Response-Adaptive Designs for Clinical Trials: Simultaneous Learning from Multiple Patients

Vishal Ahuja
The University of Chicago, Center on Aging at NORC, Chicago, IL 60637, vahuja@chicagobooth.edu

John Birge
The University of Chicago, Booth School of Business, Chicago IL 60637, john.birge@chicagobooth.edu

Abstract
Clinical trials have traditionally followed a fixed design, in which patient allocation to treatments is 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 patient allocation to treatments as the trial progresses. An ideal adaptive design is one where patients are treated as effectively as possible without sacrificing potential learning or compromising the integrity of the trial. We propose such a design, one that uses forward-looking algorithms to fully exploit learning from multiple patients simultaneously. Compared to the best existing implementable adaptive design (e.g. using Berry [1978]), we show that our proposed design improves patient outcomes by up to 8.6% under a set of considered scenarios. Further, we demonstrate our design’s effectiveness using data from a recently conducted stent trial. The design is general and applicable to any Markov decision process setting where learning takes place from multiple simultaneous individual experiments as, for example, in customized consumer offers. 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 small delays in observing patient outcomes. We do this by showing the relative performance of these schemes for maximum expected health and maximum expected learning objectives, and by demonstrating the value of a truncated-horizon approximation in a practical example.

Keywords: Adaptive clinical trials; Markov decision process; Bayesian learning; Stents.

1 Introduction

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 forty 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, and exceed $6000 per

1This 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).

2Given that 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 the trial. For example, the sales of the drug Atorvastatin (trade name: Lipitor) decreased by 42%, from $2.4 billion to $1.4 billion, after its expiration on November 30, 2011 (Forbes, 2012).
enrolled subject (Emanuel et al., 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 assigns patients to treatments in a constant proportion 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 longer times for drug approval. Consequently, regulatory bodies, such as the U.S. Food and Drug Administration, have recently encouraged the use of adaptive designs (FDA, 2010a,b).

There are 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 that 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 increase the probability of finding the successful treatment, identify ineffective and unsafe drugs sooner, and require fewer patients in the trial. As a result, adaptive designs can potentially reduce costs and shorten overall development timelines significantly. Adaptive designs 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, in particular control of Type I error, as required by FDA (FDA, 2010a,b). According to Berry and Eick (1995), 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 and Berry, 2007). Further, while frequentist analysis is not directly applicable to Bayesian settings, analogous Bayesian measures exist (Berry, 1993). Notably, the concept of predictive probability provides a Bayesian equivalent to the probability of Type I error as it provides the ability to quantify subsequent trial results given current information (Lee and Liu, 2008).

Berry and co-authors were among the first to develop a truly Bayesian response-adaptive design (see, for example, Berry, 1978, Berry and Pearson, 1985). In their design, patient allocation to treatments happens sequentially, that is, one at a time and all previous patient response(s) are known and incorporated into the allocation decision 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. Alternately, this design may be implemented when there is minimal delay in observing outcomes. Given that a vast majority of trials consist of multiple patients that need to be allocated
simultaneously before outcomes can be observed, the sequential design is rendered impractical. Our work directly addresses this gap by developing an adaptive design with multiple simultaneous randomized treatment allocations to anticipate learning through the trial horizon.

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 for clinical trials that learns simultaneously from multiple patients and optimally randomizes them to multiple treatments, consideration of a learning objective in addition to the health objective, evaluation of the relative advantage of adaptive designs over a fixed design; and the proposal of a new method (truncated horizon approximation) for solving large clinical trials. Finally, 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. §2 provides a brief overview of the literature. §3 presents the model. §4 describes various adaptive designs. In §5, we present numerical results, including application to a recently conducted clinical trial. In §6, we summarize and discuss our conclusions as well as the scope and limitations of adaptive designs.

2 Literature Overview

The majority of previous work on trial design appears in the statistic literature. 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 multi-armed 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 Gittins Index, to solve bandit problems using forward induction; Katehakis and Veinott (1987) characterize this index in a way that allows it to be calculated more easily.

Berry (1978) is one of the first studies to fully incorporate 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 et al., 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 typically focus on maximizing expected patient health.
A related stream of literature has investigated asymptotically adaptive policies for bandit problems, one that achieves 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 et al. (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 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. 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 et al. (2006) relax 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 allocation opportunity. For our model, we assume a constant delay but note that asynchronous delays could be included without changing the basic structure. 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 and Gallien, 2007) and dynamic learning about employee performance to formulate an employer’s hiring and retention decisions (Arlotto et al., 2010). Within the 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 et al. (2012). Readers are directed to Besbes and Zeevi (2009) as an example of study that uses classical statistics framework to learn about the underlying demand function. Harrison et al. (2012) discuss further connections to antecedent literature.

A paper that is particularly close to our work is Bertsimas and Mersereau (2007), who use multi-armed bandit framework in an interactive marketing context. Our work differs from theirs, as follows. First, we allow for randomized strategies, while Bertsimas and Mersereau (2007) restrict choices to integers that restricts the control space. Second, we consider a maximum expected learning objective (that provides an alternative to the Type I and Type II errors), in addition to a maximum expected successes objective that Bertsimas and Mersereau (2007) consider. Further, we analyize 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 truncated-horizon approximation, our design allows for an optimal solution developed for a smaller problem.
to be applied to a larger cohort. 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 initial study does not consider potential serial correlation in treatment effects, but our framework also applies when such dependencies are present. 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 and Berry, 2007), our design 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 finds 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 et al. (2009) is an example of a study that considers correlated arms and show 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). Unlike the classical MDP setup, the true 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.

In our model specification, we assume that the number of treatments and outcomes remain constant throughout the trial. We believe this is a reasonable assumption that reflects practical settings, although the model is generalizable to the case where the number of treatments and outcomes vary with time. Further, we assume independent and identically distributed (i.i.d) patients,

---

3We follow the terminology of Duff (2003).
and a constant number of patients allocated each period.

### 3.1 General Model specification

Let $T$ be the trial length or the total number of time periods in the trial, where a time period corresponds to the delay between initial treatment and outcome observation. 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 $J$ and $O$ be the set of treatments and outcomes, respectively. The corresponding set of health conditions, $I$, is obtained from a Cartesian product of those sets ($J \times O$).

We define the health information state as a vector $h_t \in \mathcal{H} \subseteq \mathbb{Z}^{|J|\times|O|}$, defined as follows:

$$h_t = (h_t^{11}, \ldots, h_t^{|J|,|O|}).$$

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

The controls, $u_t \in \mathcal{U} \subseteq \mathbb{R}^{|J|}_+$ are defined as follows:

$$u_t = (u_t^1, \ldots, u_t^{|J|}).$$

Here $u_t^j \in [0, 1]$ is the probability of assigning a patient to treatment $j \in J$ at time $t \in \{0, \ldots, T-1\}$ such that $\sum_{j \in J} u_t^j = 1$. The set of decisions, $d_t$, is random and obtained from the controls\(^4\) as follows:

$$d_t = (d_t^1, \ldots, d_t^{|J|}).$$

Here $d_t^j \in \mathbb{Z}_+$ is the number of patients assigned to treatment $j \in J$ at time $t \in \{0, \ldots, T-1\}$ such that $\sum_{j \in J} d_t^j = n$. $Pr(d_t|n, u_t) \sim Mu(d_t|n, u_t)^\frac{1}{|J|}$ and $Ed_t = nu_t^j$. Define $\hat{d}_t^j = Ed_t^j$ and $\hat{d}_t = (\hat{d}_t^1, \ldots, \hat{d}_t^{|J|})$.

We note that patients begin arriving at $t = 1$, and decisions for patients arriving at $t$ are made at $t-1$, and no decision is made at $t = T$.

We define the vector of probabilities as follows:

$$p_t^j = (p_t^{j,1}, \ldots, p_t^{j,|O|}).$$

Here, $p_t^{j,o}$ represents the probability of observing outcome $o \in O$ at time $t+1$ given treatment $j \in J$ at time $t$. We assume a generalized multinomial likelihood on the transition to state $h_{t+1}$ from state $h_t$, given $p_t$, and use a Dirichlet conjugate prior on $p_t$ with hyperparameters $\alpha_t = (\alpha_t^{11}, \ldots, \alpha_t^{|J|,|O|})$ for $t \in \{0, \ldots, T\}$. If we denote the initial priors by $\alpha_0 = (\alpha_0^{11}, \ldots, \alpha_0^{|J|,|O|})$ and assume that the outcomes of patients in different health conditions are not informative of each other, then each $\alpha_t^{j,o}$

\(^4\)For example, $d_t$ is obtained by tossing a $|J|$ sided dice where the probability of observing side $j$ is $u_t^j$.

\(^5\)$Mu$ denotes multinomial distribution.
can be updated independently as follows: \( \alpha^{j,o}_t = \alpha^{j,o}_0 + h^{j,o}_t \), where \( h^{j,o}_t \) captures all the (random) realizations from the past for that treatment-outcome combination.

Given the decision \( d_{t-1} \), the (random) outcomes are observed in the next period, captured in the vector \( k_t \in \mathcal{K} \subseteq \mathbb{Z}^{J \times |O|} \), that we define as the physical state, as follows:

\[
k_t^j = (k_t^{j,1}, \ldots, k_t^{j,|O|}).
\]

Here, \( k_t^{j,o} \in \mathbb{Z}_+ \) represents the number of observed patients in health condition (\( j, o \)) at a given time \( t \in \{1, \ldots, T\} \), where the treatment \( j \in J \) is given at time period \( t-1 \) and the outcome \( o \in O \) is observed in time \( t \), such that \( \sum_{j \in J, o \in O} k_t^{j,o} = n \). The above definitions directly imply the following:

for \( t = 1 \), \( h_t = k_t \) and for \( t = 2, \ldots, T \), \( h_t = h_{t-1} + k_t \).

Additionally, we define the following terms: \( \mathbf{p}_t = (p_t^{1,1}, \ldots, p_t^{j,|O|}) \), \( \alpha_t^j = (\alpha_t^{j,1}, \ldots, \alpha_t^{j,|O|}) \), and \( \alpha_0^j = (\alpha_0^{j,1}, \ldots, \alpha_0^{j,|O|}) \), for all \( o \in O, j \in J \), similar to the definition of \( h_t^j \) and \( k_t^j \) above.

The entries of the transition matrix at time \( t \in \{0, \ldots, T-1\} \), \( P_t(h_{t+1} | h_t, d_t, \alpha_0) \), representing the probability of transitioning to state \( h_{t+1} \), given \( h_t, d_t, \) and \( \alpha_0 \), is then defined as follows:

\[
P_t(h_{t+1} | h_t, d_t, \alpha_0) = \prod_{j \in J} \Pr(k_{t+1}^{j,} | h_t^{j}, d_t^{j}, \alpha_0^{j}) = \prod_{j \in J} \int_0^1 P_t(k_{t+1}^{j,} | d_t^{j}, p_t^j) g(p_t^j | h_t^{j}, \alpha_0^{j}) dp_t^j, \tag{1}
\]

if \( d_t^j \in \mathbb{Z} \) and \( k_{t+1}^{j,o} \leq d_t^j \) for all \( j \in J, o \in O \), and 0 otherwise. Here, \( P_t(k_{t+1}^{j,} | d_t^{j}, p_t^j) = Pr(k_{t+1}^{j,1}, \ldots, k_{t+1}^{j,|O|} | d_t^{j,1}, \ldots, p_t^{j,|O|}) \) is the multinomial likelihood or the marginal joint distribution of observing \( k_{t+1}^{j,1}, \ldots, k_{t+1}^{j,|O|} \) outcomes from \( d_t^j \) patients given that the probability of observing these outcomes is \( p_t^{j,1}, \ldots, p_t^{j,|O|} \), respectively, and \( g(p_t^j | h_t^{j}, \alpha_0^{j}) = g(p_t^{j,1}, \ldots, p_t^{j,|O|} | \alpha_0^{j}) = g(p_t^{j,1}, \ldots, p_t^{j,|O|} | \alpha_0^{1}, \ldots, \alpha_0^{j,|O|}) \) is the probability density function (pdf) for the Dirichlet distribution. Finally, the reward, \( R_t \), depends on the objective function chosen as follows:

(a) Patient Health (PH): \( R_T = 0 \) and \( R_t = r^T k_{t+1} \forall t \in \{0, \ldots, T-1\} \), where \( r \subseteq \mathbb{R}^{J \times |O|} \).

(b) Learning about treatment efficacy (LE): Assuming that the efficacy of a given treatment is measured based on achieving one desired outcome, \( \delta \in O \), \( R_T = \max_{j \in J} \Pr \{ p_T^j (\delta | h_T) > \max_{j \not\in \delta} \{ p_T^j (\delta | h_T) \} \} \) and \( R_t = 0 \forall t \in \{0, \ldots, T-1\} \). This definition represents the reliability of the trial in terms of conclusively determining the most efficacious treatment.

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:

\[
V_t(\alpha_t, \beta_t) = \max_{\beta_t} \{ R_t + E_{k_{t+1}} [V_{t+1}(\alpha_{t+1}, \beta_{t+1})] \}. \tag{2}
\]

We note that the following notations for value function are equivalent: \( V_t(\alpha_t, \beta_t) = \tilde{V}_t(\alpha_t, \beta_t; n, T) \).
3.2 Simple Example

Consider a clinical trial consisting of two treatments, henceforth referred to as A and B, and two mutually exclusive outcomes, henceforth referred to as success (s) and failure (f). Then, \( J = \{A, B\} \), \( O = \{s, f\} \), \( I = \{As, Af, Bs, Bf\} \). For convenience of notation, we summarize the information and physical states as follows: \( h_t = (h_t^A, h_t^B) \) and \( k_t = (k_t^A, k_t^B) \), where \( k_t^{js} = k_t^j \) and \( h_t^{ij} = h_t^i \) for \( j = \{A, B\} \). This implies \( h_t^{ij} = \sum_{t=0}^\infty d_t^j - h_t^i k_t^{ij} = d_t^j - k_t^j \). We define additional terms as follows:

\[
\alpha_t^j = \alpha_t^j, \quad \alpha_t^{ij} = \beta_t^j, \quad \alpha_t = (\alpha_t^A, \alpha_t^B), \quad \beta_t = (\beta_t^A, \beta_t^B) \quad \forall t \\
\text{and for } j \in \{A, B\}:
\]

\[
\alpha_t^j, \alpha_t^{ij} = \beta_t^j, \quad \alpha_t = (\alpha_t^A, \alpha_t^B), \quad \beta_t = (\beta_t^A, \beta_t^B) \quad \forall t \\
\text{and for } j \in \{A, B\}:
\]

Finally, the rewards are defined for each objective function. For health, following existing literature (e.g., Berry, 1978), \( r = (1, 0, 1, 0) \), implying a reward of 1 for success and 0 for failure. For learning: \( R_T = \max\{Pr(p_T^A > p_T^B), Pr(p_T^B > p_T^A)\} \).

Let \( S_t \) and \( \mathcal{P}_t \) denote the value function \((V_t)\) for the the health and learning objectives, respectively. We can then describe the dynamic program in (2) for each objective separately, as follows.

(a) **Health**: \( S_T = 0 \). 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.

\[
S_{T-1}(\alpha_{T-1}, \beta_{T-1}) = n \max_j \frac{\alpha_{T-1}^j}{\alpha_{T-1}^j + \beta_{T-1}^j}.
\]  

(3)

For \( t = 0, 1, ..., T - 2 \),

\[
S_t(\alpha_t, \beta_t) = \max_{u_t} \left\{ \sum_{j \in \{A, B\}} \frac{\alpha_t^j}{\alpha_t^j + \beta_t^j} d_t^j + \mathbb{E}_{k_{t+1}} [S_{t+1}(\alpha_{t+1}, \beta_{t+1})] \right\}.
\]  

(4)

(b) **Learning**: \( \mathcal{P}_T = \max\{Pr(p_T^A > p_T^B), Pr(p_T^B > p_T^A)\} \), and for \( t = 0, 1, ..., T - 1 \),

\[
\mathcal{P}_t(\alpha_t, \beta_t) = \max_{u_t} \mathbb{E}_{k_{t+1}} [\mathcal{P}_{t+1}(\alpha_{t+1}, \beta_{t+1})].
\]  

(5)

As before, the following equivalent notations hold: \( S_t(\alpha_t, \beta_t) = \bar{S}_t(\alpha_t, \beta_t; n, T) \), and \( \mathcal{P}_t(\alpha_t, \beta_t) = \bar{S}_t(\alpha_t, \beta_t; n, T) \).

---

7The definition of a success (typically the desired outcomes) or a failure is defined by the clinician(s) for the clinical trial. For example, an outcome may be considered a success (failure) if a patient responds (does not respond) to a particular treatment, where the response measure is defined by the clinician or determined by the existing standards.
Figure 1: State transition diagram for $n = T = 1$ (left) and the enumerated state and decision space for $n = T = 2$ (right), where $J = \{A, B\}; O = \{s, f\}$.

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 look like as follows:

$$h_5 = (8, 2, 5, 5), \{(\alpha_5^A, \beta_5^A); (\alpha_5^B, \beta_5^B)\} = \{(9, 3); (6, 6)\}.$$ 

One solution under health objective is: $u_5 = (0.7, 0.3), d_5 = (3, 1)$, and a potential next state is:

$$k_6 = (2, 1, 1, 0); h_5 = (10, 3, 6, 5); \{(\alpha_6^A, \beta_6^A); (\alpha_6^B, \beta_6^B)\} = \{(11, 4); (7, 6)\}.$$ 

The number of unique states in this setup can be viewed as a partitioning $nt$ objects (patients seen to date) into $\hat{k}$ blocks (health conditions) such that $\hat{k} = |I|$. This is precisely the number of $\hat{k}$ weak compositions of $nt$ (Bona, 2002), given by the formula $\binom{nt+\hat{k}-1}{\hat{k}-1} = \frac{(nt+\hat{k}-1)!}{nt!(\hat{k}-1)!}$. The total number of unique states that need to be solved equals $\sum_{t=1}^{T} \binom{nt+\hat{k}-1}{\hat{k}-1}$. The formula indicates that BAMDP state space increases exponentially with $n$ and $T$, commonly referred to as the curse of dimensionality.

Figure 1 depicts the state space transition diagram for one patient and one period (left) and two patients and two periods (right). While the left figure shows the transition diagram pictorially, the right figure enumerates the state and decision space for $t = \{0, 1, 2\}$ in tabular form. A consequence of the curse of dimensionality is the enhanced computational burden necessitating the need for state space approximation methods. In this paper, we use a form of truncated time-horizon to maintain tractability for a practical example (see §5.1.1), but note that other approximation techniques are possible (and are indeed the focus of ongoing work on this model).

### 3.3 General Properties

The following result states that adding the information state preserves the Markovian nature.
Proposition 2. The BAMDP defined as \( \hat{\mathcal{M}} = \{ \mathcal{H}, \mathcal{K}, \mathcal{U}, P_t(h_{t+1}|h_t, d_t), R_t \} \) is an MDP.

Proof. We follow the outline of [Bertsekas 1995, Section 5.1] and show that the addition of the (imperfect) information state reduces the problem to one with perfect state information and preserves the Markovian dynamics, i.e., \( P(h_{t+1}; k_{t+1}|h_1, \ldots, h_t; k_1, \ldots, k_t; u_1, \ldots, u_t) = P(h_{t+1}; k_{t+1}|h_t; k_t; u_t) \).

We omit the proof and refer the readers to [Bertsekas 1995].

Below, we state some properties of our model, using the simple example outlined in §3.2 (two treatments/two outcomes). Analogous results hold for multiple outcomes and treatments. The first result shows that the success probability is a nondecreasing (nonincreasing) function of \( \alpha_t \) (\( \beta_t \)).

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

Proof. 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.

\[
F_p(y|\alpha + 1, \beta) - F_p(y|\alpha, \beta) = \int_0^y \frac{1}{B(\alpha + 1, \beta)} p^\alpha (1 - p)^{\beta - 1} dp - \frac{1}{B(\alpha, \beta)} p^\alpha (1 - p)^{\beta - 1} dp
= I_y(\alpha + 1, \beta) - I_y(\alpha, \beta)
= - y^\alpha (1 - y)^\beta \frac{1}{\alpha B(\alpha, \beta)} \leq 0,
\]

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

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

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

The following two results show how the value function changes if we observe an additional success or failure on a treatment. We use the following notation: \( \alpha_t^{++} = \alpha_t^+ + 1 \), \( \alpha_t = (\alpha_t^+, \alpha_t^-) \), \( \alpha_t^+ = (\alpha_t^{++}, \alpha_t^+) \); similar definitions hold for \( \beta_t^{++}, \beta_t, \) and \( \beta_t^+ \).

Proposition 2. \( S_t(\alpha_t^+, \beta_t) \geq S_t(\alpha_t, \beta_t) \geq S_t(\alpha_t, \beta_t^+) \) for \( j \in \{ A, B \} \).
Proof. We use an induction argument to prove the first inequality. Proof of the second inequality follows uses similar arguments. Consider stage $T-1$. From (3),

$$ S_{T-1}(\alpha_{T-1}^+, \beta_{T-1}) = \sum_{j,j' \in \{A,B\}, j \neq j'} \frac{\alpha_{T-1}^j + 1}{\alpha_{T-1}^j + \beta_{T-1}^j + 1} \frac{\alpha_{T-1}^{j'}}{\alpha_{T-1}^{j'} + \beta_{T-1}^{j'}} $$

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

Now assume the inequality holds true for $t = t+1, \ldots, T-2$, i.e. $S_{t+1}(\alpha_{t+1}^+, \beta_{t+1}) \geq S_t(\alpha_t^+, \beta_t)$. Then $E_{k_{t+1}}[S_{t+1}(\alpha_{t+1}^+, \beta_{t+1})] \geq E_{k_{t+1}}[S_t(\alpha_t^+, \beta_t)]$ as follows.

$$ E_{k_{t+1}}[S_{t+1}(\alpha_{t+1}^+, \beta_{t+1})] = \sum_{k_{t+1}} S_{t+1}(\alpha_{t+1}^+, \beta_{t+1}) Pr(k_{t+1}|\alpha_{t+1}^+, \beta_{t+1}; d_t) $$

$$ \geq \sum_{k_{t+1}} S_{t+1}(\alpha_{t+1}^+, \beta_{t+1}) Pr(k_{t+1}|\alpha_{t+1}^+, \beta_{t+1}; d_t) Pr(k_{t+1}|\alpha_{t+1}^+, \beta_{t+1}; d_t) $$

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 $Pr(k_{t+1}|\alpha_{t+1}^+, \beta_{t+1}; d_t)$ is nondecreasing in $\alpha_t$, and the inequality in the last line follows from the induction argument. Finally,

$$ S_t(\alpha_t^+, \beta_t) = \max_{u_t} \left\{ \frac{\alpha_t^j + 1}{\alpha_t^j + \beta_t^j + 1} \frac{\alpha_{t+1}^j}{\alpha_{t+1}^j + \beta_{t+1}^j} + E_{k_{t+1}}[S_{t+1}(\alpha_{t+1}^+, \beta_{t+1})] \right\} $$

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

To see why $Pr(k_{t+1}^j|\alpha_t^j, \beta_t^j; d_t^j)$ is nondecreasing in $\alpha_t$, note that the probability equals

$$ \int_0^1 Pr(k_{t+1}^j|d_t^j, p_t^j) g(p_t^j|\alpha_t^j, \beta_t^j) dp_t. $$

The first term in the integral is a binomial likelihood and nondecreasing in $p_t^j$. Since $p_t^j$ is nondecreasing in $\alpha_t^j$, the result follows.
Proposition 3. If $\mathbb{E}P_T^j \geq \mathbb{E}P_T^{j'}$, $\mathcal{P}_t(\alpha_t^+,\beta_t) \geq \mathcal{P}_t(\alpha_t,\beta_t) \geq \mathcal{P}_t(\alpha_t^+,\beta_t^+)$ for $j \in \{A,B\}$.

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

(a1): $\Pr(p_T^j > p_T^{j'}) = \Pr(p_T^j > p_T^{j'}|\alpha_T,\beta_T) = \int_0^1 F(p_T^j|\alpha_T^+,\beta_T)g(p_T^{j'}|\alpha_T^+,\beta_T^+)dp_T^{j'}$,

where $F(\cdot)$ denotes the cdf, and $g(\cdot)$ denotes the pdf. (a1) implies the following:

(a2): $\Pr(p_T^j > p_T^{j'}|\alpha_T,\beta_T) \geq \Pr(p_T^j > p_T^{j'}|\alpha_T,\beta_T)$, i.e. $\Pr(p_T^j > p_T^{j'})$ is nondecreasing in $\alpha_T^+$, and

(a3): $\Pr(p_T^j > p_T^{j'}) \geq \Pr(p_T^j > p_T^{j'})$ since $\mathbb{E}p_T^j \geq \mathbb{E}p_T^{j'}$.

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

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

Proof for the second inequality, $\mathcal{P}_t(\alpha_t,\beta_t) \geq \mathcal{P}_t(\alpha_t,\beta_t^+)$, follows from similar arguments, where

the key thing to note is that that $\Pr(p_T^j > p_T^{j'})$ is nonincreasing in $\beta_T^+$.

The following two lemmas highlight useful properties of value function under the health objective. The first result shows that the value function is superadditive in $n$, i.e., the expected patient outcomes are nondecreasing in stage size. The second result shows that given a fixed $N$, spreading patients over a longer period improves expected outcomes. We omit the proofs and instead refer the readers to [Bertsimas and Mersereau (2007), Propositions 2 and 3, respectively].

Lemma 2. For any given $\alpha_0, \beta_0$, and number of patients per periods $n_1$ and $n_2$, $\tilde{S}_0(\alpha_0, \beta_0; n_1 + n_2, T) \geq \tilde{S}_0(\alpha_0, \beta_0; n_1, T) + \tilde{S}_0(\alpha_0, \beta_0; n_2, T)$.

Lemma 3. For any given $\alpha_0, \beta_0$, $N$, number of patients per periods $n_1$ and $n_2$, and trial length $T_1, T_2$, such that $\gamma n_1 = n_2$, $T_1 = \gamma T_2$ for $\gamma > 1$, and $n_1 T_1 = n_2 T_2 = N$, $\tilde{S}_0(\alpha_0, \beta_0; n_1, T_1) \geq \tilde{S}_0(\alpha_0, \beta_0; n_2, T_2)$.

4 Adaptive Designs

Standard response-adaptive designs are sequential, i.e., they randomize patients one at a time. Although such designs, that we call Perfectly Adaptive, offer the greatest learning potential, their applicability in practice is limited. Unless there are minimal observation delays, implementing a
**Perfectly Adaptive** design would result in a prohibitively long clinical trial and potentially deteriorating outcomes for the general patient population due to long approval times.

We propose a design, which we term **Jointly Adaptive** that directly addresses these limitations. A key feature of our proposed design is that it incorporates responses from the whole patient population in determining randomization probabilities, meaning that patients may be allocated to multiple treatments simultaneously.

We compare **Jointly Adaptive** design with other naive (suboptimal) implementable adaptive designs, two of which we describe below. One approach is to consider each patient in isolation at each time period such that past responses from other patients are ignored. Such a design, which we call **Isolated Adaptive**, is akin to having multiple independent clinical trials in isolation, each of which implements a **Perfectly Adaptive** design. In each such isolated trial, 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 a **Jointly Adaptive** design. However, the design is constrained by the fact that all patients in a time period receive the same treatment.

Finally, we compare our proposed design with the traditional fixed design, whose primary objective is to learn about treatment efficacy. We define a new design called **Equal Allocation** (EA), where patients are allocated to treatments in equal proportion. We summarize the designs below.

- \( \Pi_{PA} \) (Perfectly Adaptive): 1 × \( nT \) patients arrive sequentially; decision is made for one patient at a time; incorporates learning from outcomes of all previous patients.

- \( \Pi_{JA} \) (Jointly Adaptive): patients arrive in batches of \( n \) for \( T \) periods; decision is made for all \( n \) patients simultaneously; incorporates learning from outcomes of all previous patients; each patient is randomized to all available treatments.

- \( \Pi_{RA} \) (Restricted Adaptive): patients arrive in batches of \( n \) for \( T \) periods; decision is made for all \( n \) patients simultaneously; incorporates learning from outcomes of all previous patients; all \( n \) patients are restricted to receive the same treatment.

- \( \Pi_{IA} \) (Isolated Adaptive): patients arrive in batches of \( n \) for \( T \) periods, but each patient is considered in isolation; equivalent to \( n \) independent sequential trials each consisting of 1 × \( T \) patients, with no learning across trials.

- \( \Pi_{EA} \) (Equal Allocation): patients arrive in batches of \( n \) for \( T \) periods; patients allocated to each treatment in equal proportion, i.e., each treatment receives \( \frac{n}{|J|} \) patients.

Note that \( \Pi_i \), indicates a class of policies and \( \pi_i \in \Pi_i \) indicates a policy, \( i = \{PA, JA, RA, IA, EA\} \), following dynamic programming terminology [Puterman 1994].

---

9The majority of fixed designs randomize patients equally to the various treatments throughout the trial.
Theorem 1. Given $n$, $T$, $N$, $\alpha_0$, $\beta_0$, the following holds: (i) $V_0^{\pi PA} \geq V_0^{\pi JA}$, (ii) $V_0^{\pi JA} \geq V_0^{\pi RA}$, (iii) $V_0^{\pi JA} \geq V_0^{\pi JA}$, and (iv) $V_0^{\pi JA} \geq V_0^{\pi EA}$.

Proof. We first prove (i). Since $T$ is different in $\pi PA$ and $\pi JA$, we present the notation in terms of patients to provide equivalence, as follows. Let $\tilde{\pi} = (\tilde{u}_1^i, \tilde{u}_2^i, ..., \tilde{u}_N^i)$ represent the controls for patient $m = 1, 2, ..., N$ under policy $\pi_i \in \Pi_i$, $i = \{PA, JA\}$. Similarly let $\tilde{h} = (h_1^i, h_2^i, ..., h_N^i)$ denote the information available for patient $m$ under policy $\pi_i$. For each $m$, $u_{\pi}^{m} : h_{\pi}^{m} \rightarrow \mathcal{P}(U)$, i.e., each $\pi_i \in \Pi_i$ generates a probability distribution on $P^{\pi_i}(\cdot)$ on each $h^{\pi}$.

Under $\pi PA$, patient $m$ arrives at time $t$, while the arrival time for patient $m$ under $\pi JA$ is given as $t : n(t-1) < m \leq nt$, thus capturing the dependence on time. The first patient arrives at $t = 1$.

This implies the following ordering on the information states for each policy:

(a1): under $\pi PA$: for each $m > 1$, $h_{PA}^{m-1} < h_{PA}^{m} < h_{PA}^{m+1}$, and $h_{PA}^{1} < h_{PA}^{2} < ... < h_{PA}^{N}$, while

(b1): under $\pi JA$: for $m = n(t-1) + 1$ or equivalently when $((m-1) \mod n) = 0$, $h_{JA}^{m} > h_{JA}^{m-1}$, and for all $m : n(t-1) < m \leq nt$ or equivalently when $((m-1) \mod n) > 0$, $h_{JA}^{m} > h_{JA}^{m-1}$

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

(a2): when $((m-1) \mod n) = 0$, $h_{PA}^{m} = h_{JA}^{m}$ and $u_{PA}^{m} = u_{JA}^{m}$, and

(b2): when $((m-1) \mod n) > 0$, $h_{PA}^{m} > h_{JA}^{m}$, and $u_{PA}^{m} \supseteq u_{JA}^{m}$.

Together, (a2) and (b2) imply that $u_{PA}^{m} \supseteq u_{JA}^{m}$ and $u_{PA}^{m} \supseteq u_{JA}^{m}$. Since maximization under a restricted control space leads to a suboptimal solution, $V_0^{\pi PA} \geq V_0^{\pi JA}$, as elaborated below.

We reparameterize the value function as follows:

(a3): under $\pi PA$, for each $m = t$, $V_{t}^{\pi PA}(\alpha_t, \beta_t) = V_{t}^{\pi PA}(\alpha_t, \beta_t)$.

(b3): under $\pi JA$: for $m : n(t-1) < m \leq nt$, $V_{t}^{\pi JA}(\alpha_t, \beta_t) = V_{t}^{\pi JA}(\alpha_t, \beta_t)$.

We prove by induction. Consider the case when $m = N$. For health objective $\tilde{V}_{N}^{\pi PA} = \tilde{V}_{N}^{\pi JA} = 0$. For learning objective, $V_{N}^{\pi PA} = V_{N}^{\pi JA} = \max\{Pr(p_N^{PA} > p_N^{JA}), Pr(p_N^{PA} > p_N^{JA})\alpha_N, \beta_N\}$. Thus, $V_{N}^{\pi PA} \geq \tilde{V}_{N}^{\pi JA}$, i.e. the relationship holds for $N$.

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

\[
\tilde{V}_{m}^{\pi PA}(\alpha_m, \beta_m) = \max_{u_{PA}^{m}} \{R_m + \mathbb{E}_{k_{m+1}}[\tilde{V}_{m+1}^{\pi PA}(\alpha_{m+1}, \beta_{m+1})] \}
\]

\[
\geq \max_{u_{JA}^{m}} \{R_m + \mathbb{E}_{k_{m+1}}[\tilde{V}_{m+1}^{\pi JA}(\alpha_{m+1}, \beta_{m+1})] \}
\]

\[
\geq \max_{u_{JA}^{m}} \{R_m + \mathbb{E}_{k_{m+1}}[\tilde{V}_{m+1}^{\pi JA}(\alpha_{m+1}, \beta_{m+1})] \} = \tilde{V}_{m}^{\pi JA}(\alpha_m, \beta_m),
\]

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

For (ii), we revert to the notation in terms of time. Let $\tilde{u} = (\tilde{u}_1^i, \tilde{u}_2^i, ..., \tilde{u}_n^i)$ and $\tilde{h} = (h_1^i, h_2^i, ..., h_n^i)$ represent the controls and information available, respectively for patient $m = 1, 2, ..., n$ at time $t \in \{0, ..., T - 1\}$ under policy $\pi_i \in \Pi_i$, $i = \{JA, IA\}$, such that $u^{\pi}_t : h^{\pi}_t \rightarrow \mathcal{P}(U)$.

For any given $t > 0$ and for each $m = 1, 2, ..., n$,
(a4): under $\pi_{JA}$: $\mathbf{h}^{\pi_{JA}}_{t+1} = \mathbf{h}^{\pi_{JA}}_t + \sum_{m=1}^{n} \mathbf{k}^{\pi_{JA}}_{t+1,m}$ and $\mathbf{h}^{\pi_{JA}}_t = \mathbf{h}^{\pi_{JA}}_0 = \ldots = \mathbf{h}^{\pi_{JA}}_n$, and

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

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

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

$$V^{\pi_{RA}}_{t+1}(\alpha_t, \beta_t) = \max_{\mathbf{u}^{\pi_{JA}}_t} \{ R_t + \mathbb{E}_k [V^{\pi_{JA}}_{t+1}(\alpha_{t+1}, \beta_{t+1})] \}$$

s.t. $d^j_t = n$ or $d^j_t = 0 \forall j$,

$$\leq \max_{\mathbf{u}^{\pi_{JA}}_t} \{ R_t + \mathbb{E}_k [V^{\pi_{JA}}_{t+1}(\alpha_{t+1}, \beta_{t+1})] \} = V^{\pi_{JA}}_t(\alpha_t, \beta_t),$$

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

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

Below, we provide alternate proofs for parts (i) and (ii) for the case of health objective.

(i) $S_0^{\pi_{JA}}(\alpha_0, \beta_0) = S_0^{\pi_{JA}}(\alpha_0, \beta_0; n, T) \leq S_0^{\pi_{FA}}(\alpha_0, \beta_0; 1, nT) = S_0^{\pi_{FA}}(\alpha_0, \beta_0; 1, nT) = S_0^{\pi_{FA}}(\alpha_0, \beta_0)$, and

(ii) $S_0^{\pi_{IA}}(\alpha_0, \beta_0) = nS_0^{\pi_{IA}}(\alpha_0, \beta_0; 1, T) = nS_0^{\pi_{IA}}(\alpha_0, \beta_0; 1, T) \leq S_0^{\pi_{IA}}(\alpha_0, \beta_0; 1, T) = S_0^{\pi_{IA}}(\alpha_0, \beta_0)$,

where the inequality in (i) follows from Lemma 3 and in (ii) follows from Lemma 2.

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

Bounds on objective function value.

(a) **Patient Health** ($S_0^{\pi_{JA}}$): The lower bound (floor) equals $\max_{j \in J} \{ \mathbb{E} p^j_0 \}$ (see, Berry 1978). Intuitively, this is the value obtained from selecting the treatment that has the highest probability of success at the beginning and adhering to it for the remainder of the trial, even if the information during the trial may suggest otherwise. The upper bound (ceiling) equals $\mathbb{E} \{ \max p^j \}$. Intuitively, this means selecting the treatment at every period with the highest known probability of success.

(b) **Learning** ($P_0^{\pi_{JA}}$): A lower bound on the objective value is $\max_{j \in J} \{ \mathbb{E} \max_{p^j} \} \{ \max_{j \notin J} \{ p_0^{j'} \} \}$ or the expected confidence based on initial priors. The upper bound is one, which happens when we know with certainty which treatment has the highest probability of the desired outcome.

Finally, we note that the value of 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, will yield a tighter lower bound. Conversely, a weak initial prior will be heavily influenced by the information obtained during the trial.
4.1 Asymptotic Properties

The following result shows that \textit{Jointly Adaptive} design infers the “superior” treatment w.p.1 in the limit.

**Theorem 2.** \( p_{0j}^{\pi} \xrightarrow{P} 1 \) as \( T \to \infty \) or \( n \to \infty \).

\textit{Proof.} We first show that \( \pi_{J}A \) allocates infinite number of patients to each treatment in the limit (see Lemma 4 below). We complete the proof by showing that any policy with this property achieves the desired result (see Lemma 5 below).

First, we show 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 4.** Let \( \hat{p}^j \) represent the true (unknown) probabilities of success with treatment (equivalently arm) \( j \in J \), such that \( 0 < \hat{p}^j < 1 \), \( \hat{p}^j \neq \hat{p}^{j'} \) and \( \hat{p}^j > \max_{j' \in J \setminus \{j\}} \hat{p}^{j'} \) for all \( j \neq j', j, j' \in J \). Then, for any optimal policy \( \pi \in \Pi \), \( \sum_{t=0}^{T-1} d_t^j(\pi) \to \infty \) w.p.1 as \( T \to \infty \) or \( n \to \infty \).

\textit{Proof.} We prove for \( T \to \infty \) and note that same argument holds for \( n \to \infty \). We first prove for expected maximum successes objective.

Let \( \pi^* \) represent the optimal policy for this objective. Suppose \( \pi^* \) 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 \( j' \) is played i.o. Then, without loss of generality, we can assume that there exists a state \( \hat{h} \) that occurs with probability \( q^h > 0 \), from which arm \( j \) is never chosen by \( \pi^* \). Let \( \hat{p}^j = \mathbb{E}p^j \) represent the expected probability of success with arm \( j \) at state \( \hat{h} \). 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, \( \tau^j \) represents the minimum time starting at state \( \hat{h} \) when \( \hat{p}^j \) exceeds \( \hat{p}^{j'} \), and \( \omega \) is the sample path on which this happens. For any \( \epsilon < \hat{p}^j \), \( P(\tau^j < \infty) = 1 \).

Now, consider the policy \( \pi' \) that follows \( \pi^*(\tau^j + 1) \) in all states except that it chooses arm \( j \) at \( \tau^j \), which occurs with probability at least \( q^h \). Then, \( V_{\pi'}^{T+1} > V_{\pi^*}^{T+1} + q^h\epsilon \), contradicting the optimality of \( \pi^*(\tau^j + 1) \).

The proof for expected maximum learning objective follows similar argument where we note that each of the two objectives is strictly increasing in \( \max_j \hat{p}^j \) for all \( j \in J \).

The following result, consistent with \cite{Ghosal1995} Proposition 1), shows that the optimal design, that tries each treatment infinitely often, identifies the better treatment w.p.1.

**Lemma 5.** Suppose for every \( j \in J \), \( \sum_{t=1}^{T} d_t^j \xrightarrow{T \to \infty} \infty \). Then for \( n < \infty \), \( P\{p_T^j > \max_{j' \in J \setminus \{j\}} p_T^{j'}\} \xrightarrow{T \to \infty} 1 \), and for \( T < \infty \), \( P\{p_T^j > \max_{j' \in J \setminus \{j\}} p_T^{j'}\} \xrightarrow{n \to \infty} 1 \).

\footnotetext[10]{Since \( \pi^* \) never chooses arm \( j \) from \( \hat{h} \), \( \tau^j \) is the time to string of failures with arm \( j' \) until \( \hat{p}^j \) just exceeds \( \hat{p}^{j'} \).}

\footnotetext[11]{With multiple arms, some of which are played finite times, \( \hat{h} \) 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. remain, exemplified by arm \( j' \), remain and we can apply a similar argument.}

16
Proof. As above, we only prove 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) $p_T^j \xrightarrow{P} \mathbb{E} p_T^j$ as $T \to \infty$. From the definition of $\mathbb{E} p_T^j$,

$$
\mathbb{E} p_T^j = \frac{\alpha_0^j + h_T^j}{(\alpha_0^j + \beta_0^j) + nT} = \frac{\alpha_0^j}{nT} + \frac{h_T^j}{nT},
$$
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): $\frac{h_T^j}{nT} \xrightarrow{a.s. T \to \infty} \bar{p}^j$ 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): $\mathbb{E} p_T^j \xrightarrow{T \to \infty} \bar{p}^j$.

Next note that for a random variable $X$ with mean $\mu$ and variance $\sigma^2$, 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 $Var(p_T^j) = \frac{\alpha_T^{j, \pi_i} \beta_T^{j, \pi_i}}{(\alpha_T^{j, \pi_i} + \beta_T^{j, \pi_i})^2(\alpha_T^{j, \pi_i} + 1 + 1)} \xrightarrow{a.s. T \to \infty} 0$, it follows from (a5) that $\Pr(|p_T^j - \mathbb{E} p_T^j| \geq \epsilon) \xrightarrow{T \to \infty} 0$, or equivalently

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

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

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

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

4.1.1 Maximum Gain in Expected Patient Successes: Example.

The following example shows that the flexibility to naturally randomize patients to multiple treatments, inherent in an optimal adaptive design, such as Jointly Adaptive helps to avoid treatment assignment bias and can result in substantial gains in expected outcomes, relative to designs, such as Restricted Adaptive, that do not anticipate learning from multiple simultaneous outcome observations.

Consider a clinical trial with three treatments, $J = \{A, B, C\}$ and $n$ patients per period. Suppose two periods remain in the trial ($T = 2$), and the objective is to maximize expected patient successes. Suppose $p^j \in \{0, 1\}$ for all $j \in J$, i.e., $p^j$ has a two-point distribution and only one of the treatments is successful (correlated outcomes), i.e. $(p_A^1, p_B^0, p_C^0) \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. Further, suppose the priors indicate that each treatment is equally likely to succeed such that $\mathbb{E} p^j$ approximately equals $\frac{1}{3}$ and $\sum_{j \in J} \mathbb{E} p^j = 1$, but with a slight preference such that: $\mathbb{E} p_A^1 - \mathbb{E} p_B^0 = \mathbb{E} p_B^0 = \mathbb{E} p_C^0 + \frac{\epsilon}{2}$ and $\epsilon > 0$ is small. From the binary assumption, any observed outcome perfectly reveals $p^j$. We
first calculate total expected successes under $\pi_{RA}$ as follows, where we drop $\epsilon$ from the expectation calculations.

(a1): In the first period, $\pi_{RA}$ allocates all $n$ patients to treatment $A$ (based on priors) and the expected successes equals $\frac{n}{3}$.

(a2): Observe $p^A$ at the end of first period. If $p^A = 1$ w.p. $\frac{1}{3}$, $\pi_{RA}$ allocates all $n$ patients to treatment $A$ in the second period. If $p^A = 0$ w.p. $\frac{2}{3}$, $\pi_{RA}$ allocates all $n$ patients to treatment $B$, the next preferred treatment. The expected successes equals $\frac{n}{3} + \frac{2}{3} \times \frac{n}{2} = \frac{2n}{3}$, and

(a3): $S^*_{\pi_{RA}} = n$, combining (a1) and (a2).

The expected total patient successes under $\pi_{JA}$ is as follows:

(b1): In the first period, $\pi_{JA}$ randomizes fraction $w^j$ of patients to treatment $j$ such that $w^j \in (0, 1)$ and $\sum_{j \in J} w^j = 1$. The expected successes equals $\frac{n}{3}$.

(b2): Observe all $p^j, j \in J$ at the end of first period. Let $\hat{j} : p^{\hat{j}} = 1, \hat{j} \in J$ be the successful treatment identified after first period. $\pi_{JA}$ allocates all $n$ patients to $\hat{j}$ the second period. The expected successes equals $n$, and

(b3): $S^*_{\pi_{JA}} = \frac{4n}{3}$, combining (b1) and (b2).

Comparing (a3) and (b3), we observe that $\pi_{JA}$ provides a 33.3% gain in patient successes over $\pi_{RA}$. We note that when outcomes are independent (i.e., either each $p^j$ could be successful), the gain that $\pi_{JA}$ provides over $\pi_{RA}$ is lower, at 16.7%.

Finally, compared to $\pi_{EA}$, $\pi_{JA}$ provides an even greater gain of 50%, since $S^*_{\pi_{EA}} = \frac{2n}{3}$ ($\pi_{EA}$ allocates $\frac{n}{3}$ patients to each treatment in both periods); with only 2 treatments, the gain is at most 33.3% [Berry, 1978].

5 Results and Analysis

We perform numerical analyses to demonstrate the value of implementing a Jointly Adaptive design, where we continue to use the simple example in §3.2. We compare under multiple scenarios that vary in $n$, $T$, $N$, and combinations of 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,\frac{1}{2}); (\frac{1}{2},1); (2,6); (1,4)\}, resulting in 91 unique combinations of $\{(\alpha_0^A, \beta_0^A); (\alpha_0^B, \beta_0^B)\}$.

In addition, for the health objective, we compare the Jointly Adaptive design with the heuristic policies, defined below. These heuristics use learning only from past patients in contrast to adaptive designs described in §3 that are forward-looking in nature. The calculation for optimal value function is still obtained as in the case of adaptive design, by backward recursion.

- **Greedy ($\pi_{Gr}$):** This algorithm, analogous to play-the-winner rule 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.

- **GGreedy ($\pi_{GG}$):** This algorithm allocates all the patients at each period to the treatment with
the highest Gittins index, given current information, computed using an infinite horizon and 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** ($\pi_{BK}$): This algorithm, developed in Burnetas and Katehakis (1996), that we named after 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., $j^*_t = \arg\max_j\{BK^j_t\}$. We calculate $BK^j_t$ at each decision point by solving an optimization program as follows:

$$BK^j_t = \max\{p^j_t : I(\hat{p}^j_t, p^j_t) \leq \log\bar{N}_t + \alpha^j_t + \beta^j_t\},$$

where $\bar{N}_t = \sum_{j \in J}(\alpha^j_t + \beta^j_t)$, $\hat{p}^j_t = \frac{\alpha^j_t}{\alpha^j_t + \beta^j_t}$, and $I(x, y) = x\log\frac{x}{y} + (1 - x)\log\frac{1 - x}{1 - y}$ for $x$ and $y$ are Bernoulli distributed. BK is thus, an adjusted version of the estimator $\hat{p}^j_t$ in which the adjustment decreases with the number of patients already allocated to 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** ($\pi_{UC}$): This algorithm, developed by Auer et al. (2002), allocates all the patients at each period to the treatment with the highest value of upper confidence index, i.e., $j^*_t = \arg\max_j\{UCB1^j_t\}$, calculated at each decision point as follows: $UCB1^j_t = h^j_t + \sqrt{\frac{2\log nt}{\sum_{t=0}^d d^j_t}}$, 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 Jointly Adaptive design relative to other implementable designs: (a) Equal Allocation design, (b) naive adaptive designs: Restricted and Isolated Adaptive designs, and (c) heuristics: Greedy, GGreedy, BK, and UCB1, as well as the disadvantage relative to the non-implementable Perfectly Adaptive design. Berry (1978) shows numerically that $\pi_{JA}$ performs better than $\pi_{EA}$ for 25 and 50 patient observations.

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 compared Greedy and all adaptive designs perform better than the fixed design but other heuristics do not always under these scenarios.

Table 1 also shows that Jointly Adaptive design performs best amongst all implementable designs, as it must since it is optimal for these conditions. In particular, the Jointly Adaptive design performs substantially better than Equal Allocation design, and almost as well as the non-implementable
Table 1: Expected proportion of successes for a variety of problem scenarios.

<table>
<thead>
<tr>
<th>((a, b))</th>
<th>((\alpha, \beta))</th>
<th>(n)</th>
<th>(T)</th>
<th>(N)</th>
<th>FIXED</th>
<th>ADAPTIVE</th>
<th>HEURISTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>0.5000</td>
<td>0.6725</td>
<td>0.6215</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>6</td>
<td>24</td>
<td>0.5000</td>
<td>0.6259</td>
<td>0.6132</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>12</td>
<td>48</td>
<td>0.5000</td>
<td>0.6393</td>
<td>0.6333</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>24</td>
<td>96</td>
<td>0.5000</td>
<td>0.6487</td>
<td>0.6460</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>(1, 1)</td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>0.6667</td>
<td>0.6679</td>
<td>0.6679</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>(1, 1)</td>
<td>2</td>
<td>24</td>
<td>48</td>
<td>0.6667</td>
<td>0.6693</td>
<td>0.6693</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>12</td>
<td>48</td>
<td>0.6667</td>
<td>0.6693</td>
<td>0.6693</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>24</td>
<td>96</td>
<td>0.6667</td>
<td>0.6709</td>
<td>0.6707</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>6</td>
<td>24</td>
<td>0.0000</td>
<td>0.2532</td>
<td>0.2495</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>(1, 1)</td>
<td>4</td>
<td>12</td>
<td>48</td>
<td>0.0000</td>
<td>0.2632</td>
<td>0.2614</td>
</tr>
<tr>
<td>(4, 4)</td>
<td>(4, 4)</td>
<td>4</td>
<td>6</td>
<td>24</td>
<td>0.5000</td>
<td>0.5538</td>
<td>0.5480</td>
</tr>
<tr>
<td>(4, 4)</td>
<td>(4, 4)</td>
<td>4</td>
<td>12</td>
<td>48</td>
<td>0.5000</td>
<td>0.5644</td>
<td>0.5614</td>
</tr>
<tr>
<td>(4, 4)</td>
<td>(4, 4)</td>
<td>4</td>
<td>24</td>
<td>96</td>
<td>0.8000</td>
<td>0.8001</td>
<td>0.8001</td>
</tr>
<tr>
<td>(4, 4)</td>
<td>(4, 4)</td>
<td>4</td>
<td>24</td>
<td>96</td>
<td>0.8000</td>
<td>0.8744</td>
<td>0.8724</td>
</tr>
<tr>
<td>(4, 1)</td>
<td>(4, 1)</td>
<td>2</td>
<td>24</td>
<td>48</td>
<td>0.6667</td>
<td>0.6984</td>
<td>0.6980</td>
</tr>
<tr>
<td>(4, 1)</td>
<td>(4, 1)</td>
<td>4</td>
<td>24</td>
<td>96</td>
<td>0.5000</td>
<td>0.6516</td>
<td>0.6499</td>
</tr>
</tbody>
</table>

**Perfectly Adaptive** design. For example, \(\pi_{JA}\) increases patient successes by 29.2% compared to \(\pi_{EA}\) 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, \(\pi_{Gr}\) 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 percentage gain) between various designs. The table also lists the average and maximum values across 91 combinations of initial priors under the considered scenarios.

Figure 2 illustrates the variation of \(\delta_{JI}\), \(\delta_{JR}\), and \(\delta_{PJ}\) with (a) \(n\), keeping \(T\) fixed (left chart), (a) \(T\), keeping \(n\) fixed (middle chart), and (c) \(n \times T\), keeping \(N\) fixed (right chart). Some key observations from the figure are, first, the gains in successes that \(\pi_{JA}\) provides over \(\pi_{RA}\) and \(\pi_{IA}\) 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 diminished 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 Truncated-horizon approximation.

As described earlier, the adaptive design setup suffers from the the curse of dimensionality, implying that as the trial size increases, solving the fully enumerated problem becomes computationally burdensome. In a forthcoming paper (currently work in progress), we propose a state space approx-

---

Even though the percentages appear small, the absolute value of the improvement can be substantial. For example, a 2% gain in expected successes in a trial consisting of 500 patients translates into 10 potential lives saved.
Table 2: Definition of quantities comparing various designs and their values (mean, maximum) across 91 initial priors when \((n, N) = (4, 24)\).

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Definition</th>
<th>Description</th>
<th>Average</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta_{PJ})</td>
<td>(\frac{\pi_{JA} - \pi_{PA}}{\pi_{JA}})</td>
<td>Perfectly vs. Jointly Adaptive</td>
<td>0.85%</td>
<td>2.66%</td>
</tr>
<tr>
<td>(\delta_{JI})</td>
<td>(\frac{\pi_{JA} - \pi_{IA}}{\pi_{JA}})</td>
<td>Jointly vs. Isolated Adaptive</td>
<td>2.54%</td>
<td>8.64%</td>
</tr>
<tr>
<td>(\delta_{JR})</td>
<td>(\frac{\pi_{JA} - \pi_{RA}}{\pi_{JA}})</td>
<td>Jointly vs. Restricted Adaptive</td>
<td>0.26%</td>
<td>2.61%</td>
</tr>
<tr>
<td>(\delta_{Gr})</td>
<td>(\frac{\pi_{JA} - \pi_{Gr}}{\pi_{JA}})</td>
<td>Jointly Adaptive vs. Greedy</td>
<td>0.48%</td>
<td>6.03%</td>
</tr>
<tr>
<td>(\delta_{BK})</td>
<td>(\frac{\pi_{JA} - \pi_{BK}}{\pi_{JA}})</td>
<td>Jointly Adaptive vs. BK</td>
<td>2.13%</td>
<td>8.78%</td>
</tr>
<tr>
<td>(\delta_{GG})</td>
<td>(\frac{\pi_{JA} - \pi_{GG}}{\pi_{JA}})</td>
<td>Jointly Adaptive vs. GGreedy</td>
<td>3.61%</td>
<td>10.66%</td>
</tr>
<tr>
<td>(\delta_{UC})</td>
<td>(\frac{\pi_{JA} - \pi_{UC}}{\pi_{JA}})</td>
<td>Jointly Adaptive vs. UCB1</td>
<td>4.05%</td>
<td>10.61%</td>
</tr>
</tbody>
</table>

Figure 2: \(\delta_{JI}\), \(\delta_{JR}\), and \(\delta_{PJ}\) as a function of \(n\), \(T\) and \(n \times T\) with noninformative initial priors.

imation technique that addresses this computational challenge. For use in this paper, however, we propose a truncated-horizon or limited-lookahead approximation method, 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 §5.5.

In such an approach, we implement the optimal Jointly Adaptive design over a shorter number of observations that we term as \(N_{\text{short}}\) and, thereafter, use a myopic design. The expected number of successes is a combination of the optimal solution from the first \(N_{\text{short}}\) observations and a myopic solution from the remaining \((N - N_{\text{short}})\) observations. Any solution thus obtained provides a lower bound on the optimal solution. Further, this simplification results in little loss relative to a fully optimal solution. Our choice of a myopic design for the truncated-horizon approximation approach is the Greedy heuristic, described above, since it performs best among the heuristics we considered.

We illustrate the usefulness of the truncated-horizon approximation using a numerical example where \((n, N) = (4, 96)\) and \(N_{\text{short}}\) ranges from 0 to 96. Note that \(N_{\text{short}} = 0\) refers to a fully Greedy design and \(N_{\text{short}} = 96\) refers to a fully Jointly Adaptive design. Figure 3 shows the average gain in expected successes as a result of implementing truncated-horizon 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 multi-armed bandit property that the truncated-horizon approximation exploits.
Figure 3: Average gain in expected successes over a purely myopic approach as a function of \( N_{short} \) in truncated-horizon approximation, \((n, N) = (4, 96)\).

Figure 4: \( \delta_{Pr} \) as a function of \( T \) with \( n = 4 \) and noninformative initial priors.

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 \( \pi_{JA} \) vs. \( \pi_{EA} \), where we capture the difference between the two in the following quantity:

\[
\delta_{Pr} = \frac{\bar{P}_{\pi_{JA}0}(\alpha_0, \beta_0; n, T) - \bar{P}_{\pi_{EA}0}(\alpha_0, \beta_0; n, T)}{\bar{P}_{\pi_{EA}0}(\alpha_0, \beta_0; n, T)}.\]

Figure 4 plots \( \delta_{Pr} \) as a function of \( T \) when \( n = 4 \) and initial priors are noninformative. We observe that \( \delta_{Pr} \) is relatively small and quickly decreases to zero as \( T \) increases. We observe a maximum gain of 2.2% under the considered scenarios. Finally, we note that \( \bar{P}_{\pi_{JA}0} \) increases in both \( n \) and \( T \).

Managerial Insight. We have shown examples that highlight the extent of objective improvement from using \( \pi_{JA} \) over \( \pi_{EA} \) for varying parameter combinations. These observations also imply that \( \pi_{JA} \) would use same or fewer patients compared to \( \pi_{EA} \) to achieve a target objective function value.

Let \( y \in (0, 1) \) be the target learning objective. Given \( n \), let \( E_{T}\pi_{EA} = \{ \min T : \bar{P}_{\pi_{EA}}(\alpha_0, \beta_0; n, T) \geq y \} \) and \( E_{T}\pi_{JA} = \{ \min T : \bar{P}_{\pi_{JA}}(\alpha_0, \beta_0; n, T) \geq y \} \), where we note that each calculation of \( \bar{P}_{\pi_{JA}} \) is done with a fixed \( T \). Then, \( \frac{E_{T}\pi_{EA} - E_{T}\pi_{JA}}{E_{T}\pi_{EA}} \) represents the reduction in expected number of patients. To illustrate with an example, for \( y = 0.93 \), \( n = 4 \), and noninformative initial priors, we obtain \( E_{T}\pi_{EA} = 20 \) and \( E_{T}\pi_{JA} = 19 \). In other words, Jointly Adaptive design would use 5% fewer
patients in expectation to achieve a 93% learning objective target compared to a fixed design. These savings do not include the 28% 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:

$$\delta_{Lh} = \frac{P_0^{\pi JA}(LE) - P_0^{\pi JA}(PH)}{P_0^{\pi JA}(LE)}; \quad \delta_{Hl} = \frac{S_0^{\pi JA}(PH) - S_0^{\pi JA}(LE)}{S_0^{\pi JA}(PH)}.$$

Here, $P_0^{\pi JA}(LE)$ and $P_0^{\pi JA}(PH)$ refer to the values when the objective being maximized is learning (LE) and patient health (PH), respectively, where we note that $P_0^{\pi JA}(LE)$ is the optimal value. The terms $S_0^{\pi JA}(LE)$ and $S_0^{\pi JA}(PH)$ are defined similarly.

Table 3 lists the quantities defined above as a function of $N$ when $n = 4$ and initial priors are noninformative. We observe that $\delta_{Lh}$ is small compared to $\delta_{Hl}$, 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, $\delta_{Lh}$ decreases while $\delta_{Hl}$ increases, (b) this rate of change in $\delta_{Lh}$ is smaller compared to that of $\delta_{Hl}$. Together, this further highlights the benefits of implementing the optimal adaptive design.

#### 5.4 Numerically estimating clinician’s rate of learning

We numerically estimate the rate at which clinician’s initial beliefs about the success probabilities of treatments converge to the true probability values. Our estimation procedure uses simulation as follows. Starting with some initial priors, the random patient outcomes are generated according to true underlying probability distributions, that 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 “true” values, captured in the following quantity:

---

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 §1).

14 Our work provides a sense of tradeoff if decision-makers were to consider a weighted objective function (weights to health and learning objectives); doing so would require solving a constrained optimization problem.

---
Figure 5: $\epsilon_t$ vs. $t$ for various $\{(\alpha^A_0, \beta^A_0); (\alpha^B_0, \beta^B_0); T\}$ 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)

$$\epsilon_t = \frac{1}{|J|} \sum_{j \in J} |\bar{p}_j^t - E p_j^t|,$$

where $\epsilon_t$ is the Absolute Error at time $t$, $\bar{p}_j^t$, represents the true success probability of treatment $j \in \{A, B\}$ and $E p_j^t$ has been described before. The convergence rate is then the decay of $\epsilon_t$ with $t$.

Given $\bar{p}_j^t$ is unknown, we assume it for this numerical study. We consider various combinations of initial priors and trial length; for each such combination, we plot $\epsilon_t$ vs. $t$ for 91 pairs of $(\bar{p}_A^t, \bar{p}_B^t)$. Figure 5 plots three such combinations. We observe the convergence rate follows similar pattern and is approximately proportional to $t^{-\frac{1}{2}}$, consistent with Ghosal (2010).

### 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 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. 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 versus 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. (2011b) and Chimowitz et al. (2011a).

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

**Implementation.** First, we choose initial priors for both treatments. For PTAS treatment, our choice of priors yield the expected failure probability of 4.44%, as observed in the previous 45-patient trial on the same stent. For the standard treatment, we choose strong priors that result in an expected failure probability of 5.73%, 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 failures in the SAMMPRIS trial with the *Jointly Adaptive* design by first solving for the optimal policy and subsequently implementing it using the SAMMPRIS conditions with known and 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 \(\pi_{JA}\) instead of \(\pi_{EA}\).

We could directly solve for fully optimal policies for problem size up to \(N = 240\) but memory limitations precluded direct solutions of larger problems. Hence, we employed a truncated-horizon approximation approach that results in minimal deviations from optimality (see §5.1.1). We find that implementing this design with \(N = 451\) and \(N_{short} = 240\) results in 28.8 expected failures, a reduction of over 37% 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 \(N_{short}\), as would be expected. We extrapolate from the difference between the results on expected number of failures with \(N_{short} = 180\) and those with \(N_{short} = 240\), and estimate that a fully optimal design for \(N = 451\) would result in no fewer than 28.6 expected failures. This means that the truncated-horizon approximation solution objective would be within at most 0.73% of full optimality.

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 \(N_{short}\), as would be expected. We extrapolate from the difference between the results on expected number of failures with \(N_{short} = 180\) and those with \(N_{short} = 240\), and estimate that a fully optimal design for \(N = 451\) would result in no fewer than 28.6 expected failures. This means that the truncated-horizon approximation solution objective would be within at most 0.73% of full optimality.

Given \(N_{short} = 240\) and \(n = 4\) implies that clinicians have the opportunity to update their beliefs (and actions) \(T_{short} = 240/4 = 60\) times, during the trial. It may not be possible to update so often

---

16 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-arm trial, which enrolled 45 patients and tested only the PTAS treatment for efficacy, resulted in a much lower failure probability of 4.44% (FDA [2005]).

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

18 We ran our code both on a personal Macintosh computer (4GB 667Hz DDR2 SSDRAM, 250 GB HD) as well as on the research grid at Chicago Booth.

19 While truncated-horizon 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.
Table 4: Expected total failures (number and %), associated standard error, and size of the PTAS group with $\pi_{JA}$ as a function of $n$ when $N_{\text{short}} = 240$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Expected Number of Failures</th>
<th>Expected Failure Rate</th>
<th>Standard Error</th>
<th>PTAS size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.55</td>
<td>6.33%</td>
<td>0.0004%</td>
<td>39.25</td>
</tr>
<tr>
<td>2</td>
<td>28.64</td>
<td>6.35%</td>
<td>0.0007%</td>
<td>40.57</td>
</tr>
<tr>
<td>3</td>
<td>28.72</td>
<td>6.37%</td>
<td>0.0009%</td>
<td>41.72</td>
</tr>
<tr>
<td>4</td>
<td>28.81</td>
<td>6.39%</td>
<td>0.0009%</td>
<td>43.02</td>
</tr>
</tbody>
</table>

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

Figure 6: Expected total failures with $\pi_{JA}$ as a function of $T_{\text{short}}$ (left), and the distribution of failures when $n = 4$ (right).

Notes: In the right chart, (a) the failures are grouped in buckets of 1, for example the probability of observing failures between 25-26, (b) we only include the number of failures (x-axis) whose probability is at least $10^{-8}$.

(e.g., due to logistical reasons). Figure 6 (left chart) shows the total expected failures with $\pi_{JA}$ as a function of $T_{\text{short}}$. As expected, increasing $T_{\text{short}}$ results in more frequent use of the optimal policy and hence a reduction in failure rate. Figure 6 (right chart) shows the probability distribution of failures with $\pi_{JA}$ 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.

6 Discussion

Traditional clinical trial designs assign patient to treatments in a fixed proportion throughout the trial. The primary purpose of such a design is to learn; varying treatment assignment to improve patient outcomes is not a usual consideration. Response-adaptive designs, which allow clinicians to learn about treatments from patient responses during the trial and adjust patient allocation 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 treated 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 consists of a single patient every
period and there are no delays in observing outcomes, such a design is rendered impractical.

We address this gap by explicitly considering delays in observing outcomes while building on the Bayesian approach. We propose a *Jointly Adaptive* design that, at each time period learns from multiple patients and simultaneously randomizes them to multiple treatments. The key contribution of this paper is 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 *Jointly Adaptive* designs over existing designs can be significant. In addition, using the conditions of the SAMMPRIS trial, we showed that a truncated-horizon 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. 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 et al. (2009) report that protocols for about 20% of trials (34% for phase I/II trials) conducted at MD Anderson in 2000-05 used Bayesian statistical designs. Examples of such trials can be found in Krams et al. (2003), 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 *Jointly Adaptive* design and potential for substantial efficiency gains. This 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 same set of patients are being reallocated. 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. The two trials highlighted in our study - SAMMPRIS (outcome: death/major stroke), and CATIE (outcome: treatment discontinuation), highlight the versatility of our design. Finally, facilities with good information technology and logistical infrastructure are ideal for implementation of adaptive design; poor infrastructure and lack of understanding of complex

---

20Carryover 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.
statistical methodologies has been highlighted as a key barrier to adaptive design implementation (AptivSolutions, 2012; Stadler, University of Chicago).

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 exists in trials where same patients are allocated over multiple time periods. van der Graaf et al. (2012) further discusses how certain features of adaptive trials may create some potential scientific and ethical challenges.

**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 such as CATIE. 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 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. 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 three anonymous referees, an associate editor, and area editor Pinar Keskinocak.

**References**


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


Dr. Walter M. Stadler (University of Chicago). Personal conversation, November 2013.

