0.250 (0.03)	126.80 (9.36)	0.847	0.358 (0.03)	157.26 (10.3)
CB	F_a	0.047	0.269 (0.39)	126.89 (9.61)	0.213	0.677 (0.41)	184.87 (36.8)
GI	F_a	0.048	0.248 (0.18)	126.68 (9.40)	0.428	0.831 (0.10)	198.25 (13.7)
CGI	2.128	0.034	0.250 (0.02)	127.16 (9.46)	0.925	0.640 (0.08)	182.10 (12.3)
ORACLE		0.000	0.250 (0.43)	126.90 (9.42)	0.000	1.000 (0.00)	211.50 (10.3)

Table 3 Comparison of procedures in a four-arm trial of size $T = 423$ by simulation. F_a: Fisher's adjusted test; Type I E: family-wise type I error; CGI: Controlled Gittins index. Re-printed (adapted) from Villar et al. (2015, Table 6).

if using a more efficient programming language (Julia) and a more effective coding syntax, with up to $T = 4,500$ for online calculation and $T = 1,500$ for offline calculation (storing the whole (CR)DP procedure allocations in an array for saving on a hard disk).

The (CR)DP procedure could be in theory generalized to more than 2 arms, but in practice that might lead to computationally infeasible model. Alternatives which closely approximate the DP procedure are the Whittle index and the Gittins index (Villar et al., 2015; Villar, 2018; Jacko, 2019b). However, their modifications to include constraints like in the CRDP procedure have not been developed yet and may not always be possible, especially for constraints that depend on more than one arm, because the Whittle and Gittins indices crucially function by decomposing the trial-level optimization problem into single-arm optimization subproblems. Nevertheless, single-arm constraints such as about the number of observations from each arm should be implementable. If constraints are not required, then the degree of randomization can be easily implemented using the Whittle or Gittins index instead of the DP procedure in the alternative interpretation described in Williamson et al. (2021).

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256.

Bowden, J. and Trippa, L. (2017). Unbiased estimation for response aadaptive clinical trials. *Statistical Methods in Medical Research*, 26(5):2376–2388.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Burnett, T., Mozgunov, P., Pallmann, P., Villar, S. S., Wheeler, G. M., and Jaki, T. (2020). Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC medicine*, 18(1):1–21.

Cserna, B., Petrik, M., Russel, R. H., and Ruml, W. (2017). Value directed exploration in multi-armed bandits with structured priors. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*.

Deliu, N., Williams, J. J., and Villar, S. S. (2021). Efficient inference without trading-off regret in bandits: An allocation probability test for thompson sampling. *arXiv preprint arXiv:2111.00137*.

Donahue, E. and Sabo, R. T. (2021). A natural lead-in approach to response-adaptive allocation for continuous outcomes. *Pharmaceutical Statistics*, pages 1–10.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15).

Hardwick, J., Oehmke, R., and Stout, Q. F. (2006). New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference*, 136:1940–1955.

Jacko, P. (2019a). BinaryBandit: An efficient Julia package for optimization and evaluation of the finite-horizon bandit problem with binary responses. Management Science Working Paper 2019:4, Lancaster University Management School.

Jacko, P. (2019b). The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths. Management Science Working Paper 2019:3, Lancaster University Management School. arXiv:1906.10173.

Kelly, F. (1981). Multi-armed bandits with discount factor near one: the Bernoulli case. *Annals of Statistics*, 9(5):987–1001.

Melfi, V. and Page, C. (1998). Variablility in adaptive designs for estimation of success probabilities. *Lecture Notes-Monograph Series*, 34, New Developments and Applications in Experimental Design:106–114.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2018). Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Sydes, M. R., Villar, S. S., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):1–15.

Pilarski, S., Pilarski, S., and Varró, D. (2021). Optimal policy for bernoulli bandits: Computation and algorithm gauge. *IEEE Transactions on Artificial Intelligence*, 2(1):2–17.

Press, W. H. (2009). Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392.

Robertson, D. S., Lee, K. M., Lopez-Kolkovska, B. C., and Villar, S. S. (2021). Response-adaptive randomization in clinical trials: From myths to practical considerations. *arXiv preprint arXiv:2005.00564*.

Rosenberger, W. F. (1999). Randomized play-the-winner clinical trials: Review and recommendations. *Controlled Clinical Trials*, 20(4):328–342.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, 38(1):1–12.

Ryzhov, I. O. (2021). Optimal learning and optimal design. In *This book*.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P. K., Koenig, F., Krisam, J., Mozgunov, P., et al. (2020). Efficient adaptive designs for clinical trials of interventions for covid-19. *Statistics in Biopharmaceutical Research*, 12(4):483–497.

Sverdlov, O., Tymofyeyev, Y., and Wong, W. K. (2011). Optimal response-adaptive randomized designs for multi-armed survival trials. *Statistics in medicine*, 30(24):2890–2910.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Villar, S. S. (2018). Bandit strategies evaluated in the context of clinical trials in rare life-threatening diseases. *Probability in the Engineering and Informational Sciences*, 32:229–245.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215.

Wei, L. J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society, Series B*, 42(2):143–149.

Williamson, S. F., Jacko, P., and Jaki, T. (2021). Generalisations of a Bayesian decision-theoretic randomisation procedure and the impact of delayed responses. *Computational Statistics and Data Analysis*. Available online 7 December 2021. In press.

Williamson, S. F., Jacko, P., Villar, S. S., and Jaki, T. (2017). A Bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics and Data Analysis*, 113C:136–153.

Williamson, S. F. and Villar, S. S. (2020). A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, 76(1):197–209.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146.

Zhang, K., Janson, L., and Murphy, S. (2020). Inference for batched bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9818–9829. Curran Associates, Inc.