European Journal of Operational Research 248 (2016) 619–633

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients

Vishal Ahuja[1],*, John R. Birge

Booth School of Business, The University of Chicago, 5807 S Woodlawn Ave., Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Received 11 July 2014
Accepted 30 June 2015
Available online 11 July 2015

Keywords:
OR in health services
Adaptive clinical trials
Markov decision process
Bayesian learning
Stents

ABSTRACT

Clinical trials have traditionally followed a fixed design, in which randomization probabilities of patients to various treatments remains fixed throughout the trial and specified in the protocol. The primary goal of this static design is to learn about the efficacy of treatments. Response-adaptive designs, on the other hand, allow clinicians to use the learning about treatment effectiveness to dynamically adjust randomization probabilities of patients to various treatments as the trial progresses. An ideal adaptive design is one where patients are treated as effectively as possible without sacrificing the potential learning or compromising the integrity of the trial. We propose such a design, termed *Jointly Adaptive*, that uses forward-looking algorithms to fully exploit learning from multiple patients simultaneously. Compared to the best existing implementable adaptive design that employs a multiarmed bandit framework in a setting where multiple patients arrive sequentially, we show that our proposed design improves health outcomes of patients in the trial by up to 8.6 percent, in expectation, under a set of considered scenarios. Further, we demonstrate our design's effectiveness using data from a recently conducted stent trial. This paper also adds to the general understanding of such models by showing the value and nature of improvements over heuristic solutions for problems with short delays in observing patient outcomes. We do this by showing the relative performance of these schemes for maximum expected patient health and maximum expected learning objectives, and by demonstrating the value of a restricted-optimal-policy approximation in a practical example.

1. Introduction [2]

The costs of bringing a new drug to market have been estimated to be as high as $5 billion (Forbes, 2013). Clinical trials have been cited as a key factor in raising these costs, with phase III trials now representing about 40 percent of pharmaceutical companies' R&D expenditures (Roy, 2012). The total cost of a clinical trial can reach $300–$600 million (English et al., 2010), potentially an order of magnitude higher when including the value of remaining patent life,[3] and exceed $6000 per enrolled subject (Emanuel, Schnipper, Kamin, Levinson, & Lichter, 2003). Consequently, drug manufacturers face pressure to produce conclusive results faster and reduce the number of subjects required.

Traditionally, clinical trials have followed a non-adaptive or a *fixed* design that randomizes patients to treatments in a constant proportion (probabilistically) throughout the trial. Such a design, in use for several decades, is well-understood by practitioners, and provides a clean way of separating treatments. Common reasons for the prevalence of such designs include a desire to maintain low probabilities of type I error and to protect against bias. However, these designs often result in lengthy trials, poor patient outcomes, and inconclusive results, leading to longer times for drug approval. In recognition of these issues, regulatory bodies, such as the U.S. Food and Drug Administration (FDA), have encouraged the use of adaptive designs (FDA, 2010a, 2010b).

* Corresponding author. Tel.: +1 214 768 3145.
E-mail addresses: vahuja@smu.edu (V. Ahuja), john.birge@chicagobooth.edu (J.R. Birge).

[1] Present address: Cox School of Business, Southern Methodist University, 6212 Bishop Boulevard, Dallas, TX 75275, USA.

[2] This work arose through discussions with our colleagues at The University of Chicago Medical Center, who were interested in a practically implementable adaptive design for a trial such as Clinical Antipsychotic Trials of Intervention Effectiveness Study (CATIE). The NIMH-funded trial, that compared schizophrenia drugs, suffered from several shortcomings, primarily those relating to patient compliance (Lieberman et al., 2005).

[3] Given that the patent for a drug or an intervention is typically filed before clinical trials begin, shortening the trial length can significantly increase potential revenues, not to mention the potential health benefit for the patients outside of the trial. For example, the sales of the drug *Atorvastatin* (trade name: *Lipitor*) decreased by 42 percent, from $2.4 billion to $1.4 billion, after its expiration on November 30, 2011 (Forbes, 2012).

http://dx.doi.org/10.1016/j.ejor.2015.06.077
0377-2217/

There exist several types of adaptive designs (see Chow and Chang, 2008 for a comprehensive list); a commonly used design, and the focus of this work, is the *outcome-* or *response-adaptive design*. Such designs, typically Bayesian in nature, employ learn-and-confirm concepts, accumulating data on patient responses, which is then used to make procedural modifications while the trial is still underway, increasing the likelihood of selecting the *right* treatment for the *right* patient population earlier in a drug development program. Adaptive designs can potentially increase the probability of finding the successful treatment, identify ineffective and unsafe drugs sooner, and require fewer patients in the trial, thereby reducing costs and shortening development timelines. Adaptive designs can also offer a safer alternative to fixed designs, allowing patients, who are initially allocated to a relatively unsafe treatment, to be switched to the safer treatment, as and when it becomes evident during the course of the trial. Henceforth, we will use the term *adaptive* to mean *response-adaptive* design.

The inherent flexibility of a Bayesian adaptive design appears contrary to the established fixed design. Common criticisms of adaptive designs include perceptions of reduced ability to do classical tests of statistical hypotheses, especially control of type I error that FDA requires for regulatory approval (FDA, 2010a, 2010b). Berry and Eick (1995) argues that such objections are either due to a lack of understanding or involve issues that can easily be addressed, for example, by incorporating constraints into the adaptive design (Cheng & Berry, 2007). Berry and Eick (1995) and LeBlond (2010) propose the use of computer simulations to evaluate type I error rate in Bayesian approaches.[4] In fact, the cholesterol-lowering drug Pravigard PAC was the first FDA approval that took a primarily Bayesian focus (Berry, 2006). In addition, the FDA has approved a number of medical devices whose submissions utilized a Bayesian statistical method (LeBlond, 2010). Further, inferential measures such as *predictive probability* make Bayesian approaches better suited for interim analyses as they provide the ability to quantify subsequent trial results given current information (Berry, 1985; 1987; 1993; Lee & Liu, 2008).

Berry and co-authors were among the first to develop a truly Bayesian response-adaptive design (see, for example, Berry, 1978; Berry & Pearson, 1985). In their design, patient randomization to treatments happens sequentially, that is, one at a time and all previous patient response(s) are known and incorporated into the randomization decision(s) for the following patient(s). This design is reasonable for trials where a single patient is randomized at each period, as in the case of individualized therapy trials, or when there is minimal delay in observing outcomes. However, this design is not practically useful when multiple patients need to be randomized simultaneously. One could implement variations of this design, for example by randomizing all the patients at a stage with a probability calculated for a single patient using the existing sequential design. However, such designs are suboptimal in a model that recognizes available information and the timing of opportunities to gather more information, required for updating the policy. We address this gap by developing an adaptive design with multiple simultaneous randomizations to anticipate learning through the trial horizon; we call this the *Jointly Adaptive* design. Following existing literature, we assume that patients are exchangeable, their outcomes are observable before each randomization decision, there is no serial correlation in treatment effects, and the treatment effects remain the same at each stage of the trial.

1.1. Main contributions and organization of the paper

The key contribution of this paper is the development of a Bayesian MDP framework for finite-horizon problems that learns optimally from simultaneous multiple experiments, admits continuous controls, and can be used to evaluate treatments under multiple objectives. In the context of clinical trials, our contributions are the development of a practically implementable response-adaptive design (termed *Jointly Adaptive*) that learns simultaneously from multiple patients and optimally randomizes them to multiple treatments. Further contributions include consideration of a learning objective in addition to the health objective, and evaluation of the relative advantage of the *Jointly Adaptive* design over other implementable response-adaptive designs, the fixed design, and heuristics. We note that our model is generalizable to other MDP settings that involve learning from multiple simultaneous individual experiments, as in the case of customized consumer offers.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the literature. Section 3 presents the underlying model for the *Jointly Adaptive* design. Section 4 describes various adaptive designs and provides some theoretical guarantees. In Section 5, we present numerical results, including application to a recently conducted clinical trial. In Section 6, we summarize and discuss our conclusions as well as the scope and limitations of the adaptive designs.

2. Literature overview

The majority of previous work on trial design appears in the field of statistics. The class of problems involving adaptive designs has its roots in the *multi-armed bandit* problem that balances maximizing reward using knowledge already acquired with undertaking new actions to further increase knowledge, commonly referred to as the *exploitation vs. exploration* tradeoff.

The study of heuristics for the multi-armed bandit problem has a long history. Robbins (1952) is one of the earliest works on this topic that investigated the *play-the-winner* rule in a two-armed bandit problem. Bellman (1956) is one of the first to study the problem of sequential design of experiments using backward induction. Gittins (1979) employs a Dynamic Allocation Index, also called the Gittins Index, to solve bandit problems using forward induction; Katehakis and Veinott (1987) characterizes this index in a way that allows it to be calculated more easily.

Berry (1978) is one of the first studies to fully incorporate a Bayesian learning approach in a two-armed bandit. Extensions to this model include: (a) Berry and Eick (1995), which considers an objective that incorporates the conflicting goals of treating patients as effectively as possible during the trial and, with high probability, correctly identifying the relative efficacy of each treatment, and (b) Cheng and Berry (2007), which proposes a constrained adaptive design to address the “treatment assignment bias” concern raised in the literature (e.g., Chalmers, Celano, Sacks, & Harry Smith, 1983); their constraint ensures that each treatment in the trial has a certain fixed minimum probability of being chosen at each allocation decision. We refer the readers to Berry and Fristedt (1985) for further applications and note that adaptive designs have typically focused on maximizing expected patient health.

A related stream of literature has investigated asymptotically adaptive policies for bandit problems to achieve an optimal rate of regret. Lai and Robbins (1985) is a seminal study whose proposed adaptive policy achieves a $O(\log n)$ lower bound on the regret. Extensions of this study and other examples include Burnetas and Katehakis (1996), Auer, Cesa-Bianchi, and Fischer (2002), and Honda and Takemura (2010). For further details, we direct the readers to these papers and references therein.

Another stream of related literature includes evaluation of adaptive *treatment* strategies, defined by sequences of decision rules on when and how to alter the treatment of a patient in response to outcomes (Murphy, 2005). Such designs share several features with adaptive *trial designs*, for example, the use of past patient responses.

[4] A commercial software that does this simulation is called FACTS™, see www.berryconsultants.com/software/ for details.

The trials of adaptive strategies to treat a patient can either follow a *fixed design* (with adaptive strategy replacing traditional treatment) or an *adaptive design* (where adaptive treatment strategies change dynamically). In this paper, we focus on adaptive *trial* designs for specific treatments but note that this approach is also applicable to consideration of adaptive treatment strategies.

Most adaptive designs assume a constant delay in observing outcomes that corresponds with the next set of allocation decisions. Hardwick, Oehmke, and Stout (2006) relaxes this assumption by incorporating varying delays. In particular, they assume independent exponential response times such that a patient response may not be available at the next randomization opportunity. For our model, we assume a *constant* delay and no opportunities to learn before the first allocation decision. The important practical aspect that we capture in contrast to most prior work is that multiple patients receive treatment assignments simultaneously before the outcomes of the previous assignment can be observed but that delays are limited so that some sequential structure is retained.

The Bayesian learning setup appears in many other areas besides clinical trials. Examples in the OR/MS literature include work on dynamic assortments in retailing (e.g., Caro & Gallien, 2007) and dynamic learning about employee performance to formulate an employer's hiring and retention decisions (Arlotto, Chick, & Gans, 2013). Concerning dynamic pricing problems, the setup has been used to estimate unknown parameter(s) that characterize the underlying demand function. Recent examples include Aviv and Pazgal (2005), Farias and Van Roy (2010), and Harrison, Keskin, and Zeevi (2012). Readers are directed to Besbes and Zeevi (2009) as an example that uses classical statistics framework to learn about the underlying demand function. Harrison et al. (2012) discusses further connections to antecedent literature.

A paper that is close to our work is Bertsimas and Mersereau (2007), which uses multi-armed bandit framework in an interactive marketing context. Our work differs from Bertsimas and Mersereau (2007), as follows. First, we allow for randomized strategies, while Bertsimas and Mersereau (2007) restrict choices to integers that restrict the control space. Second, we consider a maximum expected learning objective, defined as the expected probability of correctly identifying the most efficacious treatment at the end of the trial, in addition to the maximum expected successes objective that Bertsimas and Mersereau (2007) considers. Further, we analyze the tradeoff between the two objectives. Third, our design provides flexibility that, in essence, differs from that offered by Bertsimas and Mersereau (2007). For example, in restricted-optimal-policy approximation, our design allows for an optimal solution developed for a smaller problem to be applied to a larger cohort using a heuristic. Finally, our work is specifically tailored to the trials context, in contrast to Bertsimas and Mersereau (2007), which is in an interactive marketing context. We evaluate and compare multiple strategies, thus making it relevant for clinicians, regulatory agencies, and other concerned parties. This is also reflected in our numerical results, where, for example, we show the value of the optimal solution under a wide variety of initial conditions, reflecting the reality that clinicians differ widely in their prior beliefs about the success probability of a specific treatment.

Our model incorporates uncertainty in parameter estimates, usually missing from fixed designs**.** While previous literature uses constraints to ensure a minimum probability of choosing a treatment (as in Cheng & Berry, 2007), our proposed *Jointly Adaptive* design (modeled in Section 3) includes such randomizations naturally. To the best of our knowledge, this is the first fully response-adaptive trial design that considers multiple patients, incorporates fixed observation delays, and can incorporate optimal solution for a learning objective.

Finally, most studies on multi-armed bandit problems, including ours, assume statistically independent arms. Although not applicable to our setting, Mersereau, Rusmevichientong, and Tsitsiklis (2009) is an example of a study that considers correlated arms and shows that the known statistical structure among arms can be exploited for higher rewards and faster convergence.

3. Model

We formulate the problem as a *Bayes-adaptive* Markov decision process (BAMDP).[5] Unlike the classical MDP setup, the *underlying* probabilities are *unknown* in BAMDP. Instead, we assume a parametric distribution on the transition probabilities at the beginning of the trial, capturing the beliefs of clinicians about each treatment. As more information is obtained in the trial, the beliefs are updated dynamically in a Bayesian fashion.

The BAMDP state is a vector with dimension equal to the number of treatment–outcome combinations, also called *health conditions*. Each component of our Markov chain state, that we call the *health information state*, represents the number of patient observations accumulated in each health condition up to a given stage. The state thus captures the information observed so far (history) and is used to derive the distributions that describe the uncertainty in the transition probabilities. The controls are the probabilities of randomizing patients to each treatment. The rewards depend on the objective function chosen; we consider two objectives: patient health and learning about the efficacy of treatments.

Below we specify the model for the simple case of a trial that consists of two treatments, henceforth referred to as treatments *A* and *B*, and two mutually exclusive outcomes, namely success (*s*) and failure (*f*). We believe that this simple case illustrates the workings of the model and indeed exemplifies many practical trials, although the model can easily be generalized to the case of multiple treatments and/or multiple outcomes. Further, we assume the following: (a) independent treatments, (b) independent and identically distributed (i.i.d.) patients, (c) no serial correlation in treatment effects, and (d) a constant number of patients allocated each period.

Remark. The use of continuous controls in our model allows for probabilistic allocation of multiple patients and helps eliminate selection bias and facilitates blinding (similar to simple randomization in traditional fixed designs). However, this could potentially lead to imbalance between various treatment groups (similar to traditional fixed designs), especially if the number of patients enrolled in the trial is small. Techniques such as block and stratified randomization can be used to address the imbalance issue in fixed designs but they have their own disadvantages.[6] In the case of adaptive design, similar restrictions or correlation structures (tailored to adaptive design) can be put in place; however, we do not address these concerns in this paper.

3.1. Model specification

We describe a basic version of our model with two alternative treatments, two possible outcomes each, and serially independent trials. Additional treatments, outcomes, and more complex dependency structures can be incorporated without materially changing the framework of the model. Let *T* be the trial length or the total number of time periods in the trial, where a time period corresponds to the (constant) time between two allocation decisions, which we assume corresponds with the delay time from initial treatment to observed outcome. We note that the first set of randomizations (decisions) take place at time $t = 0$, the first set of patient outcomes are observed at time $t = 1$, and decisions for patients arriving at time t

[5] We follow the terminology of Duff (2003).

[6] A description of such techniques can be found in many sources, for example, Suresh (2011).

are made at time $t - 1$, $t \in \{0, 1, .., T\}$. No decisions are made at time $t = T$.

Let n be the number of patients allocated per period in the trial. Then $N = nT$ represents the total number of patients (observations) in the trial. Let $J = \{A, B\}$ and $O = \{s, f\}$ be the set of independent treatments and outcomes, respectively. The corresponding set of health conditions, I, is obtained from a Cartesian product of those sets $(J \times O)$ such that $I = \{As, Af, Bs, Bf\}$.

The health information state is a vector, $\mathbf{h}_t \in \mathcal{H} \subseteq \mathbb{Z}^{|J|\times|O|}$, defined as follows:

$$\mathbf{h}_t = (h_t^{As}, h_t^{Af}, h_t^{Bs}, h_t^{Bf}).$$

Here, $h_t^{j,o} \in \mathbb{Z}_+$ represents the cumulative number of observed patients to date in health condition (j, o) at time $t \in \{0, 1, .., T\}$, for all $j \in J, o \in O$, such that $\sum_{j\in J, o\in O} h_t^{j,o} = nt$.

The controls, $\mathbf{u}_t \in \mathcal{U} \subseteq \Re_+^{|J|}$ are defined as follows:

$$\mathbf{u}_t = (u_t^A, u_t^B).$$

Here $u_t^j \in [0, 1]$ is the probability of assigning a patient to treatment $j \in J$ at time t such that $\sum_{j\in J} u_t^j = 1$. Given two mutually exclusive treatments, we only need to calculate controls for one treatment. Without loss of generality, we assume it would be for treatment A (u_t^A), such that $u_t^B = 1 - u_t^A$. The set of allocations, $\mathbf{d_t}$, is defined as:

$$\mathbf{d}_t = (d_t^A, d_t^B).$$

Here, $d_t^j \in \mathbb{Z}_+$ is the *random* number of patients assigned to treatment j at time t. If each patient assignment is independent, d_t^A is obtained from u_t^A as follows: $d_t^A \sim \text{Bin}(n; u_t^A)$,[7] such that $\mathbb{E}d_t^A = nu_t^A$ and $d_t^B = n - d_t^A$.[8]

Let p_t^j represent the (unknown) probability of observing a success with treatment $j \in J$ at time t. The vector of probabilities is defined as follows:

$$\mathbf{p}_t = (p_t^A, p_t^B).$$

Following existing literature, p_t^j is assumed to be Beta distributed with hyperparameters (α_t^j, β_t^j) such that $g(p_t^j) \sim \text{Beta}(\alpha_t^j, \beta_t^j)$ and $\mathbb{E}p_t^j = \dfrac{\alpha_t^j}{\alpha_t^j + \beta_t^j}$, where $g(\cdot)$ represents the probability density function (pdf).

Given allocation $\mathbf{d}_t$, the (random) outcomes are observed in the next period, captured in the vector $\mathbf{k}_{t+1} \in \mathcal{K} \subseteq \mathbb{Z}^{|J|\times|O|}$, that we define as the *physical state*, as follows:

$$\mathbf{k}_{t+1} = (k_{t+1}^{As}, k_{t+1}^{Af}, k_{t+1}^{Bs}, k_{t+1}^{Bf}).$$

Here, $k_{t+1}^{j,o} \in \mathbb{Z}_+$ represents the number of observed patients in health condition (j, o) at time $t + 1$, where the treatment $j \in J$ is given at time t and the outcome $o \in O$ is observed at time $t + 1$, such that $\sum_{j\in J, o\in O} k_{t+1}^{j,o} = n$.

Given p_t^j, the likelihood of observing k_{t+1}^{js} successes out of d_t^j is binomially distributed, i.e. $Pr(k_{t+1}^{js}|d_t^j, p_t^j) \sim \text{Bin}(k_{t+1}^{js}; d_t^j; p_t^j)$. Since the Beta distribution serves as a conjugate prior for the Binomial distribution, the posterior distribution of p_{t+1}^j is given as follows: $g(p_{t+1}^j) \sim \text{Beta}(\alpha_t^j + k_{t+1}^{js}, \beta_t^j + k_{t+1}^{jf})$.

Let $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \{(\alpha_t^A, \beta_t^A), (\alpha_t^B, \beta_t^B)\}$. If we denote the initial values (at $t = 0$) of Beta distribution hyperparameters by $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B)\}$ and assume that the outcomes of patients in different health conditions are not informative of each other, then each (α_0^j, β_0^j) can be updated independently as follows: $\alpha_t^j = \alpha_0^j + h_t^{j,s}$ and $\beta_t^j = \beta_0^j + h_t^{j,f}$, where $h_t^{j,o}$ captures all the (random) realizations from the past for that treatment–outcome combination. In the absence of any knowledge of treatment efficacy, a commonly assumed initial prior is non-informative, i.e., $(\alpha_0^j, \beta_0^j) = (1, 1)$ for all $j \in J$, equivalent to a uniform [0,1] distribution. Going forward we will use the term "initial priors" to mean "initial values of Beta distribution hyperparameters (at $t = 0$)" and use the two terms interchangeably.[9]

The above definitions directly imply the following: for $t = 1$, $\mathbf{h}_t = \mathbf{k}_t$ and for $t = 2, \cdots, T$, $\mathbf{h}_t = \mathbf{h}_{t-1} + \mathbf{k}_t$.[10]

The entries of the transition matrix at time t, $P_t(\mathbf{h}_{t+1}|\mathbf{h}_t, \mathbf{d}_t, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, representing the probability of transitioning to state $\mathbf{h}_{t+1}$, given $\mathbf{h}_t$, $\mathbf{d}_t$, and $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, is then the product of individual probabilities, defined as follows:

$$\begin{aligned} P_t(\mathbf{h}_{t+1}|\mathbf{h}_t, \mathbf{d}_t, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) &= \prod_{j\in J} Pr(k_{t+1}^{js}|h_t^{js}, h_t^{jf}, d_t^j, \alpha_0^j, \beta_0^j) \\ &= \prod_{j\in J} \int_0^1 Pr(k_{t+1}^{js}|d_t^j, p_t^j) g(p_t^j|\alpha_t^j, \beta_t^j) dp_t^j, \quad (1) \end{aligned}$$

if $d_t^j \in \mathbb{Z}$ and $k_{t+1}^{js} \le d_t^j$ for all $j \in J$, and 0 otherwise.

Finally, the reward function, R_t, depends on the objective function chosen (described below in Sections 3.1.1 and 3.1.2). The entire formulation is a dynamic program, in which the objective is to maximize the expected value function (V_t) that captures expected total reward and solves the Bellman equation as follows:

$$V_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \max_{\mathbf{u}_t} \{R_t + \mathbb{E}_{\mathbf{k}_{t+1}}[V_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})]\}. \quad (2)$$

The optimal policy obtained from (2) serves as a basis for our proposed *Jointly Adaptive*design, described in Section 4.

Below, we define the reward and value functions separately for the two objective functions we consider: patient health and learning.

3.1.1. Objective: patient health

The **patient health** objective **(PH)** aims to maximize the number of patient successes *in* the trial. Following existing literature (e.g., Berry & Eick, 1995; Ning & Huang, 2010), we assign a reward of 1 for success and 0 for failure. Formally, the reward function is defined as: $R_T = 0$ and $R_t = (k_{t+1}^{As} + k_{t+1}^{Bs})$ for $t = 0, 1, \cdots, T - 1$.

Let $\mathcal{S}_t$ denote the value function (V_t) for the patient health objective. The dynamic program in (2) can be expressed as follows.

For the terminal period (T), where no decision needs to be made, $\mathcal{S}_T = 0$. For $T - 1$, the terminal *decision* stage, the optimal strategy is to allocate all patients to the treatment with highest expected success probability as follows:

$$\mathcal{S}_{T-1}(\boldsymbol{\alpha}_{T-1}, \boldsymbol{\beta}_{T-1}) = n \max_j \frac{\alpha_{T-1}^j}{\alpha_{T-1}^j + \beta_{T-1}^j}. \quad (3)$$

For $t = 0, 1, \cdots, T - 2$,

$$\mathcal{S}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \max_{\mathbf{u}_t} \left\{ \sum_{j\in\{A,B\}} \frac{\alpha_t^j}{\alpha_t^j + \beta_t^j} d_t^j + \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})] \right\}. \quad (4)$$

3.1.2. Objective: learning

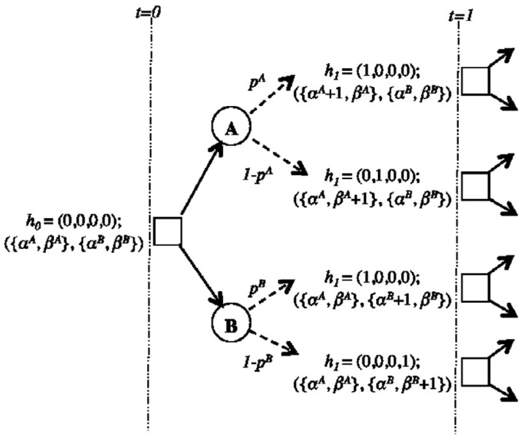
The **learning** objective **(LE)** represents the probability of correctly identifying the most efficacious treatment at the end of the trial. Formally, the reward function is defined as: $R_T = \max\{Pr(p_T^A > p_T^B), Pr(p_T^B > p_T^A)\}$ and $R_t = 0$ for $t = 0, 1, \cdots, T - 1$.

[7] Bin denotes binomial distribution.

[8] If additional information can be used in each patient assignment, for example, the assignments of other patients, then d_t^j could correspond to a random selection of the integers above and below nu_t^j, such that the expectation is nu_t^j.

[9] The initial priors represent clinicians' prior beliefs about each treatment's efficacy.

[10] To illustrate with a simple numerical example, suppose $n = 4$, $t = 5$, $T = 10$, and $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B)\} = \{(1, 1); (1, 1)\}$. Then, a state at $t = 5$ may appear as follows: $\mathbf{h}_5 = (8, 2, 5, 5)$, $\{(\alpha_5^A, \beta_5^A); (\alpha_5^B, \beta_5^B)\} = \{(9, 3); (6, 6)\}$. One solution under the health objective is: $\mathbf{u}_5 = (0.7, 0.3)$, $\mathbf{d}_5 = (3, 1)$, and a potential next state is $\mathbf{k}_6 = (2, 1, 1, 0)$, implying the following: $\mathbf{h}_6 = (10, 3, 6, 5)$ and $\{(\alpha_6^A, \beta_6^A); (\alpha_6^B, \beta_6^B)\} = \{(11, 4); (7, 6)\}$.


t=0
t=1
$h_0 = (0,0,0,0)$;
$(\{\alpha^A, \beta^A\}, \{\alpha^B, \beta^B\})$
A
B
p^A
$1\text{-}p^A$
p^B
$1\text{-}p^B$
$h_1 = (1,0,0,0)$;
$(\{\alpha^A+1, \beta^A\}, \{\alpha^B, \beta^B\})$
$h_1 = (0,1,0,0)$;
$(\{\alpha^A, \beta^A+1\}, \{\alpha^B, \beta^B\})$
$h_1 = (1,0,0,0)$;
$(\{\alpha^A, \beta^A\}, \{\alpha^B+1, \beta^B\})$
$h_1 = (0,0,0,1)$;
$(\{\alpha^A, \beta^A\}, \{\alpha^B, \beta^B+1\})$

t=0		t=1		t=2	
$(h_0^{As},h_0^{Af},h_0^{Bs},h_0^{Bf})$	(d_0^A,d_0^B)	$(h_1^{As},h_1^{Af},h_1^{Bs},h_1^{Bf})$	(d_1^A,d_1^B)	$(h_2^{As},h_2^{Af},h_2^{Bs},h_2^{Bf})$	
(0, 0, 0, 0)	(2,0) (1,1) (0,2)	(0, 2, 0, 0) (1, 1, 0, 0) (2, 0, 0, 0) (0, 1, 0, 1) (0, 1, 1, 0) (1, 0, 0, 1) (1, 0, 1, 0) (0, 0, 0, 2) (0, 0, 1, 1) (0, 0, 2, 0)	(2,0) (1,1) (0,2)	(0, 4, 0, 0) (1, 3, 0, 0) (2, 2, 0, 0) (3, 1, 0, 0) (4, 0, 0, 0) (0, 3, 0, 1) (0, 3, 1, 0) (1, 2, 0, 1) (1, 2, 1, 0) (2, 1, 0, 1) (2, 1, 1, 0) (3, 0, 0, 1) (3, 0, 1, 0) (0, 2, 0, 2) (0, 2, 1, 1) (0, 2, 2, 0) (1, 1, 0, 2) (1, 1, 1, 1)	(1, 1, 2, 0) (2, 0, 0, 2) (2, 0, 1, 1) (2, 0, 2, 0) (0, 1, 0, 3) (0, 1, 1, 2) (0, 1, 2, 1) (0, 1, 3, 0) (1, 0, 0, 3) (1, 0, 1, 2) (1, 0, 2, 1) (1, 0, 3, 0) (0, 0, 0, 4) (0, 0, 1, 3) (0, 0, 2, 2) (0, 0, 3, 1) (0, 0, 4, 0)
Total States = 1		**Total States = 10**		**Total States = 35**	

Fig. 1. State transition diagram for $n = T = 1$ (left) and the enumerated state and decision space for $n = T = 2$ (right), where $J = \{A, B\}$; $O = \{s, f\}$. *Notes*: left figure: squares (□) represent decision points and circles (○) represent random outcomes.

Letting $\mathcal{P}_t$ denote the value function (V_t) for the learning objective, the dynamic program in (2) can be expressed as follows:

$$\mathcal{P}_T = \max\,\{Pr(p_T^A > p_T^B), Pr(p_T^B > p_T^A)\}, \text{ and}$$

for $t = 0, 1, \cdots, T-1$,

$$\mathcal{P}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \max_{\mathbf{u}_t} \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{P}_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})]. \tag{5}$$

Curse of dimensionality. The total number of unique states in this setup that need to be solved equals $\sum_{t=1}^{T}\frac{(nt+\hat{k}-1)!}{nt!\,(\hat{k}-1)!}$ (Bona, 2002), indicating that BAMDP state space increases exponentially with n and T, commonly referred to as the *curse of dimensionality*. Fig. 1 depicts the state space transition diagram for one patient and one period (left) and two patients and two periods (right). A consequence of the curse of dimensionality is the enhanced computational burden necessitating the use of state space approximation methods. While several approximation techniques are possible (and are indeed the focus of ongoing work), in this paper, we use a particular form of approximate dynamic approach to maintain tractability for a practical example. This particular approach, that we term restricted-optimal-policy approximation, utilizes the fully optimal policy for a restricted number of periods early on and a myopic policy for the remainder of the time periods (see Section 5.1.1).

3.2. General properties

Below, we state some properties of our model, starting with the first result that shows that adding the information state preserves the Markovian nature. Unless stated otherwise, all proofs can be found in the Appendix.

Proposition 1. *The BAMDP defined as* $\hat{\mathcal{M}} = \{\mathcal{H}, \mathcal{K}, \mathcal{U}, P_t(\mathbf{h}_{t+1}|\mathbf{h}_t, \mathbf{d}_t), R_t\}$ *is an MDP.*

The proof of the above statement follows the outline of Bertsekas (1995, Section 5.1), which shows that that the addition of the (imperfect) information state reduces the problem to one with perfect state information and preserves the Markovian dynamics, i.e., $P(\mathbf{h}_{t+1}; \mathbf{k}_{t+1}|\mathbf{h}_1, .., \mathbf{h}_t; \mathbf{k}_1, .., \mathbf{k}_t; \mathbf{u}_1, .., \mathbf{u}_t) = P(\mathbf{h}_{t+1}; \mathbf{k}_{t+1}|\mathbf{h}_t; \mathbf{k}_t; \mathbf{u}_t)$. We omit the proof and refer the readers to Bertsekas (1995).

The following result shows that the success probability is a nondecreasing (nonincreasing) function of α_t^j (β_t^j).

Lemma 1. *Let* $Pr^{\alpha_t^j,\beta_t^j} = Pr(p_t^j > y|\alpha_t^j, \beta_t^j)$ *represent the probability of success with treatment* $j \in J$ *at time* t *such that* $0 \le y < 1$. *Then the following stochastic order prevails:* $Pr^{\alpha_t^j+1,\beta_t^j} \ge Pr^{\alpha_t^j,\beta_t^j} \ge Pr^{\alpha_t^j,\beta_t^j+1}$.

The following corollary is a direct consequence of the above lemma.

Corollary 1. *Let* $p_t^j = Pr(s|j, \alpha_t^j, \beta_t^j)$, $p_t^{j+} = Pr(s|j, \alpha_t^j + 1, \beta_t^j)$, *and* $p_t^{j-} = Pr(s|j, \alpha_t^j, \beta_t^j + 1)$. *Then, the following stochastic order prevails:* $p_t^{j+} \ge p_t^j \ge p_t^{j-}$.

The following two results show how the value function changes if we observe an additional success or failure with a treatment. We use the following notation: $\alpha_t^{j+} = \alpha_t^j + 1$, $\boldsymbol{\alpha}_t = (\alpha_t^j, \alpha_t^{j'})$, $\boldsymbol{\alpha}_t^+ = (\alpha_t^{j+}, \alpha_t^{j'})$; similar definitions hold for β_t^{j+}, $\boldsymbol{\beta}_t$, and $\boldsymbol{\beta}_t^+$.

Proposition 2. $\mathcal{S}_t(\boldsymbol{\alpha}_t^+, \boldsymbol{\beta}_t) \ge \mathcal{S}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) \ge \mathcal{S}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t^+)$ *for* $j, j' \in \{A, B\}$.

Proposition 3. *If* $\mathbb{E}p_t^j \ge \mathbb{E}p_t^{j'}$, $\mathcal{P}_t(\boldsymbol{\alpha}_t^+, \boldsymbol{\beta}_t) \ge \mathcal{P}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) \ge \mathcal{P}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t^+)$ *for* $j, j' \in \{A, B\}$.

4. Comparison of adaptive designs

We compare the *Jointly Adaptive* design modeled in Section 3.1 with the existing sequential response-adaptive design that randomizes patients one at a time (e.g., Berry & Eick, 1995). While such a design, that we call *Perfectly Adaptive*, offers the greatest learning potential, its applicability in practice is limited. Implementing this design would result in a prohibitively long clinical trial and potentially deteriorating outcomes for the general patient population due to long approval times.

We also compare the *Jointly Adaptive* design with two naive (suboptimal) implementable versions of the *Perfectly Adaptive* design. In the first design, that we call *Isolated Adaptive*, each patient is considered in isolation at each time period. Such a design is akin to having multiple independent clinical trials in isolation, each of which implements a *Perfectly Adaptive* design. The information set is reduced by a factor of n, implying reduced opportunities for learning and inferior outcomes.

Another approach is to impose a constraint such that all patients are allocated to a single treatment at each time period. Such a design, that we call *Restricted Adaptive*, incorporates responses from all past patients similar to the *Jointly Adaptive* design. However, the design is constrained by the fact that all patients in a time period receive the same treatment.

Finally, we benchmark against the traditional fixed design, primarily used to learn about treatment efficacy. We define a new

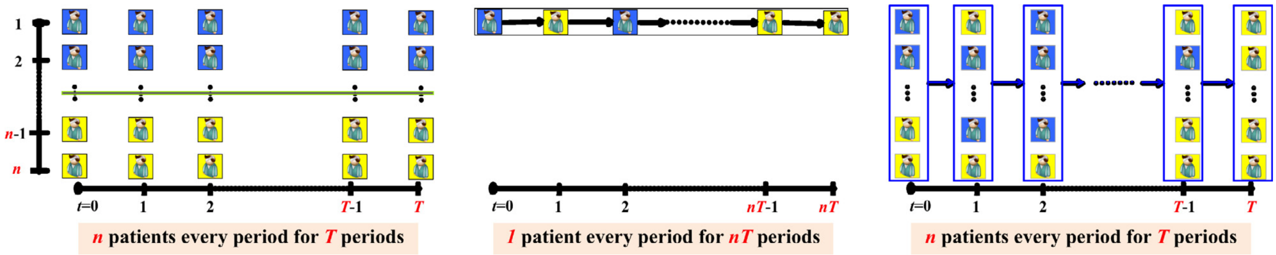

1
2
n-1
n
t=0
1
2
T-1
T
n patients every period for T periods
t=0
1
2
nT-1
nT
1 patient every period for nT periods
t=0
1
2
T-1
T
n patients every period for T periods

Fig. 2. A pattern of learning and patient allocation under various designs. *Notes*: the three designs, from left to right are Π_{EA} *(Equal Allocation)*, Π_{PA} *(Perfectly Adaptive)*, and Π_{JA} *(Jointly Adaptive)*.

design called *Equal Allocation* (EA), where patients are allocated to treatments in equal proportion.[11] We summarize the designs below.

- Π_{PA} *(Perfectly Adaptive)*:= 1 × nT patients arrive sequentially; decision is made for one patient at a time; incorporates learning from outcomes of all previous patients.
- Π_{JA} *(Jointly Adaptive)*:= n patients per period × T periods; decision is made for all n patients simultaneously; incorporates learning from outcomes of all previous patients; each patient is randomized to all the available treatments.
- Π_{RA} *(Restricted Adaptive)*:= n patients per period × T periods; decision is made for all n patients simultaneously; incorporates learning from the outcomes of all previous patients; all n patients receive the same treatment at each period.
- Π_{IA} *(Isolated Adaptive)*:= n patients per period × T periods; each patient is considered in isolation; equivalent to n independent T period sequential trials with no learning across trials.
- Π_{EA} *(Equal Allocation)*:= n patients per period × T periods; $\frac{n}{|J|}$ patients allocated to each treatment.

Note that Π_i indicates a class of policies, and $\pi_i \in \Pi_i$, $i = \{PA, JA, RA, IA, EA\}$ indicates the optimal policy that belongs with that class of policy. Fig. 2 visually illustrates the differences between the three designs: Π_{EA}, Π_{PA}, and Π_{JA}. Below, we list some structural properties of the *Jointly Adaptive* design and provide numerical comparisons in Section 5.

Theorem 1. *Given n, T, N,* $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, *the following holds: (i)* $V_0^{\pi_{PA}} \geq V_0^{\pi_{JA}}$, *(ii)* $V_0^{\pi_{JA}} \geq V_0^{\pi_{IA}}$, *(iii)* $V_0^{\pi_{JA}} \geq V_0^{\pi_{RA}}$, *and (iv)* $V_0^{\pi_{JA}} \geq V_0^{\pi_{EA}}$.

Note that when $n = 1$, $V_0^{\pi_{PA}} = V_0^{\pi_{JA}} = V_0^{\pi_{RA}} = V_0^{\pi_{IA}}$.

Remark. The initial values of Beta distribution hyperparameters or initial priors can play a significant role. A strong initial prior, whose values are numerically large and not likely to be affected significantly with the observed information from trial. Conversely, a weak initial prior will be heavily influenced by the information obtained during the trial.

Asymptotic properties

The following result shows that under any policy that belongs to the class of optimally adaptive policies, each treatment is applied infinitely often (i.o.) in the limit.

Lemma 2. *Let* $\bar{p}^j$ *represent the (unknown) underlying probabilities of success with treatment (equivalently arm) j, such that* $0 < \bar{p}^j < 1$, $\bar{p}^j \neq \bar{p}^{j'}$, *and* $j, j' \in J$. *Then, for any optimal policy* $\pi \in \Pi$, $\sum_{t=0}^{T-1} d_t^{j,\pi} \xrightarrow{a.s.} \infty$ *w.p.* 1 *as* $T \to \infty$ *or* $n \to \infty$.

The following result, consistent with Ghosal, Ghosh, and Samanta (1995, Proposition 1), shows that the optimal design, that tries each treatment infinitely often, identifies the better treatment w.p.1.

Lemma 3. *Suppose for every* $j \in J$, $\sum_{t=1}^{T} d_t^j \xrightarrow[T\to\infty]{a.s.} \infty$. *Then, for* $n < \infty$, $Pr\{p_T^j > \max_{j' \in J\setminus\{j\}} p_T^{j'}\} \xrightarrow[T\to\infty]{P} 1$, *and for* $T < \infty$, $Pr\{p_T^j > \max_{j' \in J\setminus\{j\}} p_{T,}^{j'}\} \xrightarrow[n\to\infty]{P} 1$ *for all* $j, j' \in J$.

The following result shows that the *Jointly Adaptive* design infers the "superior" treatment *w.p.* 1 in the limit.

Theorem 2. $\mathcal{P}_0^{\pi_{JA}} \xrightarrow{P} 1$ *as* $T \to \infty$ *or* $n \to \infty$.

5. Results and analysis

We perform numerical analyses to demonstrate the value of implementing the *Jointly Adaptive* design under multiple scenarios that vary in n, T, N, and combinations of initial values of Beta distribution hyperparameters or initial priors. We use 13 different values of initial priors, the same set as in Berry (1978): {(4,1); (6,2); (1,$\frac{1}{2}$); (2,1); ($\frac{1}{2}$,$\frac{1}{2}$); (1,1); (2,2); (4,4); (6,6); (1,2); ($\frac{1}{2}$,1); (2,6); (1,4)}, resulting in 91 unique combinations of $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B)\}$. We note that $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B)\} = \{(1, 1); (1, 1)\}$, equivalently a uniform [0,1] distribution on p_0^A and p_0^B and henceforth referred to as "noninformative initial priors", represents the case when clinician's beliefs at the beginning of the trial are such that each treatment is equally likely to succeed. Unless otherwise stated, all results are based on solving a fully enumerated problem using backward recursion (Puterman, 1994).

Further, for the health objective, we compare the *Jointly Adaptive* design with the heuristic policies described below. Such heuristics use learning only from past patients, in contrast to forward-looking adaptive designs, described in §4.[12] The first two heuristics are used in Bertsimas and Mersereau (2007).

- *Greedy* (π_{Gr}): This algorithm, analogous to the play-the-winner rule of Robbins (1952), allocates all the patients at each period to the treatment with the highest expected probability of success, given current information. In case of a tie, *Greedy* divides the patients equally among the treatments with the highest expected success probability.
- *GGreedy* (π_{GG}): This algorithm allocates all the patients in each period to the treatment with the highest Gittins index, given current information, computed using an infinite horizon and a discount rate of 0.90. Our procedure for computing the indices for this problem follows the discussion in the first chapter of Gittins (1989). In case of a tie, *GGreedy* divides the patients equally among the treatments.
- *BK* (π_{BK}): This algorithm, developed in Burnetas and Katehakis (1996), that we named after the authors' last initials, allocates all the patients at each period to the treatment with the highest value of an index that we call *BK*, i.e., $j_t^* = \arg\max_j \{BK_t^j\}$. We calculate BK_t^j at each decision point by solving an optimization program as

[11] The majority of fixed designs randomize patients equally to the various treatments throughout the trial.

[12] The calculation for optimal value function is still obtained by backward recursion.

Table 1
Expected proportion of successes for a variety of problem scenarios.

					FIXED	ADAPTIVE				HEURISTICS			
(α^A, β^A)	(α^B, β^B)	n	T	N	Equal (π_{EA})	Perfectly (π_{PA})	Jointly (π_{JA})	Restricted (π_{RA})	Isolated (π_{IA})	Greedy (π_{Gr})	GGreedy (π_{GG})	UCB1 (π_{UC})	BK (π_{BK})
(1,1)	(1,1)	2	12	24	0.5000	0.6259	0.6215	0.6205	0.6077	0.6202	0.6202	0.6162	0.6123
(1,1)	(1,1)	4	6	24	0.5000	0.6259	0.6132	0.6084	0.5847	0.6127	0.6127	0.6096	0.6049
(1,1)	(1,1)	4	12	48	0.5000	0.6393	0.6333	0.6299	0.6077	0.6321	0.6321	0.6250	0.6232
(1,1)	(1,1)	4	24	96	0.5000	0.6487	0.6460	0.6439	0.6259	0.6433	0.6433	0.6328	0.6366
(2,1)	(1,4)	2	12	24	0.6667	0.6679	0.6679	0.6678	0.6670	0.6676	0.6280	0.6240	0.6630
(2,1)	(1,4)	2	24	48	0.6667	0.6693	0.6693	0.6692	0.6679	0.6685	0.6481	0.6385	0.6631
(2,1)	(1,4)	4	12	48	0.6667	0.6693	0.6691	0.6690	0.6670	0.6686	0.6410	0.6339	0.6611
(2,1)	(1,4)	4	24	96	0.6667	0.6709	0.6707	0.6705	0.6679	0.6695	0.6564	0.6452	0.6630
(1,4)	(1,4)	4	6	24	0.2000	0.2532	0.2495	0.2478	0.2297	0.2482	0.2482	0.2476	0.2414
(1,4)	(1,4)	4	12	48	0.2000	0.2632	0.2614	0.2605	0.2417	0.2603	0.2603	0.2574	0.2531
(4,4)	(4,4)	4	6	24	0.5000	0.5538	0.5480	0.5470	0.5304	0.5479	0.5479	0.5434	0.5438
(4,4)	(4,4)	4	12	48	0.5000	0.5644	0.5614	0.5607	0.5421	0.5608	0.5608	0.5502	0.5547
(4,1)	(1,4)	4	12	48	0.8000	0.8001	0.8001	0.8000	0.8000	0.8000	0.7642	0.7593	0.7958
(4,1)	(4,1)	4	24	96	0.8000	0.8744	0.8724	0.8719	0.8588	0.8709	0.8709	0.8612	0.8676
(1,0.5)	(1,2)	2	24	48	0.6667	0.6984	0.6980	0.6979	0.6920	0.6955	0.6854	0.6773	0.6898
(0.5,0.5)	(6,6)	4	24	96	0.5000	0.6516	0.6499	0.6497	0.6349	0.6444	0.6402	0.6314	0.6416

follows:

$$BK_t^j = \max\left\{ \underline{p}_t^j : \mathbf{I}(\hat{p}_t^j, \underline{p}_t^j) \leq \frac{\log \bar{N}_t}{\alpha_t^j + \beta_t^j} \right\},$$

where $\bar{N}_t = \sum_{j \in J}(\alpha_t^j + \beta_t^j)$, $\hat{p}_t^j = \dfrac{\alpha_t^j}{\alpha_t^j + \beta_t^j}$, and $\mathbf{I}(x, y) = x\log\dfrac{x}{y} + (1-x)\log\dfrac{(1-x)}{(1-y)}$, for x and y Bernoulli distributed. *BK* is thus an adjusted version of the estimator $\hat{p}_t^j$, in which the adjustment decreases with the number of patients already allocated to the treatment j (and incorporates prior information). In case of a tie, *BK* divides the patients equally among the treatments. For an exact specification of the algorithm, see the original paper.

- *UCB1* (π_{UC}): This algorithm, developed in Auer et al. (2002), allocates all the patients at each period to the treatment with the highest value of upper confidence index, i.e., $j_t^* = \arg\max_j \{UCB1_t^j\}$, calculated at each decision point as follows: $UCB1_t^j = h_t^j + \sqrt{\dfrac{2\log nt}{\sum_{t=0}^{t} d_t^j}}$. In case of a tie, *UCB1* divides the patients equally among the treatments. For further details, see the original paper.

5.1. Objective: patient health

We use this numerical study to investigate the advantage of implementing the *Jointly Adaptive* design relative to other implementable designs: (a) the *Equal Allocation* design, (b) naive adaptive designs*: Restricted* and *Isolated Adaptive* designs, and (c) heuristics*: Greedy*, *GGreedy, BK,* and *UCB1*. As a purely theoretical exercise, we also show the disadvantage of the *Jointly Adaptive* design relative to the *Perfectly Adaptive* design, a non-implementable design, unless all treatment outcomes are immediately observable.

Table 1 lists the expected proportion of successes for various designs. The results clearly show that implementing adaptive designs and heuristics improve expected patient successes compared to the fixed design, although the magnitude of the improvement varies. We also note that*Greedy* and all adaptive designs perform better than the fixed design but other heuristics do not always do so under these scenarios.

Further, the *Jointly Adaptive* design performs best among all implementable designs. In particular, the *Jointly Adaptive* design performs substantially better than the *Equal Allocation* design, and almost as well as the non-implementable *Perfectly Adaptive* design. For example π_{JA} increases patient successes by 29.2 percent compared to π_{EA} when $(n, N) = (4, 48)$ and initial priors are noninformative. While none of the three heuristics is optimal, neither is any a clear winner. On average, π_{Gr} actually performs best among the heuristic policies, although the gain is decreasing in N, suggesting that other heuristics may be more beneficial for larger problem sizes.

Table 2 lists various quantities ($\delta' s$) that capture the magnitude of the difference (expressed as a percentage gain) between various designs. The table also lists the average and maximum values across 91 combinations of initial priors under the considered scenarios.

Fig. 3 illustrates the variation of δ_{JI}, δ_{JR}, and δ_{PJ} with (a) n, keeping T fixed at 12 time periods (left chart), (a) T, keeping n fixed at 4 patients (middle chart), and (c) $n \times T$, keeping N fixed at 48 total patients (right chart). Some key observations from the figure are, first, the gains in expected patient successes that π_{JA} provides over π_{RA} and π_{IA} are increasing in n, which can be attributed to the fact that additional patients provide additional learning opportunities. Second, all three quantities are nonincreasing in T, essentially due to the fact that the bulk of the learning happens earlier in the trial and additional periods provide diminishing opportunities to learn. Third, the effect of n dominates that of T. The results demonstrate the potential for enhanced patient outcomes by implementing the *Jointly Adaptive* design, especially considering that a typical clinical trial consists of a large number of patients.

5.1.1. Restricted-optimal-policy approximation

As described earlier, the adaptive design setup suffers from the curse of dimensionality, implying that, as the trial size increases, solving the fully enumerated problem becomes computationally burdensome. While we are working on a state space approximation technique to addresses this challenge, in this paper, we propose a restricted-optimal-policy approximation method, a form of approximate dynamic programming approach, that is useful for solving large problems with relatively short observation delays (i.e., relatively small n and large T), such as the one described in Section 5.5.

In this approach, we fully enumerate all states, actions, and transitions for initial (restricted) number of periods and then assume that a myopic policy is followed in the subsequent periods. More specifically, we implement the optimal *Jointly Adaptive* design over the initial shorter horizon that we term as T_{short} and, thereafter, use a myopic design. The expected number of successes is a combination of the optimal solution from the first T_{short} periods and a myopic

Table 2
Definition of quantities comparing various designs and their values (average, maximum) across 91 initial priors when $(n, N) = (4, 24)$.

Quantity	Definition	Description	Average (percent)	Maximum (percent)
δ_{PJ}	$\frac{\mathcal{S}^{\pi_{PA}} - \mathcal{S}^{\pi_{JA}}}{\mathcal{S}^{\pi_{JA}}}$	*Perfectly* vs. *Jointly Adaptive*	0.85	2.66
δ_{JI}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{IA}}}{\mathcal{S}^{\pi_{IA}}}$	*Jointly* vs. *Isolated Adaptive*	2.54	8.64
δ_{JR}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{RA}}}{\mathcal{S}^{\pi_{RA}}}$	*Jointly* vs. *Restricted Adaptive*	0.26	2.61
δ_{Gr}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{Gr}}}{\mathcal{S}^{\pi_{Gr}}}$	*Jointly Adaptive* vs. *Greedy*	0.48	6.03
δ_{BK}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{BK}}}{\mathcal{S}^{\pi_{BK}}}$	*Jointly Adaptive* vs. *BK*	2.13	8.78
δ_{GG}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{GG}}}{\mathcal{S}^{\pi_{GG}}}$	*Jointly Adaptive* vs. *GGreedy*	3.61	10.66
δ_{UC}	$\frac{\mathcal{S}^{\pi_{JA}} - \mathcal{S}^{\pi_{UC}}}{\mathcal{S}^{\pi_{UC}}}$	*Jointly Adaptive* vs. *UCB1*	4.05	10.61

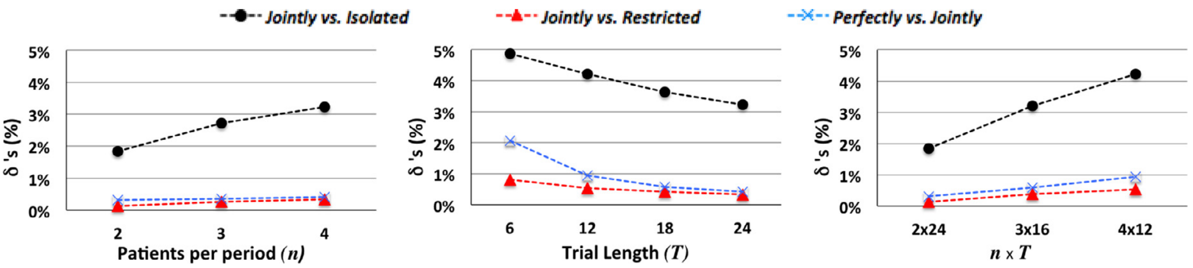


Fig. 3. δ_{JI}, δ_{JR}, and δ_{PJ} as a function of (left) n (with fixed $T = 12$), (middle) T (with fixed $n = 4$) and (right) $n \times T$ (with fixed $N = 48$) with noninformative initial priors.

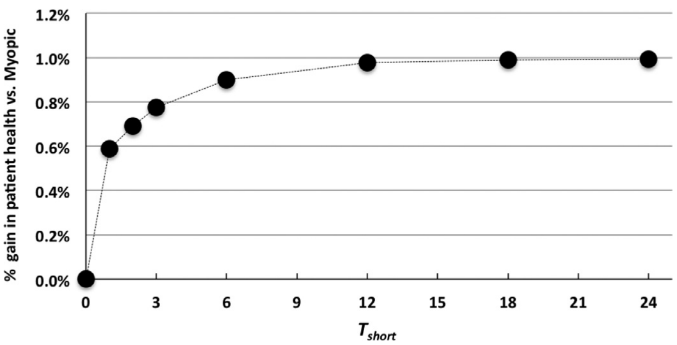


Fig. 4. Average gain in expected successes over a purely myopic approach as a function of T_{short} in restricted-optimal-policy approximation, $(n, T) = (4, 24)$.

solution from the remaining $(T - T_{\text{short}})$ periods. Any solution thus obtained provides a lower bound on the optimal solution. We note that this approximation results in little loss relative to a fully optimal solution since most of the uncertainty is resolved in the states attained in which the myopic policy is used. Our choice of a myopic design for the restricted-optimal-policy approximation approach is the *Greedy* heuristic, described earlier, since it performs best among the heuristics we considered.

We illustrate the usefulness of the restricted-optimal-policy approximation using a numerical example, where $(n, T) = (4, 24)$ or equivalently $(n, N) = (4, 96)$ and T_{short} ranges from 0 to 24. Note that $T_{\text{short}} = 0$ refers to a fully *Greedy* design and $T_{\text{short}} = 24$ refers to the optimal *Jointly Adaptive* design. Fig. 4 shows the average gain in expected successes as a result of implementing the restricted-optimal-policy approximation over a purely *Greedy* approach, where we note that this gain is increasing in time. Further, the gain increases sharply early on but then plateaus afterward, a result of the multi-armed bandit property that the restricted-optimal-policy approximation exploits.

5.2. Objective: learning about treatments

As mentioned earlier, a fixed design is primarily focused on learning and hence any comparison should be benchmarked against it. In this section, we present numerical results comparing π_{JA} vs. π_{EA} with the following quantity capturing the difference between the two:

$$\delta_{Pr} = \frac{\mathcal{P}_0^{\pi_{JA}} - \mathcal{P}_0^{\pi_{EA}}}{\mathcal{P}_0^{\pi_{EA}}}.$$

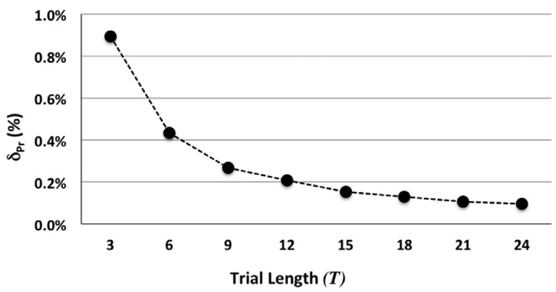


Fig. 5. δ_{Pr} as a function of T with $n = 4$ and noninformative initial priors.

Fig. 5 plots δ_{Pr} as a function of T when $n = 4$ and initial priors are noninformative. We observe that δ_{Pr} is relatively small and quickly decreases to zero as T increases. We observe a maximum gain of 2.2 percent under the considered scenarios. Finally, we note that $\mathcal{P}_0^{\pi_{JA}}$ increases in both n and T.

Managerial insight. We have shown examples that highlight the extent of the objective improvement from using π_{JA} over π_{EA} for varying parameter combinations. These observations also imply that π_{JA} would use the same or fewer patients compared to π_{EA} to achieve a target objective function value, as explained below.

Let $y \in (0, 1)$ be the target expected learning objective. Given n, let $\mathbb{E}T^{\pi_{EA}} = \{\min T : \mathcal{P}_0^{\pi_{EA}} \geq y\}$ and $\mathbb{E}T^{\pi_{JA}} = \{\min T : \mathcal{P}_0^{\pi_{JA}} \geq y\}$ denote the expected time required to achieve the target objective value under π_{EA} and π_{JA}, respectively, where we note that each calculation of $P_0^{\pi_{JA}}$ is done with a fixed T. Then, $\frac{\mathbb{E}T^{\pi_{EA}} - \mathbb{E}T^{\pi_{JA}}}{\mathbb{E}T^{\pi_{EA}}}$ represents the reduction in expected number of patients. To illustrate with an example, for $y = 0.93$, $n = 4$, and noninformative initial priors, $\mathbb{E}T^{\pi_{EA}} = 20$, and $\mathbb{E}T^{\pi_{JA}} = 19$. In other words, the *Jointly Adaptive* design would use 5 percent fewer patients in expectation to achieve a 93 percent

Table 3
Cross-objective impact as a function of N, with $n = 4$ and noninformative initial priors.

N	$\mathcal{P}_0^{\pi_{JA}}(LE)$	$\mathcal{P}_0^{\pi_{JA}}(PH)$	$\mathcal{S}_0^{\pi_{JA}}(PH)$	$\mathcal{S}_0^{\pi_{JA}}(LE)$	δ_{Lh} (percent)	δ_{Hl} (percent)
24	0.8797	0.8363	0.6132	0.5105	4.9	16.8
36	0.9001	0.8567	0.6261	0.5177	4.8	17.3
48	0.9128	0.8714	0.6333	0.5200	4.5	17.9
60	0.9215	0.8816	0.6381	0.5135	4.3	19.5
72	0.9281	0.8895	0.6415	0.5095	4.2	20.6

learning objective target compared to a fixed design.[13] These savings do not include the 28 percent gain in expected patient successes.

5.3. Cross-objective impact

We numerically evaluate the impact on the alternative objective function value of maximizing a chosen objective. Specifically, we first calculate the change in expected proportion of successes when the objective function being maximized changes from patient health to learning and vice-versa. The following quantities capture this tradeoff:

$$\delta_{Lh} = \frac{\mathcal{P}_0^{\pi_{JA}}(LE) - \mathcal{P}_0^{\pi_{JA}}(PH)}{\mathcal{P}_0^{\pi_{JA}}(LE)};\ \delta_{Hl} = \frac{\mathcal{S}_0^{\pi_{JA}}(PH) - \mathcal{S}_0^{\pi_{JA}}(LE)}{\mathcal{S}_0^{\pi_{JA}}(PH)}.$$

Here, $\mathcal{P}_0^{\pi_{JA}}(LE)$ and $\mathcal{P}_0^{\pi_{JA}}(PH)$ refer to the value of expected learning in the trial under the objectives of learning (LE) and patient health (PH), respectively. $\mathcal{S}_0^{\pi_{JA}}(LE)$ and $\mathcal{S}_0^{\pi_{JA}}(PH)$ are defined similarly for the value of expected patient health under the two objectives. We note that $\mathcal{P}_0^{\pi_{JA}}(LE)$ and $\mathcal{S}_0^{\pi_{JA}}(PH)$ are the *optimal* values for learning and patient health, respectively.

Table 3 lists the quantities defined above as a function of N when $n = 4$ and initial priors are noninformative. We observe that δ_{Lh} is small compared to δ_{Hl}, highlighting a key feature of adaptive design: patients are treated as effectively as possible without a significant loss in learning. In addition, we observe that (a) as N increases, δ_{Lh} decreases while δ_{Hl} increases, (b) this rate of change in δ_{Lh} is smaller compared to that of δ_{Hl}. Together, this further highlights the benefits of implementing the optimal adaptive design.

We further illustrate the tradeoff between the patient health (PH) and learning (LE) objectives using a numerical example below, where we calculate the expected learning and expected proportion of successes in a trial when the objective function is a convex combination of the two objectives (PH *and* LE).

Let $CC = \alpha PH + (1 - \alpha)LE$ denote the convex combination of the two objectives, where α and $(1 - \alpha)$ denote the weights on patient health (PH) and learning (LE) objectives, respectively. Note that $\alpha = 0$ and $\alpha = 1$ refer to the cases of pure learning and pure patient health objectives, respectively. We calculate the expected proportion of successes in the trial, $\mathcal{S}_0^{\pi_{JA}}(CC)$, and the expected probability of correctly identifying the most efficacious treatment at the end of the trial, $\mathcal{P}_0^{\pi_{JA}}(CC)$, under this combined objective. Fig. 6 depicts the tradeoff between the two objectives when $n = 4$ and $N \in \{24, 96\}$, with noninformative initial priors. We observe that even a small shift from pure learning to the patient health objective (represented by movement from $\alpha = 0$ to $\alpha = 0.1$ on the curve) results in a substantial gain in the expected patient successes in the trial, while only causing a negligible loss in the expected learning. However, increasing the weight on the patient health objective (α) further, while increasing expected patient successes in the trial, requires an increasingly larger sacrifice on the expected learning from the trial. The result reinforces the benefits from implementing the optimal adaptive design and giving consideration to the health of the patients in the trial. The figure is consistent with Hardwick and Stout (2008, Fig. 1), which analyzes a two-armed bandit problem under a combination of two similar objectives and where they use a modification of an ideal multiarmed bandit design.

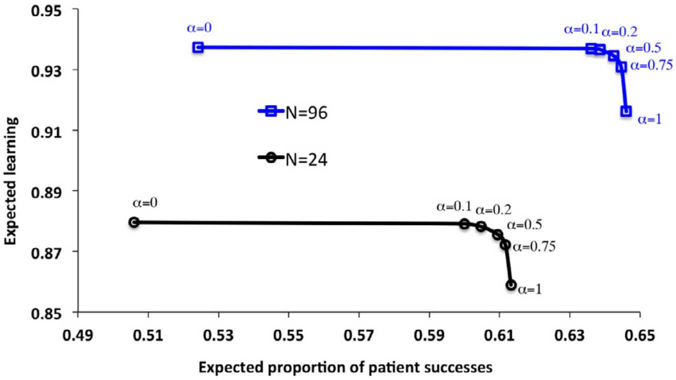


Fig. 6. Tradeoff between competing objectives (learning vs. patient health) for $N = 24$ and $N = 96$ with $n = 4$ and noninformative initial priors.

5.4. Numerically estimating the clinician's rate of learning

We numerically attempt to estimate the rate at which the clinician's initial beliefs about the success probabilities of treatments converge to the underlying probability values, where we emphasize that the underlying probabilities are not known to the researcher. Our estimation procedure uses simulation as follows. Starting with initial values of Beta distribution hyperparameters or initial priors, the random patient outcomes are generated according to the underlying Binomial distribution, which are then used to update the priors. We define the convergence rate as the rate at which the expected values of the success probabilities derived from the posteriors at each time period converge to the underlying probabilities, captured in the following quantity:

$$\epsilon_t = \frac{1}{|J|} \sum_{j \in J} |\bar{p}^j - \mathbb{E} p_t^j|,$$

where ϵ_t is the *Absolute Error* at time t, $\bar{p}_t^j$, represents the underlying success probability of treatment $j \in \{A, B\}$ and $\mathbb{E}p_t^j$ is the current expectation conditioned on the information state, as described earlier. The convergence rate is then the decay of ϵ_t with t.

We conduct our numerical simulations using various combinations of initial priors and trial length $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B); T\}$; for each such combination, we calculate ϵ_t vs. t for 91 pairs of $(\bar{p}_t^A, \bar{p}_t^B)$. Fig. 7 plots three such combinations, where we plot the average value of ϵ_t across the 91 combinations. We observe the convergence rate follows a similar pattern and is approximately proportional to $t^{-\frac{1}{2}}$, consistent with Ghosal (2010, chap. 2).

5.5. Application to a clinical trial

Trial background and description. We implement the *Jointly Adaptive* design ex-post on a recently conducted stent study, the *Stenting and*

[13] This implies the *Jointly Adaptive* design has the potential to lower the cost of conducting the clinical trial, given that patient and time costs are key contributors to a high cost of clinical trial (see Section 1).

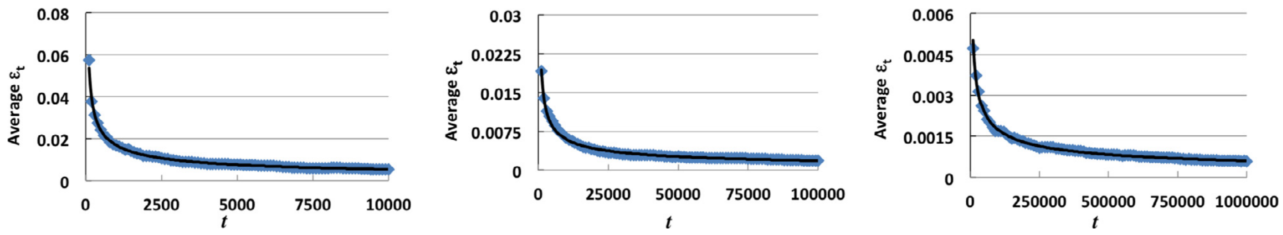


Fig. 7. Average ϵ_t vs. t for the following $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B); T\}$ combinations: $\{(1, 1), (1, 1); 10^4\}$ (left), $\{(1, 4), (4, 1); 10^5\}$ (middle), and $\{(6, 2), (1, \frac{1}{2}); 10^6\}$ (right). *Notes*: the average ϵ_t is across 91 combinations of $(\bar{p}_t^A, \bar{p}_t^B)$; the solid lines on each of the graphs are the trend line fitted to the data points.

Aggressive Medical Management for Preventing Recurrent Stroke in Intracranial Stenosis (SAMMPRIS) trial. The trial, that lasted for approximately two and a half years, evaluated whether adding *Percutaneous Transluminal Angioplasty and Stenting* (PTAS) to the standard treatment improves patient outcomes when treating Intracranial Arterial Stenosis. The trial compared the efficacy of two treatments in preventing the primary endpoint.[14] A treatment was considered a *failure* if a primary endpoint was observed on a patient, and a *success* otherwise.

A total of 451 patients were randomized, approximately equally, to the two arms as follows: (a) aggressive medical management alone (227 patients) and (b) PTAS plus aggressive medical management (224 patients). The trial concluded that adding PTAS to the standard treatment provides no benefit. More critically, it resulted in much worse expected patient outcomes, based on 33 failures in the PTAS group vs. 13 failures in the medical-management group, for a total of 46 failures. To support our assumption of short observation delays, all 33 failures in the PTAS group occurred within a week while 25 failures occurred within 1 day of administering the procedure. Additional details can be found in Chimowitz et al. (2011a) and Chimowitz et al. (2011b).

The SAMMPRIS trial provides an ideal setting for applying and testing our model for several reasons. First, it employed the traditional fixed design, thus offering a good basis for comparison. Second, the trial parameters make it computationally feasible to implement our design with a restricted-optimal-policy approximation: $|J| = |O| = 2$ and an average of $n = 4$ patients enrolled and received treatment every week, implying $t = 1$ week (delay in observing outcomes).[15]

Implementation. First, we choose initial priors for both treatments. For PTAS treatment, our choice of priors yield the expected failure probability of 4.44 percent at the beginning of the trial. This is equal to the failure rate *observed* in the previous 45-patient trial on the same stent.[16] For the standard treatment, we choose strong initial priors that result in an expected failure probability of 5.73 percent, as *observed* in the SAMMPRIS trial, where we assume that the failure probability is known with a high degree of certainty and any observations from the trial have negligible impact on this probability.

Our goal was to calculate the expected number of failures in the SAMMPRIS trial with the *Jointly Adaptive* design. We do this by first solving for the optimal policy based on our choice of initial priors and subsequently implementing this policy using the SAMMPRIS conditions with fixed failure probabilities. In other words, clinicians start the trial with the assumption that PTAS is better than standard treatment; however, patients' randomization to treatments is based on π_{JA} instead of π_{EA}. Our calculations assume that the failure rates from the SAMMPRIS trial revealed the underlying probabilities of success with each treatment.

We could directly solve for fully optimal policies for problem size up to $N = 240$ (equivalently $T = 60$) but memory limitations precluded direct solutions of larger problems.[17] Hence, we employed a restricted-optimal-policy approximation approach that results in minimal deviations from optimality (see Section Section 5.1.1). We find that implementing this design with $N = 451$ and $T_{\text{short}} = 60$ (equivalently 240 observations) results in 28.8 expected failures, a reduction of over 37 percent in expectation.

Table 4 lists the total expected failures, standard error and the number of patients allocated to PTAS group as a function of n. Further, the total expected failures decrease with T_{short}, as would be expected. We extrapolate from the difference between the results on expected number of failures with $T_{\text{short}} = 45$ and those with $T_{\text{short}} = 60$, and estimate that a fully optimal design for $N = 451$ would result in no fewer than 28.6 expected failures.[18] This means that the restricted-optimal-policy approximation solution objective would be within at most 0.73 percent of the full optimality.[19]

Given $N = 451$, clinicians have the opportunity to update their beliefs (and actions) 112 times during the trial.[20] It may not be possible to update so often, for example, due to logistical reasons. Fig. 8 (left chart) shows the total expected failures with π_{JA} as a function of the number of updates, where we consider four cases: (a) 112 updates or update every week, (b) 28 updates or update every 4 weeks, (c) 14 updates or update every 8 weeks, and (d) 7 updates or update every 16 weeks. As expected, increasing the number of updates results in a more frequent use of the optimal policy and hence a reduction in failure rate. Fig. 8 (right chart) shows the probability distribution of failures with π_{JA} when $n = 4$, where we grouped the failures in buckets of 1. Note that the chance of 46 or more failures (as occurred in the original trial) is negligibly small at 0.00088.

[14] The primary endpoint was defined as the following (negative) outcome: any stroke or death within 30 days after enrollment or after any revascularization procedure of the qualifying lesion during follow-up, or stroke in the territory of the symptomatic intracranial artery beyond 30 days.

[15] This stent trial also generated controversy in the mainstream media, for example, in a 2011 *New York Times* article (NYTimes, 2011), providing an additional motivation to choose this trial for our study. The controversy can be attributed to the fact that results from the SAMMPRIS trial were in direct contrast to the results from an earlier smaller trial on the same stent that led the FDA to approve the use of the stent in 2005 under the Humanitarian Device Exemption (HDE) program. The earlier smaller single-armed trial, which enrolled 45 patients and tested only the PTAS treatment for efficacy, resulted in a much lower failure probability of 4.44 percent (FDA, 2005).

[16] In doing so, we assume that this smaller trial itself started with a noninformative initial prior and that the outcomes were an unbiased sample, as evidently was assumed by the FDA when approving the stent under the HDE program.

[17] We ran our code on a personal Macintosh computer (4 gigabyte 667 hertz DDR2 SSDRAM, 250 gigabyte HD) as well as on the research grid at Chicago Booth.

[18] Extrapolation allows us to consider the worst-case scenario, although we expect the number of failures to be less than 28.6 given that the expected number of failures are convex, decreasing in T_{short}.

[19] While the restricted-optimal-policy approximation performs well in expectation, for any specific example, it is trivially best to always apply the better treatment and that *Greedy* or other naive heuristics may perform equally well on any particular sample path.

[20] $112 = \lfloor \frac{451}{4} \rfloor$

Table 4
Expected total failures (number and percent), associated standard error, and the size of the PTAS group with π_{JA} as a function of n when $T_{\text{short}} = 60$.

n	Expected number of failures	Expected failure rate (percent)	Standard error (percent)	PTAS size
1	28.55	6.33	0.027	39.25
2	28.64	6.35	0.048	40.57
3	28.72	6.37	0.061	41.72
4	28.81	6.39	0.066	43.02

Notes: (a) actual total failures in the trial is 46 (10.2 percent), (b) PTAS size refers to the number of patients allocated to the stent group, (c) $n = 1$ is equivalent to π_{PA}.

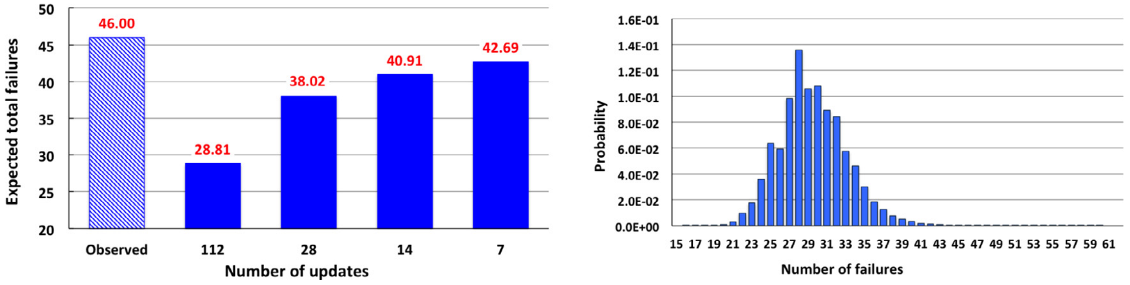


Fig. 8. Expected total failures with π_{JA} as a function of T_{short} (left), and the distribution of failures when $n = 4$ (right). *Notes*: in the right chart, the failures are grouped in buckets of one, for example, the probability of observing failures between 25 and 26; we only include the failures (x-axis) with at least 10^{-8} probability of occurrence.

6. Discussion

Traditional clinical trial designs randomize patients to treatments with a fixed probability throughout the trial. The primary purpose of such a design is to learn; varying randomization probabilities to improve patient outcomes is not a usual consideration. Response-adaptive designs, which allow clinicians to learn about the treatments from patient responses during the trial and adjust patient randomization accordingly, offer an alternative. Adaptive designs can improve patient outcomes, reduce overall development costs, and bring treatments to market sooner.

Existing adaptive designs are sequential in nature, i.e., patients are randomized one at a time. This design follows the classical two-armed Bernoulli bandit problem that exemplifies the tradeoff between the cost of gathering information and the benefit of exploiting the information already gathered, the so-called *exploration vs. exploitation* dilemma. Unless the trial involves a single patient randomization every period and there are minimal delays in observing outcomes, such a design is not practically implementable. While implementable variations of this design exist, they are not fully optimal. We address this gap by proposing the *Jointly Adaptive* design that, at each time period, learns from multiple patients and simultaneously randomizes them to multiple treatments. Our proposed design is forward looking and explicitly considers delays in observing outcomes while building on the Bayesian approach. The key contribution of this paper is, thus, the proposal of a Bayesian MDP framework for finite-horizon problems that learns optimally from simultaneous multiple experiments while allowing for continuous controls, and evaluation of treatments for multiple objectives.

Our proposed design performs better compared to the fixed design, other naive adaptive designs, and heuristics, on both patient health and learning objectives. Consideration of the expected maximum learning objective is another contribution of our work. A key feature of this design is that it naturally allows for mixtures of treatments without imposing constraints artificially. Our numerical results also show that the magnitude of improvement for the *Jointly Adaptive* design over the existing designs can be significant. In addition, using the conditions of the SAMMPRIS trial, we show that a restricted-optimal-policy approximation of the *Jointly Adaptive* design enables effective computation for realistically sized trials and provides close-to-optimal performance with the potential for significant reductions in mortality and morbidity from failed treatments.

Scope. While the majority of clinical trials in practice use fixed designs, adaptive designs, particularly those employing the Bayesian approach, are increasingly being used. According to the FDA, a Bayesian approach, when correctly employed, may be less burdensome than a frequentist approach; for example, a recent guidance document highlighted two examples of device trials that have successfully used Bayesian methods: TRANSCAN (P970033) and INTERFIX (P970015) (FDA, 2010b).

The University of Texas MD Anderson Cancer Center is a pioneer in this area, particularly for cancer trials, owing to a rise in genetic and biological biomarkers that can be used as clinical end points (Berry, 2006). Biswas, Liu, Lee, and Berry (2009) report that protocols for about 20 percent of trials (34 percent for phase I/II trials) conducted at MD Anderson in 2000–2005 used Bayesian statistical designs. Examples of such trials can be found in Muss et al. (2009) and Wilber et al. (2010), highlighting the diversity of diseases and clinical settings.

While Bayesian adaptive designs can be implemented in a wide variety of settings, some trials offer a more conducive environment to the use of the *Jointly Adaptive* design and potential for efficiency gains. These include trials with: (a) relatively rapid observation of patient responses, such as investigations of acute disease interventions, (b) treatments aimed at changes in specific clinical measurements, such as cholesterol and blood pressure, (c) clearly observable primary endpoints such as mortality or treatment discontinuation, (d) newly recruited patient population at each period since that eliminates any issues related to carryover effects when the same patients are reallocated.[21] The scope of the application of adaptive designs is indeed large and covers a number of areas including migraine, oncology, rheumatoid arthritis, diabetes, obesity, stroke, HIV, hepatitis, and Alzheimer' s. Finally, facilities with good information technology and logistical infrastructure are ideal for implementation of adaptive design; poor infrastructure and lack of understanding of complex statistical methodologies has been highlighted as a key barrier to adaptive design implementation (AptivSolutions, 2012; Stadler, 2013).

[21] Carryover effect is defined as the effect of the treatment on the patient from the previous time period on the response of the same patient in the current time period.

We hope to bring further visibility to the importance of response-adaptive designs and expand its applicability in practice. The media attention on I-SPY2 trial (WSJ, 2010) highlights the potential gains from such designs for all interested parties including: (a) the pharmaceutical industry, primarily interested in reducing costs, (b) medical professionals, primarily concerned with bringing effective treatments sooner to patients, and (c) regulatory agencies, primarily concerned with setting policies to ensure that drug and medical device introductions are safe and maximize social welfare.

Limitations. Given underlying assumptions, Bayesian response-adaptive designs may not be fully applicable in all trial contexts. Examples of settings that offer limited benefit include when: (a) patient outcomes cannot be observed until the end of the trial, (b) blinding cannot be fully maintained, (c) frequent analyses create undue burdens on trial participants, (d) time to observation of the primary endpoint is a random variable, (e) drugs are prohibitively expensive and uncertainty in allocation requires large safety stocks at sites, (f) carryover effects exist in trials where same patients are allocated over multiple time periods. van der Graaf, Roes, and van Delden (2012) further discuss how certain features of adaptive trials may create some potential scientific and ethical challenges. Scott and Baker (2007) discuss the challenges in implementing adaptive designs and how those can be addressed.

An essential element in the FDA evaluation of the validity of Bayesian inference for regulatory studies is the demonstration of frequentist operating characteristics, particularly control of type I error. While frequentist measures do not directly translate to Bayesian analysis, alternate approaches such as computer simulations can be used to evaluate type I error rate (or some analog of it) in Bayesian approaches (Berry & Eick, 1995; LeBlond, 2010). Further, in trials that have long-term outcomes as in the case of some phase III studies, variations of adaptive designs such as block randomization can be used (Korn & Freidlin, 2011).

Future work. For extensions of this work, we are currently investigating other approximation methods to address the increased complexity resulting from large numbers of patients and time periods. This dynamic design will retain the key properties of the *Jointly Adaptive* design. We also plan to "retrospectively" implement adaptive designs on other large trials. In addition, we plan to extend our model to include patient heterogeneity, multiple treatments and multiple outcomes, as well as asynchronous delays in observing outcomes (all of which increase the state space). Finally, extending the model to other MDP settings where learning takes place with some observation delay presents opportunities for further development.

Acknowledgments

The authors gratefully acknowledge the financial support provided by The University of Chicago Booth School of Business. This research was supported by funding from the National Institutes of Health, including the National Institute on Aging (T32AG000243; P30AG012857). We also thank Anirban Basu, Don Eisenstein, Varun Gupta, Elbert Huang, Michael Pavlin, Chris Ryan, Linus Schrage, Wei Su, Matt Taddy, and participants of the operations management workshop at Chicago Booth for their helpful comments and suggestions. The authors thank two anonymous referees for their constructive suggestions, and the editor, Jean-Charles Billaut.

Appendix. Proofs

Proof of Lemma 1. We prove the first inequality. Proof of the second inequality uses similar arguments.

It is sufficient to show that $Pr(p > y|\alpha+1,\beta) - Pr(p > y|\alpha,\beta) \geq 0$, or equivalently $F_p(y|\alpha+1,\beta) - F_p(y|\alpha,\beta) \leq 0$, where we removed the superscript j and subscript t for convenience, and $F(\cdot)$ denotes the cdf.

$$\begin{aligned} &F_p(y|\alpha+1,\beta) - F_p(y|\alpha,\beta) \\ &= \int_0^y \left[\frac{1}{B(\alpha+1,\beta)} p^{\alpha}(1-p)^{\beta-1} - \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] dp \\ &= I_y(\alpha+1,\beta) - I_y(\alpha,\beta) \\ &= -\frac{y^{\alpha}(1-y)^{\beta}}{\alpha B(\alpha,\beta)} \leq 0, \end{aligned}$$

where $I_y(\alpha,\beta) = \dfrac{B(y;\alpha,\beta)}{B(\alpha,\beta)}$ is the regularized incomplete beta function, $B(y;\alpha,\beta) = \int_0^y p^{\alpha-1}(1-p)^{\beta-1}$ is the incomplete beta function, $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function, $\Gamma\alpha = (\alpha-1)!$, and equality in the last line follows from the definition of $I_y(\alpha,\beta)$. □

Proof of Proposition 2. We use an induction argument to prove the first inequality. Proof of the second inequality uses similar arguments. Consider stage $T-1$. From (3),

$$\begin{aligned} \mathcal{S}_{T-1}(\boldsymbol{\alpha}^{+}_{T-1},\boldsymbol{\beta}_{T-1}) &= n \max_{j,j'\in\{A,B\},j'\neq j} \left\{ \frac{\alpha^{j}_{T-1}+1}{\alpha^{j}_{T-1}+\beta^{j}_{T-1}+1}, \frac{\alpha^{j'}_{T-1}}{\alpha^{j'}_{T-1}+\beta^{j'}_{T-1}} \right\} \\ &\geq n \max_{j\in\{A,B\}} \frac{\alpha^{j}_{T-1}}{\alpha^{j}_{T-1}+\beta^{j}_{T-1}} = \mathcal{S}_{T-1}(\boldsymbol{\alpha}_{T-1},\boldsymbol{\beta}_{T-1}). \end{aligned}$$

Now assume that the inequality holds true for $t = t+1,\cdots,T-2$, i.e. $\mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1}) \geq \mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1})$. Then $\mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1})] \geq \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1})]$ as follows.

$$\begin{aligned} &\mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1})] \\ &= \sum_{\mathbf{k}_{t+1}=0}^{\mathbf{d}_t} \mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1}) Pr(\mathbf{k}_{t+1}|\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1};\mathbf{d}_t) \\ &= \sum_{\mathbf{k}_{t+1}=0}^{\mathbf{d}_t} \mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1}) Pr(k^{j}_{t+1}|\alpha^{j}_{t+1}+1,\beta^{j}_{t+1};d^{j}_{t}) \\ &\quad Pr(k^{j'}_{t+1}|\alpha^{j'}_{t+1},\beta^{j'}_{t+1};d^{j'}_{t}) \\ &\geq \sum_{\mathbf{k}_{t+1}=0}^{\mathbf{d}_t} \mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1}) Pr(k^{j}_{t+1}|\alpha^{j}_{t+1},\beta^{j}_{t+1};d^{j}_{t}) \\ &\quad Pr(k^{j'}_{t+1}|\alpha^{j'}_{t+1},\beta^{j'}_{t+1};d^{j'}_{t}) \\ &= \sum_{\mathbf{k}_{t+1}=0}^{\mathbf{d}_t} \mathcal{S}_{t+1}(\boldsymbol{\alpha}^{+}_{t+1},\boldsymbol{\beta}_{t+1}) Pr(\mathbf{k}_{t+1}|\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1};d_t) \\ &\geq \sum_{\mathbf{k}_{t+1}=0}^{\mathbf{d}_t} \mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1}) Pr(\mathbf{k}_{t+1}|\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1};d_t) \\ &= \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1})], \end{aligned}$$

where the equality in the second line follows from the fact that treatments are independent of each other, the inequality in the third line follows from the fact that $Pr(k^{j}_{t+1}|\alpha^{j}_{t},\beta^{j}_{t};d^{j}_{t})$ is nondecreasing in α_t, and the inequality in the last line follows from the induction argument.[22] Finally,

$$\begin{aligned}\mathcal{S}_t(\boldsymbol{\alpha}_t^+, \boldsymbol{\beta}_t) &= \max_{\mathbf{u}_t} \{\frac{\alpha_t^j + 1}{\alpha_t^j + \beta_t^j + 1} d_t^j + \frac{\alpha_t^{j'}}{\alpha_t^{j'} + \beta_t^j} d_t^{j'} \\ &\quad + \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1}^+, \boldsymbol{\beta}_{t+1})]\} \\ &\geq \max_{\mathbf{u}_t} \{ \sum_{j=\{A,B\}} \frac{\alpha_t^j}{\alpha_t^j + \beta_t^j} d_t^j + \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1}^+, \boldsymbol{\beta}_{t+1})]\} \\ &\geq \max_{\mathbf{u}_t} \{ \sum_{j=\{A,B\}} \frac{\alpha_t^j}{\alpha_t^j + \beta_t^j} d_t^j + \mathbb{E}_{\mathbf{k}_{t+1}}[\mathcal{S}_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})]\} \\ &= \mathcal{S}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t),\end{aligned}$$

where the inequality in the last line follows from the induction argument. □

Proof of Proposition 3. We use an induction argument to prove the first inequality. Consider $t = T$. We first note that by the definition of Beta distribution,

(a1): $Pr(p_T^j > p_T^{j'}) = Pr(p_T^j > p_T^{j'} | \boldsymbol{\alpha}_T, \boldsymbol{\beta}_T) = \int_0^1 F(p_t^j | \alpha_T^{j'}, \beta_T^{j'})$ $g(p_T^j | \alpha_T^j, \beta_T^j) dp_T^j$, where $F(\cdot)$ denotes the cdf, and $g(\cdot)$ denotes the pdf. (a1) implies the following:

(a2): $Pr(p_T^j > p_T^{j'} | \boldsymbol{\alpha}_T^+, \boldsymbol{\beta}_T) \geq Pr(p_T^j > p_T^{j'} | \boldsymbol{\alpha}_T, \boldsymbol{\beta}_T)$, i.e. $Pr(p_T^j > p_T^{j'})$ is nondecreasing in α_T^j, and

(a3): $Pr(p_T^j > p_T^{j'}) \geq Pr(p_T^{j'} > p_T^j)$ since $\mathbb{E} p_T^j \geq \mathbb{E} p_T^{j'}$ (given). We refer the readers to Cook (2008) for additional details on why (a2) and (a3) hold. Since $|J| = 2$, $Pr(p_T^j > p_T^{j'}) = 1 - Pr(p_T^{j'} > p_T^j)$. This, combined with (a3) yields the following:

(a4): $Pr(p_T^j > p_T^{j'}) \geq \frac{1}{2}$. Next, we note that for $x > 0$, $\max(x, y) = \frac{x + y + |x - y|}{2}$. Therefore, $\max(x, 1 - x) = \frac{1 + |2x - 1|}{2} \geq 0$ if $x \geq \frac{1}{2}$. This implies the following: (a5): $\max(x, 1 - x)$ is increasing in x if $x \geq \frac{1}{2}$.

Setting $x = Pr(p_T^j > p_T^{j'})$ in (a5), we get the following:

(a6): $\max\{Pr(p_T^j > p_T^{j'}), Pr(p_T^{j'} > p_T^j)\} = \mathcal{P}_T$ is increasing in $Pr(p_T^j > p_T^{j'})$.

Together, (a2) and (a6) yield the desired result. Proof for $t = 0, \cdots, T - 1$ follows from the induction argument and is omitted.

Proof for the second inequality, $\mathcal{P}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) \geq \mathcal{P}_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t^+)$, follows from similar arguments, where the key thing to note is that that $Pr(p_T^j > p_T^{j'})$ is nonincreasing in β_T^j. □

Proof of Theorem 1. We first prove (i). Since T is different in π_{PA} and π_{JA}, we present the notation in terms of patients in order to make the designs equivalent, as follows. Let $\mathbf{u}^{\pi_i} = (\mathbf{u}_1^{\pi_i}, \mathbf{u}_2^{\pi_i}, \cdots, \mathbf{u}_N^{\pi_i})$ represent the controls for patient $m = 1, 2, \cdots, N$ under policy $\pi_i \in \Pi_i$, $i = \{PA, JA\}$. Similarly let $\mathbf{h}^{\pi_i} = (\mathbf{h}_1^{\pi_i}, \mathbf{h}_2^{\pi_i}, \cdots, \mathbf{h}_N^{\pi_i})$ denote the information available for patient m under policy π_i. For each m, $\mathbf{u}_m^{\pi_i} : \mathbf{h}_m^{\pi_i} \to \mathscr{P}(\mathcal{U})$, i.e., each $\pi_i \in \Pi_i$ generates a probability distribution on $P^{\pi_i}(\,\cdot\,)$ on each $\mathbf{h}^{\pi_i}$.

Under π_{PA}, patient m arrives at time $t + 1$, while the arrival time for patient m under π_{JA} is given as $t : nt < m \leq n(t + 1)$, thus capturing the dependence on time. The first patient arrives at $t = 0$. This implies the following ordering on the information states for each policy:

(a1): under π_{PA}: for each $m > 1$, $\mathbf{h}_{m-1}^{\pi_{PA}} < \mathbf{h}_m^{\pi_{PA}} < \mathbf{h}_{m+1}^{\pi_{PA}}$, and $\mathbf{h}_1^{\pi_{PA}} < \mathbf{h}_2^{\pi_{PA}} < \ldots < \mathbf{h}_N^{\pi_{PA}}$, while

(b1): under π_{JA}: for $m = nt + 1$ or equivalently when $((m - 1) \bmod n) = 0$, $\mathbf{h}_m^{\pi_{JA}} > \mathbf{h}_{m-1}^{\pi_{JA}}$, and $\mathbf{h}_m^{\pi_{JA}} = \mathbf{h}_{m-1}^{\pi_{JA}}$, otherwise,

where the comparison between information states is componentwise. Together, (a1) and (b1) imply the following:

(a2): when $((m - 1) \bmod n) = 0$, $\mathbf{h}_m^{\pi_{PA}} = \mathbf{h}_m^{\pi_{JA}}$ and $\mathbf{u}_m^{\pi_{PA}} = \mathbf{u}_m^{\pi_{JA}}$, and

(b2): when $((m - 1) \bmod n) > 0$, $\mathbf{h}_m^{\pi_{PA}} > \mathbf{h}_m^{\pi_{JA}}$. and $\mathbf{u}_m^{\pi_{PA}} \supseteq \mathbf{u}_m^{\pi_{JA}}$.

Together, (a2) and (b2) imply that $\mathbf{h}^{\pi_{PA}} \geq \mathbf{h}^{\pi_{JA}}$ and $\mathbf{u}^{\pi_{PA}} \supseteq \mathbf{u}^{\pi_{JA}}$. Since maximization under a restricted control space leads to a suboptimal solution, $V_0^{\pi_{PA}} \geq V_0^{\pi_{JA}}$. We elaborate on the proof below.

We reparameterize the value function as follows:

(a3): under π_{PA}, for each $m = t + 1$, let $\tilde{V}_m^{\pi_{PA}}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m) = V_{t+1}^{\pi_{PA}}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$,

(b3): under π_{JA}: for $m : nt < m \leq n(t + 1)$, let $\tilde{V}_m^{\pi_{JA}}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m) = V_t^{\pi_{JA}}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$.

We prove by induction. Consider the case when $m = N$. For the health objective $\tilde{V}_N^{\pi_{PA}} = \tilde{V}_N^{\pi_{JA}} = 0$. For the learning objective, $\tilde{V}_N^{\pi_{PA}} = \tilde{V}_N^{\pi_{JA}} = \max\{Pr(p_N^A > p_N^B), Pr(p_N^B > p_N^A) | \boldsymbol{\alpha}_N, \boldsymbol{\beta}_N\}$. Thus, $\tilde{V}_N^{\pi_{PA}} \geq \tilde{V}_N^{\pi_{JA}}$, i.e. the relationship holds for N.

Now assume that the relationship holds for $m = m + 1, \cdots, N$, i.e. $\tilde{V}_{m+1}^{\pi_{PA}} \geq \tilde{V}_{m+1}^{\pi_{JA}}$. Then, for all $m = 1, 2, .., m$,

$$\begin{aligned}&\tilde{V}_m^{\pi_{PA}}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m) \\ &\quad = \max_{\mathbf{u}_m^{\pi_{PA}}} \{R_m + \mathbb{E}_{\mathbf{k}_{m+1}}[\tilde{V}_{m+1}^{\pi_{PA}}(\boldsymbol{\alpha}_{m+1}, \boldsymbol{\beta}_{m+1})]\} \\ &\quad \geq \max_{\mathbf{u}_m^{\pi_{JA}}} \{R_m + \mathbb{E}_{\mathbf{k}_{m+1}}[\tilde{V}_{m+1}^{\pi_{PA}}(\boldsymbol{\alpha}_{m+1}, \boldsymbol{\beta}_{m+1})]\} \\ &\quad \geq \max_{\mathbf{u}_m^{\pi_{JA}}} \{R_m + \mathbb{E}_{\mathbf{k}_{m+1}}[\tilde{V}_{m+1}^{\pi_{JA}}(\boldsymbol{\alpha}_{m+1}, \boldsymbol{\beta}_{m+1})]\} = \tilde{V}_m^{\pi_{JA}}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m),\end{aligned}$$

where the first inequality follows from the fact that $\mathbf{u}^{\pi_{PA}} \supseteq \mathbf{u}^{\pi_{JA}}$ and the second inequality follows from the induction assumption. This completes the proof of (i).

For (ii), we revert to the notation in terms of time. Let $\mathbf{u}_t^{\pi_i} = (\mathbf{u}_{t_1}^{\pi_i}, \mathbf{u}_{t_2}^{\pi_i}, \cdots, \mathbf{u}_{t_n}^{\pi_i})$ and $\mathbf{h}_t^{\pi_i} = (\mathbf{h}_{t_1}^{\pi_i}, \mathbf{h}_{t_2}^{\pi_i}, \cdots, \mathbf{h}_{t_n}^{\pi_i})$ represent the controls and information available, respectively, for patient $m = 1, 2, \cdots, n$ at time $t \in \{0, \cdots, T - 1\}$ under policy $\pi_i \in \Pi_i$, $i = \{JA, IA\}$, such that $\mathbf{u}_t^{\pi_i} : \mathbf{h}_t^{\pi_i} \to \mathscr{P}(\mathcal{U})$.

For any given $t > 0$ and for each $m = 1, 2, \cdots, n$,

(a4): under π_{JA}: $\mathbf{h}_{t+1_m}^{\pi_{JA}} = \mathbf{h}_{t_m}^{\pi_{JA}} + \sum_{m=1,\cdots,n} \mathbf{k}_{t+1_m}^{\pi_{JA}}$ and $\mathbf{h}_{t_1}^{\pi_{JA}} = \mathbf{h}_{t_2}^{\pi_{JA}} = \ldots = \mathbf{h}_{t_n}^{\pi_{JA}}$, and

(b4): under π_{IA}: $\mathbf{h}_{t+1_m}^{\pi_{IA}} = \mathbf{h}_{t_m}^{\pi_{IA}} + \mathbf{k}_{t+1_m}^{\pi_{IA}}$, and each $\mathbf{h}_{t_m}^{\pi_{IA}}$ is independently updated.

Together, (a4) and (b4) imply that $\mathbf{h}_t^{\pi_{JA}} \geq \mathbf{h}_t^{\pi_{IA}}$, and $\mathbf{u}_t^{\pi_{JA}} \supseteq \mathbf{u}_m^{\pi_{IA}}$. The rest of the proof is similar to (i) where the key argument is that optimization over the restricted control space leads to a suboptimal solution.

For (iii), we use the induction argument again. First, the statement holds true for all $t = t + 1, \cdots, T$ for both objectives, as shown in (i). Then, from the definition of π_{RA},

$$\begin{aligned}V_t^{\pi_{RA}}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) &= \max_{\mathbf{u}_t^{\pi_{JA}}} \{R_t + \mathbb{E}_{\mathbf{k}_{t+1}}[V_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})]\} \\ &\qquad \text{s.t.} \quad d_t^j = n \text{ or } d_t^j = 0 \; \forall j, \\ &\qquad \leq \max_{\mathbf{u}_t^{\pi_{JA}}} \{R_t + \mathbb{E}_{\mathbf{k}_{t+1}}[V_{t+1}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1})]\} = V_t^{\pi_{JA}}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t),\end{aligned}$$

where the inequality above stems from the fact that adding a constraint leads to a suboptimal solution.

Finally for (iv), the argument is the same as in (iii) except with a different (and stronger) constraint: $d_t^j = \frac{n}{2}$ for all $j \in J$ and $t \in \{0, \cdots, T - 1\}$. □

Proof of Lemma 2. We prove the result for $T \to \infty$ and note that the same argument holds for $n \to \infty$. We first prove the result for the patient health objective.

[22] To see why $Pr(k_{t+1}^j | \alpha_t^j, \beta_t^j; d_t^j)$ is nondecreasing in α_t, note that the probability equals $\int_0^1 Pr(k_{t+1}^j | d_t^j, p_t^j) g(p_t^j | \alpha_t^j, \beta_t^j) dp_t^j$. The first term in the integral is a binomial likelihood and nondecreasing in p_t^j. Since p_t^j is nondecreasing in α_t^j, the result follows.

Let π^* represent the optimal policy for this objective. Suppose π^* plays arm j a finite number of times with probability $q^j > 0$, i.e., $\sum_{t=0}^{T-1} d_t^{j,\pi^*} < \infty$ as $T \to \infty$ $w.p.\, q^j$. Note that arm j' is played i.o. Then, without loss of generality, we can assume that there exists a state $\hat{\mathbf{h}}$ that occurs with probability $q^h > 0$, from which arm j is never chosen by π^*. Let $\hat{p}^j = \mathbb{E}p^j$ represent the expected probability of success with arm j at state $\hat{\mathbf{h}}$. Consider time $\tau^j = \inf(t(\omega) : \hat{p}^j > \hat{p}^{j'}(t, \omega, \pi^*(t+1)) + \epsilon)$ for all $j \neq j'$ and $\epsilon > 0$. Here, τ^j represents the minimum time starting at state $\hat{\mathbf{h}}$ when $\hat{p}^j$ exceeds $\hat{p}^{j'}$, and ω is the sample path on which this happens.[23] For any $\epsilon < \hat{p}^j$, $P(\tau^j < \infty) = 1$.[24]

Now, consider the policy π' that follows $\pi^*(\tau^j + 1)$ in all states except that it chooses arm j at τ^j, which occurs with probability at least q^h. Then, $V_{\tau^j+1}^{\pi'} > V_{\tau^j+1}^{\pi^*} + q^h\epsilon$, contradicting the optimality of $\pi^*(\tau^j + 1)$.

The proof for the learning objective follows a similar argument where we note that each of the two objectives is strictly increasing in $\max_j \{\hat{p_T}^j\}$ for all $j \in J$. □

Proof of Lemma 3. As above, we only prove the result for $T \to \infty$. The proof relies on showing (a) $\mathbb{E}p_T^j$ is a consistent estimator of $\bar{p}^j$ and (b) $p_T^j \overset{P}{\to} \mathbb{E}p_T^j$ as $T \to \infty$. From the definition of $\mathbb{E}p_T^j$,

$$\mathbb{E}p_T^j = \frac{\alpha_0^j + h_T^{j,s}}{(\alpha_0^j + \beta_0^j) + nT} = \frac{\frac{\alpha_0^j}{nT}}{\frac{(\alpha_0^j + \beta_0^j)}{nT} + 1} + \frac{\frac{h_T^{js}}{nT}}{\frac{(\alpha_0^j + \beta_0^j)}{nT} + 1},$$

we observe:

(a1): $\frac{(\alpha_0^j + \beta_0^j)}{nT} \xrightarrow[T\to\infty]{0}$, and $\frac{(\alpha_0^j + \beta_0^j)}{nT} + 1 \xrightarrow[T\to\infty]{1}$, by applying Slutsky's Theorem,

(a2): $\frac{h_T^{js}}{nT} \xrightarrow[T\to\infty]{a.s.} \bar{p}^j$ by Strong Law of Large Numbers, and $\frac{\alpha_0^j}{nT} \xrightarrow[T\to\infty]{0}$.

Combining (a1) and (a2) yields the following:

(a3): $\mathbb{E}p_T^j \xrightarrow[T\to\infty]{a.s.} \bar{p}^j$.

For a random variable X with mean μ and variance σ^2, the Chebyshev inequality states that for any $k > 0$, $Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$. Choosing $X = p_T^j$, $k = \frac{\epsilon}{\sigma}$, where $\epsilon > 0$ and $\sigma^2 = Var(p_T^j)$, we obtain,

(a4): $Pr(|p_T^j - \mathbb{E}p_T^j| \geq \epsilon) \leq \frac{Var(p_T^j)}{\epsilon^2}$.

Since $Var(p_T^j) = \frac{\alpha_T^j \beta_T^j}{(\alpha_T^j + \beta_T^j)^2(\alpha_T^j + \beta_T^j + 1)} \xrightarrow[T\to\infty]{a.s.} 0$, it follows from (a4) that $Pr(|p_T^j - \mathbb{E}p_T^j| \geq \epsilon) \xrightarrow[T\to\infty]{0}$, or equivalently,

(a5): $p_T^j \xrightarrow[T\to\infty]{P} \mathbb{E}p_T^j$, by definition of convergence in probability.

Combining (a3) and (a5), and applying Slutsky's theorem twice,

(a6): $p_T^A - p_T^B \xrightarrow[T\to\infty]{P} \mathbb{E}p_T^A - \mathbb{E}p_T^B \xrightarrow[T\to\infty]{a.s.} \bar{p}^A - \bar{p}^B > 0$.

Since a.s. convergence implies convergence in probability, $p_T^A - p_T^B \xrightarrow[T\to\infty]{P} \bar{p}^A - \bar{p}^B > 0$, giving us the desired result. □

Proof of Theorem 2. Since π_{JA} is the optimal policy that belongs to the Π_{JA} class of policy, it follows from Lemma 2 that π_{JA} allocates infinite number of patients to each treatment in the limit. The proof is complete by noting that any policy with this property achieves the desired result, as per Lemma 3. □

[23] Since π^* never chooses arm j from $\hat{\mathbf{h}}$, τ^j represents the time it takes for a string of failures with arm j' until $\hat{p}^j$ just exceeds $\hat{p}^{j'}$.

[24] With multiple arms, some of which are played a finite number of times, $\hat{\mathbf{h}}$ represents the state at which the last of these finitely played arms, represented by arm j, is played, and only arms that are played i.o., exemplified by arm j', remain; thus we can apply a similar argument.

References

AptivSolutions (2012). Top barriers to adaptive trial implementation. URL: http://www.aptivsolutions.com/blog/adaptive-trials/2012/05/top-barriers-to-adaptive-trial-implementation.

Arlotto, A., Chick, S. E., & Gans, N. (2013). Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science, 60*(1), 110–129.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*(2), 235–256.

Aviv, Y., & Pazgal, A. (2005). A partially observed Markov decision process for dynamic pricing. *Management Science, 51*(9), 1400–1416.

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933–1960), 16*(3/4), 221–229.

Berry, D. (1978). Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association, 73*(362), 339–345.

Berry, D. (1985). Interim analyses in clinical trials: Classical vs. Bayesian approaches. *Statistics in Medicine, 4*(4), 521–526.

Berry, D. (1987). Interim analysis in clinical trials: The role of the likelihood principle. *American Statistician, 41*(2), 117–122.

Berry, D. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine, 12*(15–16), 1377–1393.

Berry, D. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery, 5*(1), 27–36.

Berry, D., & Eick, S. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine, 14*(3), 231–246.

Berry, D., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments*. London: Chapman and Hall.

Berry, D., & Pearson, L. (1985). Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine, 4*(4), 497–508.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*: 1. Belmont: Athena Scientific.

Bertsimas, D., & Mersereau, A. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research, 55*(6), 1120–1135.

Besbes, O., & Zeevi, A. (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research, 57*(6), 1407–1420.

Biswas, S., Liu, D. D., Lee, J. J., & Berry, D. A. (2009). Bayesian clinical trials at the University of Texas MD Anderson Cancer Center. *Clinical Trials, 6*(3), 205–216.

Bona, M. (2002). *A walk through combinatorics* (2nd). River Edge, NJ: World Scientific Publishing Company Inc..

Burnetas, A. N., & Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics, 17*(2), 122–142.

Caro, F., & Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science, 53*(2), 276.

Chalmers, T. C., Celano, P., Sacks, H. R., & Harry Smith, J. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine, 309*, 1358–1361.

Cheng, Y., & Berry, D. (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika, 94*(3), 673.

Chimowitz, M. I., Lynn, M. J., Derdeyn, C. P., Turan, T. N., Fiorella, D., Lane, B. F., Janis, L. S., Lutsep, H. L., Barnwell, S. L., Waters, M. F., et al. (2011). Stenting versus aggressive medical therapy for intracranial arterial stenosis. *New England Journal of Medicine, 365*(11), 993–1003.

Chimowitz, M. I., Lynn, M. J., Turan, T. N., Fiorella, D., Lane, B. F., Janis, S., & Derdeyn, C. P. (2011). Design of the stenting and aggressive medical management for preventing recurrent stroke in intracranial stenosis trial. *Journal of Stroke and Cerebrovascular Diseases, 20*(4), 357–368.

Chow, S., & Chang, M. (2008). Adaptive design methods in clinical trials—A review. *Orphanet Journal of Rare Diseases, 3*, 11.

Cook, J. D. (2008). *Numerical computation of stochastic inequality probabilities*. UT MD Anderson Cancer Center, Department of Biostatistics. Technical report 46.

Duff, M. (2003). Design for an optimal probe. In *Proceedings of the twentieth international conference on machine learning* (pp. 131–138).

Emanuel, E. J., Schnipper, L. E., Kamin, D. Y., Levinson, J., & Lichter, A. S. (2003). The costs of conducting clinical research. *Journal of Clinical Oncology, 21*(22), 4145–4150.

English, R., Lebovitz, Y., Griffin, R., et al. (2010). *Transforming clinical research in the United States: Challenges and opportunities: Workshop summary*. National Academies Press.

Farias, V., & Van Roy, B. (2010). Dynamic pricing with a prior on market response. *Operations Research, 58*(1), 16–29.

FDA (2005). Summary of safety and probable benefit. URL: www.accessdata.fda.gov/cdrh_docs/pdf5/H050001b.pdf.

FDA (2010). Adaptive design clinical trials for drugs and biologics, guidance for industry. URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf.

FDA (2010). Guidance for the use of Bayesian statistics in medical device clinical trials, guidance for industry and FDA staff. URL: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf.

Forbes (2012). Pfizer Q1 earnings slump on Lipitor patent expiry. URL: http://www.forbes.com/sites/greatspeculations/2012/05/02/pfizer-q1-earnings-slump-on-lipitor-patent-expiry.

Forbes (2013). The cost of creating a new drug now 5 billion, pushing big pharma to change. URL: www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine.

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, & S. G. Walker (Eds.), *Bayesian nonparametrics*. In *Cambridge series in statistical and probabilistic mathematics* (pp. 35–79). Cambridge: Cambridge University Press.

Ghosal, S., Ghosh, J., & Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics, 23*(6), 2145–2152.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological), 41*(2), 148–177.

Gittins, J. C. (1989). Multi-armed bandit allocation indices. *Wiley Interscience series in systems and optimization*. Chichester, New York.

Hardwick, J., Oehmke, R., & Stout, Q. F. (2006). New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference, 136*(6), 1940–1955.

Hardwick, J. Stout, Q. F. (2008). Response adaptive designs for balancing complex objectives. Working paper, University of Michigan.

Harrison, J. M., Keskin, N. B., & Zeevi, A. (2012). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science, 58*(3), 570–586.

Honda, J., & Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *COLT* (pp. 67–79).

Katehakis, M. N., & Veinott, A. F. (1987). The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research, 12*(2), 262–268.

Korn, E. L., & Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology, 29*(6), 771–776.

Lai, T., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics, 6*(1), 4–22.

LeBlond, D. (2010). Statistical viewpoint: FDA Bayesian statistics guidance for medical device clinical trials—Application to process validation. *Journal of Validation Technology, 16*(4), 24.

Lee, J. J., & Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials, 5*(2), 93–106.

Lieberman, J., Stroup, T., McEvoy, J., Swartz, M., Rosenheck, R., Perkins, D., Keefe, R., Davis, S., Davis, C., Lebowitz, B., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine, 353*(12), 1209–1223.

Mersereau, A., Rusmevichientong, P., & Tsitsiklis, J. (2009). A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control, 54*(12), 2787–2802.

Murphy, S. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24*(10), 1455–1481.

Muss, H. B., Berry, D. A., Cirrincione, C. T., Theodoulou, M., Mauer, A. M., Kornblith, A. B., Partridge, A. H., Dressler, L. G., Cohen, H. J., Becker, H. P., et al. (2009). Adjuvant chemotherapy in older women with early-stage breast cancer. *New England Journal of Medicine, 360*(20), 2055–2065.

Ning, J., & Huang, X. (2010). Response-adaptive randomization for clinical trials with adjustment for covariate imbalance. *Statistics in Medicine, 29*(17), 1761–1768.

NYTimes (2011). Study is ended as a stent fails to stop strokes. URL: http://www.nytimes.com/2011/09/08/health/research/08stent.html?_r=0.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming* (1st). New York, NY, USA: John Wiley & Sons, Inc.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society, 58*(5), 527–535.

Roy, A. S. A. (2012). Stifling new cures: The true cost of lengthy clinical drug trials. URL: http://www.manhattan-institute.org/html/fda_05.htm.

Scott, C. T., & Baker, M. (2007). Overhauling clinical trials. *Nature Biotechnology, 25*(3), 287–292.

Stadler, W. M. (2013). The University of Chicago. Personal conversation, November 2013.

Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences, 4*(1), 8.

van der Graaf, R., Roes, K. C., & van Delden, J. J. (2012). Adaptive trials in clinical research: Scientific and ethical issues to consider. *Journal of the American Medical Association, 307*(22), 2379–2380.

Wilber, D. J., Pappone, C., Neuzil, P., De Paola, A., Marchlinski, F., Natale, A., Macle, L., Daoud, E. G., Calkins, H., Hall, B., et al. (2010). Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation. *JAMA: The Journal of the American Medical Association, 303*(4), 333–340.

WSJ (2010). A new Rx for medicine. URL: http://online.wsj.com/news/articles/SB10001424052748703882404575520190576846812.