# Sequential Patient Recruitment and Allocation for Adaptive Clinical Trials

**Onur Atan**
UCLA

**William R. Zame**
UCLA

**Mihaela van der Schaar**
University of Cambridge,
UCLA , Alan Turing Institute

## Abstract

Randomized Controlled Trials (RCTs) are the gold standard for comparing the effectiveness of a new treatment to the current one (the control). Most RCTs allocate the patients to the treatment group and the control group by uniform randomization. We show that this procedure can be highly suboptimal (in terms of learning) if – as is often the case – patients can be recruited in cohorts (rather than all at once), the effects on each cohort can be observed before recruiting the next cohort, and the effects are heterogeneous across identifiable subgroups of patients. We formulate the patient allocation problem as a finite stage Markov Decision Process in which the objective is to minimize a given weighted combination of type-I and type-II errors. Because finding the exact solution to this Markov Decision Process is computationally intractable, we propose an algorithm – *Knowledge Gradient for Randomized Controlled Trials* (RCT-KG) – that yields an approximate solution. Our experiment on a synthetic dataset with Bernoulli outcomes shows that for a given size of trial our method achieves significant reduction in error, and to achieve a prescribed level of confidence (in identifying whether the treatment is superior to the control), our method requires many fewer patients.

## 1 Introduction

Randomized Controlled Trials (RCTs) are the gold standard for evaluating new treatments. Phase I trials are used to evaluate safety and dosage, Phase II trials are used to provide some evidence of efficacy, and Phase III trials (which are the subject of this paper) are used to evaluate the effectiveness of the new treatment in comparison to the current one. RCTs are useful because they create a treatment group and a control group that are as similar as possible except for the treatment used. Most RCTs recruit patients from a prescribed target population and uniformly randomly assign the patients to treatment groups using repeated uniform randomization (perhaps adjusted to deal with chance imbalances). This approach is optimal if all patients are recruited at once or if the outcomes for previous patients cannot be observed when recruiting new patients or if the patient population is (or is thought to be) homogeneous. However, in many circumstances, patients are (or can be) recruited in cohorts, the outcomes for patients in previous cohorts can be observed when recruiting a new cohort, and the population contains identifiable subgroups for which differences in effects might be expected. (For example, different effects of treatment might be expected for patients with different genetic mutations [Moss et-al. (2015)]; see [Hébert et al.(1999)] and Table 1 for additional examples.) In such situations, the information learned from previous cohorts can be used in recruiting and allocating patients in the new cohort: the optimal policy should not necessarily recruit the same number of patients to each identifiable subgroup or allocate equal numbers of patients within a subgroup to the treatment and the control. We illustrate this point in the Experiments.

Our goal in this paper is to develop a procedure that prescribes two things: (i) the number of patients in each cohort to recruit from each subgroup, (ii) the allocation of these patients to treatment or control in order to minimize the error (type-I or type-II or a given convex combination) in identifying the patient subgroups for which the treatment is more/less effective than the control. Our work differs from recent work on Bayesian clinical trials [Berry et al.(2006)] in which the the information obtained from the previous cohorts are used only for treatment allocation; in our work, the infor-

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

mation from previous cohorts is used both for patient recruitment and for allocation to treatment or control. As an example, consider the RCT setting in [Barker et al.(2009)] for neoadjuvant chemotherapy in which two subgroups are identified based on the hormone receptor status, human epidermal growth factor receptor 2 (HER2) status, and MammaPrint11,12 status. If our procedure were to be used in this actual trial to recruit 100 patients in each cohort, the initial cohort would consist of 50 patients from each subgroup and would allocate them uniformly to treatment/control. However, in the second and succeeding cohorts, we would use the observed outcomes from earlier cohorts to recruit more patients from the subgroup in which uncertainty about treatment efficacy was larger and, within each subgroup, we would allocate more patients to whichever of treatment/control had displayed larger variance.

Our first contribution is to formalize the learning problem as a finite stage Markov Decision Problem (MDP) in which the designer recruits $N$ patients over $K$ steps and observes the outcomes of step $k-1$ before taking step $k$. However, because the action and state spaces of this MDP are very large, solving this MDP by dynamic programming is computationally intractable. We therefore propose a computationally tractable greedy algorithm *Knowledge Gradient for Randomized Controlled Trials* (RCT-KG) that yields an approximate solution. We illustrate the effectiveness of our RCT-KG algorithm in a set of experiments using synthetic data in which outcomes are drawn from a Bernoulli distribution with unknown probabilities. In particular, we show that, keeping the sizes of the trial and of the cohorts fixed, RCT-KG yields significantly smaller expected error; conversely, in order to achieve a given level of confidence in identifying whether the treatment is better than the control, RCT-KG requires many fewer patients/cohorts.

Our approach makes a number of assumptions. The first is that patients can be recruited in cohorts, and not all at once. The second is that the outcomes for patients in each cohort are realized and can be observed before recruiting and allocating patients for the succeeding cohort. The third is that subgroups are identified in advance. The fourth is that, for each cohort (after the first), the number of patients recruited in each subgroup and the allocation of patients (to treatment or to control) within each subgroup can be chosen to depend on the observations made from previous cohorts. These assumptions are strong and certainly are not satisfied for all RCTs, but they are satisfied for *some* RCTs, and for those our approach offers very significant improvements over previous approaches to the speed and accuracy of learning.

## 2 Related Work

The most commonly used procedure to allocate patients into treatment and control groups is "repeated fair coin-tossing" (uniform randomization). One potential drawback to this approach is the possibility of unbalanced group sizes when the set of patients (or the set of patients in an identified subgroup) is small [Friedman et al.(1998)]. Hence, investigators often follow a restricted randomization procedure for small RCTs, such as *blocked randomization* [Lachin et al.(1988)] or *adaptive bias-coin randomization* [Schulz and Grimes(2002)]. These procedures have the effect of assigning more patients to treatment/control groups to prevent an imbalance between the groups. An alternative, but less frequently used procedure is *covariate-adaptive randomization* in which patients are assigned to treatment groups to minimize *covariate imbalance* [Moher et al.(2012)]. These approaches are efficient when patients are recruited at one time or the outcomes of previous cohorts are not observed . However, as we have noted and can be seen in Table 1, it is often the case that patients are recruited sequentially in cohorts and the outcomes of previous cohorts can be observed before recruiting the next cohort. There is a substantial literature concerning such settings [Lewis and Bessen(1990), Whitehead(1997)], but it focuses on the decision of whether to terminate the trial, rather than how to allocate the next cohort of patients to the treatment or control groups, which is the focus of our paper.

*Response adaptive randomization* uses information about previous cohorts to allocate patients in succeeding cohorts: the probability of being assigned to a treatment group is increased if responses of prior patients in that particular group has favorable [Hu and Rosenberger(2006), Berry et al.(2006)]. However, the goal of the most of these approaches is to improve the benefit to the patients in the trial rather than to learn more about the comparison between the treatment and the control. Multi-Armed Bandits (MABs) constitute a general mathematical decision framework for resource (patient) allocation with the objective of maximizing the cumulative outcomes (patient benefits) [Auer et al.(2002), Gittins et al.(2011), Agrawal and Goyal(2012)]. However, [Villar et al.(2015)] shows that although MABs achieve greater patient benefit, they suffer from poor learning performance because they allocate most of the patients to favorable actions (treatments). Hence, [Villar et al.(2015)] proposes a variation on MAB algorithms that allocates patients to control action at times to maintain patient benefit while mitigating the poor learning performance. Some of the existing work on Bayesian RCTs [Berry et al.(2006)] does use the information from previous cohorts to allocate patients in order to improve the

| Study | Size | Treatment | Primary Outcome | Result |
|---|---|---|---|---|
| [Hacke et al.(1995)] | 620 | rt-PA (alteplase) | Barthel Index (BI) at 90 days | Treatment is effective in improving the outcome in a defined subgroup of stroke patients. |
| [Hébert et al.(1999)] | 838 | Red cell transfusion | 30 days mortality | effective among the patients with Apache 2 score less than equal to 20 and age less than 55. |
| [Hacke et al.(2008)] | 821 | Intravenous thrombolysis | disability at 90 days | As compared with placebo, intravenous alteplase improved clinical outcomes significantly. |
| [Moss et-al. (2015)] | 69 | Ivacaftor | ppFEV1 in week 24 | Ivacaftor significantly improves lung function in adult patients with R117H-CFTR. |

Table 1: RCT Examples in the Literature

learning performance. The work closest to ours may be [Bhatt et al.(2006)] which specifically addresses the special case of a trial with two identified subgroups. However the only adaptation considered is to entirely cease recruiting patients into one of the subgroups, not to change the number of patients recruited into each subgroup. Our paper prescribes a principled algorithm to improve the learning performance that adjusts both the number of patients recruited into each subgroup and the proportion of patients allocated to treatment/control within each subgroup.

Another line of literature relevant to ours is family of policies known as Optimal Computing Budget Allocation (OCBA) [Chen(1995), Chen et al.(2003)]. These policies are derived as an optimization problem to maximize the probability of later identifying the best alternatives by choosing the measurements. The approach in these policies is by approximating the objective function with lower and upper bounds. In this paper, we provide a Bayesian approach to RCT design and model the problem as a finite stage MDP. The optimal patient allocation policy can be obtained by solving the Dynamic Programming (DP); for the case of discounting, the Gittins Index [Gittins et al.(2011), Whittle(1980)] provides an (optimal) closed form solution to the MDP. However, both of these approaches are computationally intractable for problems as large as typical RCTs.

Knowledge Gradient (KG) policies provide an approximate, greedy approach to the MDP. However, in our setting, the action space for the MDP, which contains all possible allocation choices, is very large and hence KG policies are again not computationally tractable. (Morever, KG policies typically assume multivariate normal priors for the measurements [Frazier et al.(2008), Frazier et al.(2009)].). [Chen et al.(2013), Chen et al.(2015)] proposes a variation of KG policies that they deemed Opt-KG and that selects the action that may generate maximum possible reduction in the errors. In our setting, their approach would require that patients be recruited one at a time and that the outcome of the action (treatment or control) chosen for each patient from each subgroup be observable before the next patient is recruited. Those requirements are not appropriate for our setting, in which we observe only a (noisy) signal about the true outcome of the selected treatment and we recruit patients in cohorts. not one at a time. (It would typically be completely impractical to allocate patients one at a time: for any realistic time to observation and total number of patients, doing so would result in a clinical trial that would last for many years.) Moreover, our approach allow for a much broader class of outcome distributions (exponential families, which includes Bernoulli distributions and many others) and we allocate all patients in a cohort at each stage. Our RCT-KG algorithm might be viewed as generalizing Opt-KG in all of the aspects mentioned.

## 3 A Statistical Model for a Randomized Clinical Trial

Our statistical model for a RCT has four components: the patient population, the given patient subgroups, the treatments and the treatment outcomes. Write $\mathcal{W}$ for the patient population and $X$ for a prescribed partition of $\mathcal{W}$ into patient subgroups. Write $Y = \{0, 1\}$ for the action space; 0 represents the control action and 1 represents the treatment action. Let $Z$ be the outcome space; without much loss we assume $Z \subset \mathbb{R}$ with minimum $z_{\min}$ and maximum $z_{\max}$. ($Z$ might be continuous or discrete.)

We wish to allocate a total of $N$ patients in $K$ steps/cohorts over a total time $T$ in order to identify the more efficacious treatment for each subgroup. Throughout we make the following assumptions:

1. The outcome distribution belongs to the exponential family.
2. The outcomes for patients in each cohort are realized before making the allocation decision for the succeeding cohort.

In the next subsection, we give a brief description of the exponential family of distributions and Jeffrey's prior on their parameters.

### 3.1 Exponential Families and Jeffrey's Prior

Let $\Theta$ be a parameter space. Fix functions $G : Z \to \mathbb{R}^d$ and $h : Z \to \mathbb{R}$. The $d$-dimensional parameter exponential family with sufficient statistic $G$ and parametrization $\theta$, relative to $h$, is the family of distributions defined by the following densities: $p(z|\theta) = \Phi(\theta)h(z)\exp(\theta{\cdot}G(z))$ where $\Phi(\theta)$ is uniquely determined by the requirement that $p(\cdot|\theta)$ is a probability density; hence $\Phi(\theta) = \left(\int_{-\infty}^{\infty} h(z)\exp(\theta{\cdot}G(z))d(z)\right)^{-1}$ An alternative expression is:

$$p(z|\theta) = h(z)\exp(\theta{\cdot}G(z) - F(\theta))$$

where $F(\theta) = -\log\Phi(\theta)$. Write $\mu$ for the expectation: $\mu(\theta) = \mathbb{E}_{Z|\theta}[Z]$. Different choices of the various ingredients lead to a wide variety of different probability densities; for example choosing $G(z) = z, h(z) = 1, \theta = \ln\frac{q}{1-q}$ generates a Bernoulli distribution. Other choices lead to Poisson, exponential and Gaussian distributions.

In this paper, we want a Bayesian "non-informative" prior that is invariant under re-parametrization of the parameter space. To be specific, we use Jeffrey's prior, which is proportional to the square root of the Fisher information $I(\theta)$. In the case of the exponential family, the Fisher information is the second derivative of the normalization function, i.e., $I(\theta) = \sqrt{|F''(\theta)|}$. Under Jeffrey's prior, the posterior on the parameter $\theta$ after $n$ observations is given by $p(\theta|z_1,\ldots,z_n) \propto \sqrt{|F''(\theta)|}\exp(\sum_{i=1}^n \theta{\cdot}G(z_i) - nF(\theta))$ Given $n$ outcomes $(z_1, z_2, \ldots, z_n)$, we can summarize the information needed for the posterior distribution by $s = [s_0, s_1] = [\sum_{i=1}^n G(z_i), n]$. The posterior is then proportional to $\sqrt{|F''(\theta)|}\exp(\theta{\cdot}s_0 - s_1F(\theta))$.

Given a subgroup $x \in X$ and a treatment $y \in Y$ Let $\theta_{x,y}$ be the true parameter for subgroup $x$ and treatment $y$. (Of course $\theta_{x,y}$ is not known.) The result of treatment $y$ on a patient in subgroup $x$ is referred to as the *treatment outcome* and is assumed to be drawn according to the exponential family distribution, i.e., $Z \sim p(\cdot|\theta_{x,y})$. Write $\mu(\theta_{x,y})$ for the true expected outcome of the treatment $y$ on subgroup $x$ and define the *treatment effect* for the parameters $\theta_0, \theta_1$ as $E(\theta_0, \theta_1) = \mu(\theta_1) - \mu(\theta_0)$; the treatment effect on subgroup $x$ is $E(\theta_{0,x}, \theta_{1,x})$.

### 3.2 Treatment Effectiveness

Given a threshold $\tau \geq 0$ (set be the designer), we define

$$\nu(x) = \begin{cases} 1 & \text{if } \frac{E(\theta_{0,x},\theta_{1,x})}{\mu(\theta_{x,0})} \geq \tau \\ 0 & \text{if otherwise} \end{cases}$$

so $\nu(x) = 1$ if the treatment is sufficiently better than the control, in which case we say the treatment is *effective*. (For example, see [Farrar et al.(2000)], in which the goal is to identify whether reduction in pain by the treatment with respect to control is more than 33%.) We define the *positive set* $H^+ = \{x \in X : \nu(x) = 1\}$ to be the set of subgroups for which the treatment is effective and the *negative set* $H^- = X \setminus H^+$ to be the complementary set of subgroups for which the treatment is ineffective.

Given the dataset, any algorithm can only produce a set of subgroups in which treatment is *estimated* to be effective; write $H^+_{\text{est}}$ for this set of subgroups and $H^-_{\text{est}}$ for the complementary set of subgroups. A type-I error occurs if a subgroup $x \in H^+$ is in $H^-_{\text{est}}$ (i.e. treatment is actually effective but is estimated to be ineffective); a type-II occurs if a subgroup $x \in H^-$ is in $H^+$ (i.e. treatment is actually ineffective but is estimated to be effective). For a given estimated set $H^+_{\text{est}}$, the magnitudes of type-I and type-II errors are

$$\mathrm{e}_1^K = \sum_{x\in X} 1\left(x \in H^-\right) 1\left(x \in H^+_{\text{est}}\right)$$
$$\mathrm{e}_2^K = \sum_{x\in X} 1\left(x \in H^+\right) 1\left(x \in H^-_{\text{est}}\right)$$

Given $\lambda \in [0,1]$ the *total error* is: $\mathrm{e}^K = \lambda\mathrm{e}_1^K + (1-\lambda)\mathrm{e}_2^K$ where $\lambda$ is a parameter that is selected by the designer based on the designer's view of the importance of type-I and type-II errors. (We use the superscript $K$ to indicate that we are computing errors after $K$ cohorts have been recruited.)

## 4 Design of a RCT as a Markov Decision Problem

In this subsection, we model the RCT design problem as a finite step non-discounted MDP. A finite step MDP consists of number of steps, a state space, an action space, transition dynamics and a reward function. We need to define all the components of the MDP.

We are given a budget of $N$ patients to be recruited in $K$ steps. At time step $k$, the designer decides to recruit $M_k$ patients; of these $u_k(x, y)$ are from subgroup $x$ and are assigned to treatment $y$, so $\sum_{x\in X}\sum_{y\in Y} u_k(x, y) = M_k$. Having made a decision

$U_k = \{u_k(x,y)\}$, the designer observes the outcomes $W_k = \{W_k(x,y) = \sum_{j=1}^{u_k(x,y)} G(Z_j) : Z_j \sim \mathbb{P}(\cdot|\theta_{x,y})\}$.

Write $\bar{M}_{k-1} = \sum_{\ell=0}^{k-1} M_\ell$ for the number of patients recruited through step $k-1$. We define a filtration $(\mathcal{F}_k)_{k=0}^K$ by setting $\mathcal{F}^k$ to be the sigma-algebra generated by the decisions and observations through step $k-1$: $\{U^0, W^0, U^1, W^1 \ldots, U^{k-1}, W^{k-1}\}$. We write $\mathbb{E}_k[\cdot] = \mathbb{E}\left[\cdot|\mathcal{F}^k\right]$ and $\mathrm{Var}_k[\cdot] = \mathrm{Var}\left[\cdot|\mathcal{F}^k\right]$. Recruitment and allocation decisions are restricted to be $\mathcal{F}^k$-measurable so that decisions to be made at each step depends only on information available from previous steps.

The state space for the MDP is the space of all possible distributions under consideration for $\{\theta_{x,y}\}$. Let $S^k$ denote the $2X \times (d+1)$ *state matrix* that contains the hyper-parameters of posterior distribution of the outcomes for both treatment and control actions for all $(x,y) \in X \times Y$ in the $k$th step. Define $\mathcal{S}^k$ to be the all possible states at the $k$th step, that is, $\mathcal{S}^k = \left\{ S^k = \left[s^k_{x,y}\right] : s^k_{x,y} = \left[s^k_{x,y,0}, s_{x,y,1}\right] \right\}$ where $s^k_{x,y,0}$ is the $d$-dimensional cumulative sufficient statistic and $s_{x,y,1}$ is the number of samples from subgroup $x$ with treatment action $y$.

The action space at step $k$ is the set of all possible pairs of $(M_k, U_k)$ with $M_k \leq N - \bar{M}_{k-1}$ and $\sum_{x,y} u_k(x,y) = M_k$. Taking an action $a_k = (M_k, U_k)$ means recruiting $M_k$ patients in total, of whom $u_k(x,y)$ will be from subgroup $x$ and assigned to treatment $y$. Fix the decision stage as $k$. When the designer selects one of the actions, we use Bayes rule to update the distribution of $\theta_{x,y}$ conditioned on $F^k$ based on outcome observations of $W_k$, obtaining a posterior distribution conditioned on $\mathcal{F}^{k+1}$. Thus, our posterior distribution for $\theta_{x,y}$ is proportional to $\sqrt{\left|F''(\theta)\right|}\exp\left(\theta \cdot s^k_{x,y,0} - F(\theta) s^k_{x,y,1}\right)$. The parameters of the posterior distribution can be written as a function of $s^k$ and $W_k$. Define $S^{k+1} = T(s^k, a_k, W_k)$ to be the transition function given observed treatment outcome $W_k$ and $\mathbb{P}(S^{k+1}|S^k, a_k)$ to be the posterior state transition probabilities conditioned on the information available at step $k$. Having taken an allocation action $a_k = (M_k, U_k)$, the state transition probabilities are: $S^{k+1} = s^k + \left[W^k, U^k\right]$ with posterior predictive probability $\mathbb{P}(W|s^k)$.

For a state $s = (s_0, s_1)$, the action space is the set of pairs of $(m,u)$ with $m$ less than or equal to the remaining patient budget, and elements in $u$ summing up to $m$. Denote this set as $A(s)$.

Given the state vector $s$, write $P_x(s)$ for the posterior probability that the treatment is effective, conditional on $\theta_{x,0}$ and $\theta_{x,1}$ being drawn according to the posterior distributions. Formally,

$$P_x(s) = \mathbb{P}\left( \frac{E(\theta_0, \theta_1)}{\mu(\theta_0)} \geq \tau \middle| \begin{matrix} \theta_0 \sim \mathcal{P}_{\theta|s_{x,0,0}, s_{x,0,1}} \\ \theta_1 \sim \mathcal{P}_{\theta|s_{x,1,0}, s_{x,1,1}} \end{matrix} \right).$$

We can now compute estimated positive and negative sets, $H^+_{est}, H^-_{est}$. If $x \in H^-_{est}$ (i.e. the treatment is estimated to be ineffective for the subgroup $x$) then the probability that $x \in H^+$ (i.e. the treatment is actually effective) is $P_x(s)$. Similarly, if $x \in H^+_{est}$ then the probability that $x \in H^-$ is $1 - P_x(s)$. Hence, given $H^+_{est}, H^-_{est}$ the posterior expected total error is:

$$\sum_{x \in X} \lambda P_x(s) 1(x \in H^-_{est}) + (1-\lambda)(1 - P_x(s)) 1(x \in H^+_{est}) \quad (1)$$

The following proposition identifies the set that minimizes this posterior expected total error.

**Proposition 1.** *Given the terminal state $s$, the set that minimizes this posterior expected total error is $H^+_{est} = \{x \in X : P_x(s) \geq 1 - \lambda\}$.*

Proof of the proposition is given in the supplementary material. Now define $g$ by

$$g(x;\lambda) = \lambda(1-x)1(x \geq 1-\lambda) + (1-\lambda)x 1(x < 1-\lambda).$$

Then, the posterior expected total error can be written as $\mathrm{e}^K = \sum_{x \in X} g\left(P_x(s^K); \lambda\right)$. We'll omit using $\lambda$ in the rest of the paper for notational brevity. We define the reward function $R : S \times A \to \mathbb{R}$ as the decrease in the posterior expected total error that results from taking a particular action in state $s$; i.e.

$$R(s,a) = \sum_{x \in X} \mathbb{E}\left[ g\left(P_x(s)\right) - g\left(P_x(S^{k+1})\right) \middle| a^k = a \right]$$

where the expectation is taken with respect to the outcome distributions of the treatment and control actions given the state vector $s$. A *policy* is a mapping $\pi : S \to A$ from the state space to the action space, prescribing the number of patients to recruit from each subgroup and how to assign them to treatment groups. The value function of a policy $\pi$ beginning from state $s^0$ is $V^\pi(s^0) = B(s^0) - \sum_{x \in X} \mathbb{E}^\pi\left[g\left(P_x(s^k); \lambda\right)\right] = \sum_{k=0}^K \mathbb{E}^\pi\left[R(S^k, A_k)\right]$ where $B(s^0) = \sum_{x \in X} g\left(P_x(s^0); \lambda\right)$ and the expectation is taken with respect to the policy $\pi$. Our learning problem is the $K$-stage MDP with tuple: $\{K, \{S^k\}, \{A(s)\}, T(s^k, a_k, W_k), R(s^k, a_k)\}$ and our goal is to solve the following optimization problem:

$$\begin{aligned} &\text{maximize}_\pi \sum_{k=1}^K \mathbb{E}^\pi\left[R(S^k, \pi(S^k))\right] \\ &\text{subject to } \pi(S_k) \in A(S_k) \quad \text{for all } k \end{aligned}$$

In the next subsection, we propose a Dynamic Programming (DP) approach for solving the $K$-stage MDP defined above.

### 4.1 Dynamic Programming (DP) solution

In the dynamic programming approach, the value function is defined as the optimal value function given a particular state $S^k$ at a particular stage $k$, and is determined recursively through Bellman's equation. If the value function can be computed efficiently, the optimal policy can also be computed from it. The optimal value function at the terminal stage $K-1$ (the stage in which the last patient is recruited) is given by: $V^{K-1}(s) = \max_{a\in A(s)} R(s,a)$ The dynamic programming principle tells us that the loss function at other indices $0 \le k < K-1$ is given recursively by $Q^k(s,a) = \mathbb{E}_k\left[V^k(T(s,a,W))\right]$, $V^k(s) = \max_{a\in A(s)} Q^k(s,a)$ where the expectation is taken with respect to the distribution of the cumulative treatment outcomes $W$. The dynamic programming principle tells us that any policy that satisfies the following is optimal: $A_k^* = \arg\max_{a\in A(S^k)} = Q^k(S^k,a)$.

However, it is computationally intractable to solve the DP because the state space contains all possible distributions under consideration and so is very large. In what follows of the paper, we propose an (approximate) solution under the restriction that the total number $M$ of patients to be recruited at each step is fixed and determined by the designer. (As we show in the experiments, the choice of $m$ can have a large impact on the performance.) Our approach is greedy but computationally tractable. The proposed algorithm, RCT-KG, computes for each action $a \in A$ the one-stage reward that would be obtained by taking $a$, and then selects the action with the maximum reward.

## 5 The RCT-KG Algorithm

Because solving the MDP is intractable, we offer a tractable approximate solution. We focus on the setting where the number of patients $M$ in each cohort is fixed: $M = T/K$. In this circumstance, what is to be decided for each cohort (each step) is the number of patients to recruit in each subgroup and the group-specific assignments of these patients (subject to the constraint that the total number of patients in each cohort is $M$). The action set is therefore $A = \left\{u : \sum_x \sum_y u(x,y) = M\right\}$ and the size of this action set is $|A| = \binom{M+2X-1}{2X-1}$. The Rand-KG algorithm computes an expected improvement in the terminal value function by taking an action $a \in A$. The Knowledge Gradient (KG) policy selects the action that makes the largest expected improvement in the value function at each instance, that is,

$$A_k^{KG}(s) = \arg\max_{a\in A(s)} \mathbb{E}_k\left[V^K(P(s,a,W)) - V^K(s)\right]$$
$$= \arg\max_{a\in A(s)} \int_w V^K(P(s,a,w))\mathbb{P}(w|S^k)\,dw$$

**Algorithm 1** Optimistic Action Computation

**Input :** Current state vector: $s$
Set optimal action $u^* = 0$
**for** $m = 1,\dots,M$ **do**
  **for** $(x,y) \in X \times Y$ **do**
    Set $s_1 = T(s, u^*, u^* \odot G(z_{\max}))$
    Set $s_2 = T(s, u^*, u^* \odot G(z_{\min}))$
    Set $\tilde{u} = u^* + 1_{(x,y)}$
    Compute $v_1 = V^K(T(s,\tilde{u},\tilde{u} \odot G(z_{\max}))) - V^K(s_1)$
    Compute $v_2 = V^K(T(s,\tilde{u},\tilde{u} \odot G(z_{\min}))) - V^K(s_2)$
    Compute $q(x,y) = \max(v_1, v_2)$.
  **end for**
  Compute $(x^*, y^*) = \arg\max_{x,y} q(x,y)$.
  Update $u^* = u^* + 1_{(x^*,y^*)}$.
**end for**
Return $A^{RCT-KG}(s) = u^*$.

However, computing the KG policy requires computing the posterior predictive distribution and posterior expectation for each action in $A$. This is a computationally intractable procedure because the size of the action space is on the order of $\mathcal{O}\left(M^{2X-1}\right)$. Hence we propose the RCT-KG algorithm which computes *optimistic* improvements in the value functions. Algorithm 1 shows a tractable way of computing these optimistic improvements. At each iteration $m$, the procedure computes the maximum improvement in the value function that can be obtained from an additional sample from the pair $(x,y)$ and increments that index by 1. The complexity of computing $A^{RCT-KG}(s)$ is only $\mathcal{O}(MX)$.

**Algorithm 2** The RCT-KG Algorithm

**Input :** $K$, $S^0$
**for** $k = 1,\dots,K$ **do**
  Compute $U_k^* = A^{RCT-KG}(S^k)$ (Algorithm 1).
  Recruit the patients based on $U_k^*$, observe cumulative treatment outcome $W_k^*$.
  Update $S^{k+1} = S^k + (W_k^*, U_k^*)$.
**end for**
Compute $H_{\text{est}}^+ = \{x \in \mathcal{X} : P_x(S^K) \ge 1-\lambda\}$.
**Output :** $H_{\text{est}}^+$.

At each step $k$, the best action computes using the procedure in Algorithm 1 and patients are recruited and

assigned accordingly. At the end of the current decision step, the state vector is updated based on the observed treatment outcomes. When the patient budget is exhausted, our algorithm outputs (as the estimated positive set) the set of subgroups with clinically relevant improvements: $H^+_{\text{est}} = \{x \in X : P_x(s^K) \geq 1-\lambda\}$. The pseudo-code for RCT-KG is given in Algorithm 2.

# 6 Experiments

In all of our experiments, we use assume that the outcome is either success or failure of the treatment for that patient, and assume the outcomes follow a Bernoulli distribution; we continue to assume that the outcomes for each cohort of patients is observable before the succeeding cohort must be recruited and allocated. (The clinical study in [Hébert et al.(1999)] provides a real-life example.) We assume there are identifiable subgroups (e.g. distinguished as to male/female and/or young.old). In each experimental setting we generate 1000 independent experiments and report the average over these 1000 experiments. We compare the results of our algorithm with those Uniform Allocation (UA), aka repeated fair coin tossing, that uniformly randomly recruits the patients from subgroups uniformly assigns the patients to treatment groups, and (where appropriate) with Thompson Sampling (TS) [Agrawal and Goyal(2012)] that draws the parameters of the outcomes for each action and then selects the action with the best sample outcomes and an adaptive randomization, a variant of DexFEM [Warner et al. (2015)], that shifts the treatment allocation ratio towards treatments with higher posterior variance. We use TS and DexFEM only for treatment allocation.

For the first experiment we consider a setting with two subgroups; for the remaining experiments, we consider a setting with four subgroups.

## 6.1 Error Rates with Two Subgroups

We begin with a setting with 2 subgroups 0, 1. We assume the true parameters for the subgroups are $\theta_{x,0} = 0.5$ for $x \in \{0,1\}$, $\theta_{1,1} = 0.7$ and we vary $\theta_{0,1}$ from 0.51 to 0.70. (Note that identifying the best treatment is more challenging for subgroup 0 than for subgroup 1.) We recruit 100 patients in each of 10 cohorts – 1000 patients in total. Figure 2 compares the performance of our RCT-KG with UA in terms of error rates; as can be seen, our algorithm outperforms UA throughout the range of the parameter $\theta_{0,1}$, and the improvement in performance is greatest in the middle range of this parameter, when the difference between treatment and control among the subgroups is greatest. This improvement is achieved because our algorithm recruits and samples more frequently from subgroup 0, which represents the more challenging learning problem

## 6.2 Error Rates and Confidence Levels with Four Subgroups

We now turn to a setting in which there are 4 subgroups 0,1,2,3. We take the parameters to be $\theta_{x,0} = 0.5$ for all $x$ and $\theta_{x,1} = 0.3, 0.45, 0.55, 0.7$ for $x = 0, 1, 2, 3$. Note that there are two subgroups in which the treatment action is ineffective and two subgroups in which it is effective, and that identification is easier in subgroups $0, 3$ than in subgroups $1, 2$. We examine a number of aspects.

### 6.2.1 Confidence Levels Across Cohorts

In the first experiment in this setting, we recruit 100 patients in each of 10 cohorts, and compare the confidence levels achieved for each cohort and for each subgroup; the results for 4 horizons are shown in Table 2. As can be seen, the RCT-KG, UA and DexFEM algorithms achieve very similar confidence levels (probability of correctly identifying the actual label) for subgroups 0,3 at each of these time horizons, but RCT-KG algorithm achieves significantly better confidence levels for subgroups 1,2 – the subgroups for which identification is more difficult. RCT-KG achieves superior confidence levels because it recruits more patients to subgroups 1,2 and allocates patients in each subgroup more informatively. To illustrate, we refer to Table 3, which shows the total number of patients recruited to each of the subgroups and the allocation of patients to control and treatment within the subgroups. (Remember that we are reporting averages over 1000 experiments.)

| Algorithm/ SG | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| RCT-KG | 98.79 | 82.92 | 83.56 | 98.78 |
| DexFEM | 98.94 | 79.28 | 79.25 | 99.00 |
| UA | 98.92 | 78.93 | 78.96 | 98.94 |

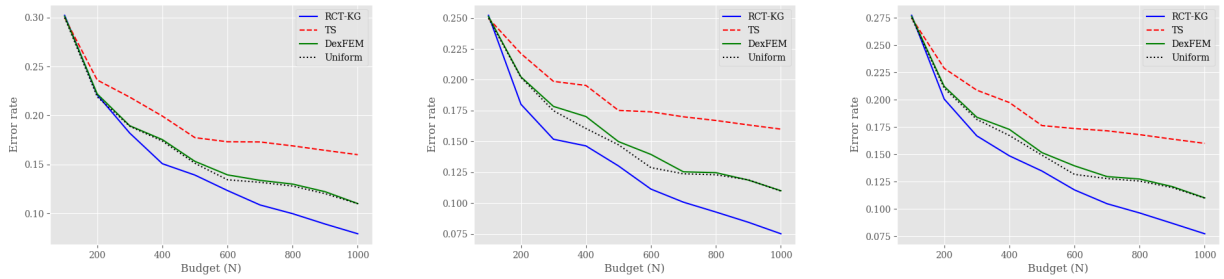
Table 2: Comparison of confidence levels on subgroups

| SG | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| RCT-KG | $81, 60$ | $190, 181$ | $186, 177$ | $81, 54$ |
| DexFEM | $137, 118$ | $128, 127$ | $128, 127$ | $137, 118$ |
| UA | $128, 127$ | $127, 128$ | $128, 127$ | $127, 128$ |

Table 3: Recruitment and allocation in each subgroup # allocated to control, # allocated to treatment

### 6.2.2 Achieving a Prescribed Confidence Level

In this experiment, we recruited 100 patients in each cohort and continued recruiting patients until a pre-




(a) Type-I error (b) Type-II error (c) Total error

Figure 1: Error Comparisons with Benchmarks

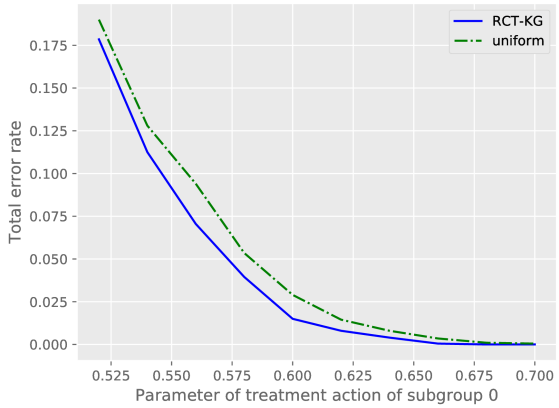



Figure 2: Total error rates for different parameter

scribed (average) confidence level $\beta = 0.90, 0.95$ was achieved (i.e., until $\frac{1}{X}\sum_{x\in\mathcal{X}} g(P_x(S^k)) < 1-\beta$.) Table 4 shows the number of cohorts necessary for each algorithm to achieve the prescribed confidence level; as can be seen, RCT-KG achieves the same confidence level as UA and DexFEM using fewer cohorts of patients, which means that a RCT could be carried out with fewer patients and completed in less time.

| Algorithm | $\beta = 0.95$ | $\beta = 0.90$ |
|---|---|---|
| RCT-KG | 12.6 | 7.2 |
| DexFEM | 22.5 | 10.2 |
| UA | 22.9 | 10.7 |

Table 4: Comparison of length for a confidence level

#### 6.2.3 Error Rates and Patient Budgets

In this experiment we recruited 100 patients in each cohort and computed the type-I, type-II and overall error rates for various total patient budgets. As seen from Figure 1, RCT-KG significantly outperformed the UA, DexFEM and TS algorithms for all budgets. (TS did especially poorly when the patient budget is large because TS aims to maximize the patient benefit, not the learning performance, and so allocated more patients to the treatment that has been found to be better at each stage, which slows learning.)

#### 6.2.4 Cohort Size

In this experiment, we compared the performance in terms of total error of RCT-KG and UA when $m$ patients were recruited in each cohort with a total budget of 500 patients. As seen in Table 5, the error rate of UA is independent of $m$ because all patients are recruited and allocated in the same way in every cohort, but the error rate of RCT-KG is significantly smaller when $m$ is smaller because smaller cohorts provide more opportunities to learn.

| $m$ | 25 | 50 | 100 | 250 |
|---|---|---|---|---|
| RCT-KG | 0.1245 | 0.1281 | 0.1292 | 0.1411 |
| UA | 0.1484 | 0.1484 | 0.1484 | 0.1484 |

Table 5: Total Errors for Different Cohort Sizes

## 7 Conclusion and Future Work

This paper makes three main contributions. (1) We formalize the problem of allocating patients in a RCT as a finite stage MDP. (2) We provide a greedy computationally tractable algorithm RCT-KG that provides an approximately optimal solution to this problem. (3) We illustrate the effectiveness of our algorithm in a collection of experiments using synthetic datasets for which outcomes are drawn from a Bernoulli distribution with unknown probabilities.

### 7.1 Acknowledgements

This research is supported by the Office of Naval Research (ONR) and the NSF (Grant number: ECCS1462245, ECCS1533983, and ECCS1407712).

## References


[Agrawal and Goyal(2012)] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 1–39, 2012.

[Auer et al.(2002)] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[Barker et al.(2009)] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, L. J. Esserman. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 47 (86-1):97-100, 2009.

[Berry et al.(2006)] D. A. Berry. Bayesian clinical trials. *Nature reviews Drug discovery*, 47(5-1):27, 2006.

[Bhatt et al.(2006)] D. A. Berry. Adaptive designs for clinical trials. *New England Journal of Medicine*, 375:65–74, 2006.

[Cannon et al.(2004)] C. P. Cannon, E. Braunwald, C. H. McCabe, D. J. Rader, J. L. Rouleau, R. Belder, S. V. Joyal, K. A. Hill, M. A. Pfeffer, and A. M. Skene. Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *New England journal of medicine*, 350 (15):1495–1504, 2004.

[Chen(1995)] C.-H. Chen. An effective approach to smartly allocate computing budget for discrete event simulation. In *Decision and Control, Proceedings of the IEEE Conference on*, volume 3, pages 2598–2603, 1995.

[Chen et al.(2003)] C.-H. Chen, K. Donohue, E. Yücesan, and J. Lin. Optimal computing budget allocation for monte carlo simulation with application to product design. *Simulation Modelling Practice and Theory*, 11(1):57–74, 2003.

[Chen et al.(2013)] X. Chen, Q. Lin, and D. Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *International Conference on Machine Learning*, pages 64–72, 2013.

[Chen et al.(2015)] X. Chen, Q. Lin, and D. Zhou. Statistical decision making for optimal budget allocation in crowd labeling. *The Journal of Machine Learning Research*, 16(1):1–46, 2015.

[Farrar et al.(2000)] J. T. Farrar, R. K. Portenoy, J. A. Berlin, J. L. Kinman, B. L. Strom. Defining the clinically important difference in pain outcome measures. *Pain*, 21(4):599–613, 2009.

[Frazier et al.(2009)] P. Frazier, W. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(88-3):287–294, 2000. 599–613, 2009.

[Frazier et al.(2008)] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[Friedman et al.(1998)] L. M. Friedman, C. Furberg, D. L. DeMets, D. Reboussin, and C. B. Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

[Gittins et al.(2011)] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[Hacke et al.(1995)] W. Hacke, M. Kaste, C. Fieschi, D. Toni, E. Lesaffre, R. Von Kummer, G. Boysen, E. Bluhmki, G. Höxter, M.-H. Mahagne, et al. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke: the european cooperative acute stroke study (ecass). *JAMA*, 274(13):1017–1025, 1995.

[Hacke et al.(2008)] W. Hacke, M. Kaste, E. Bluhmki, M. Brozman, A. Dávalos, D. Guidetti, V. Larrue, K. R. Lees, Z. Medeghri, T. Machnig, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *New England Journal of Medicine*, 359(13):1317–1329, 2008.

[Hébert et al.(1999)] P. C. Hébert, G. Wells, M. A. Blajchman, J. Marshall, C. Martin, G. Pagliarello, M. Tweeddale, I. Schweitzer, E. Yetisir, and T. R. in Critical Care Investigators for the Canadian Critical Care Trials Group. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *New England Journal of Medicine*, 340(6): 409–417, 1999.

[Hu and Rosenberger(2006)] F. Hu and W. F. Rosenberger. *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons, 2006.

[Lachin et al.(1988)] J. M. Lachin, J. P. Matts, and L. Wei. Randomization in clinical trials: conclusions and recommendations. *Controlled clinical trials*, 9(4):365–374, 1988.

[Lewis and Bessen(1990)] R. J. Lewis and H. A. Bessen. Sequential clinical trials in emergency medicine. *Annals of emergency medicine*, 19(9): 1047–1053, 1990.

[Moher et al.(2012)] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, 10(1):28–55, 2012.

[Schulz and Grimes(2002)] K. F. Schulz and D. A. Grimes. Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, 359(9306):614–618, 2002.

[Moss et-al. (2015)] R. B. Moss, P. A. Flume, J. S. Elborn, J. Cooke, S. M. Rowe, S. A. McColley, R. C. Rubenstein, M. Higgins. Efficacy and safety of ivacaftor treatment: randomized trial in subjects with cystic fibrosis who have an R117H-CFTR mutation. *The Lancet*, 3(9306):614–618, 2002.

[Villar et al.(2015)] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199, 2015.

[Warner et al. (2015)] P. Warner, C J. Weir, C. H. Hansen, A. Douglas, M. Madhra, S. G. Hillier, P.T. K. Saunders, J. P. Iredale, S. Semple, B. R. Walker and others . *Low-dose dexamethasone as a treatment for women with heavy menstrual bleeding: protocol for response-adaptive randomised placebo-controlled dose-finding parallel group trial (DexFEM). BMJ open*, 30(5-1):6837, 2015.

[Whitehead(1997)] J. Whitehead. *The design and analysis of sequential clinical trials.* John Wiley & Sons, 1997.

[Whittle(1980)] P. Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 143–149, 1980.