Budgeted Learning For Developing Personalized Treatment

Kun Deng
Department of Statistics
University of Michigan
Michigan, USA
kundeng@umich.edu

Russ Greiner
Department of Computer Science
University of Alberta
Alberta, Canada
rgreiner@ualberta.ca

Susan Murphy
Department of Statistics
University of Michigan
Michigan, USA
samuprhy@umich.edu

***Abstract*—There is increased interest in using patient-specific information to personalize treatment. Personalized treatment decision rules can be learned using data from standard clinical trials, but such trials are very costly to run. This paper explores the use of budgeted learning techniques to design more efficient clinical trials, by effectively determining which type of patients to recruit, at each time, throughout the duration of the trial. We propose a Bayesian bandit model and discuss the computational challenges and issues pertaining to this approach. We compare our budgeted learning algorithm, which approximately minimizes the Bayes risk, using both simulated data and data modeled after a clinical trial for treating depressed individuals, with other plausible algorithms. We show that our budgeted learning algorithm demonstrated excellent performance across a wide variety of situations.**

***Keywords*-Budgeted Learning; Bayesian; Active Learning; Personalized Treatment;Reinforcement Learning**

I. Introduction

As every patient is different, a treatment that works well for one patient may be completely ineffective or even harmful for other patients with the same disease. The goal of personalized medicine is to treat a patient with the treatment that is best for *her*, based on *her* own characteristics [1]. Unfortunately, medicine is still a long way from this goal. Over the past decade, medical science has taken a staged approach [2], where the first step has been to identify useful characteristic (biomarker) or combinations of characteristics (biomarker profiles). These characteristics can be any biological, psychological, social and genetic factor that is likely to influence the effectiveness of the treatment. The second step is to determine which treatment is best for each biomarker profile. Hence, a central challenge in clinical research is the development of clinical trial designs that efficiently evaluate the clinical usefulness of the putative biomarker profiles for personalizing treatment.

As an example, suppose researchers in depression disorders have conjectured that the best treatment for several biomarker profiles might be different. Here a natural question is how best to design a clinical trial to learn the best treatment for each profile; see Section IV-B. To reduce the cost of running such trials, many researchers have started to utilize modern clinical infrastructure such as electronic medical records to recruit their participants, thus permitting targeted recruitment of participants in each biomarker profile. We can therefore formulate the design of the clinical trial as a Bayesian budgeted bandit problem, where each intermediate decision step specifies which type of patients (that is which biomarker profile) to recruit, anytime during the trial. The goal is to optimize the overall quality of the resulting personalized treatment. Below we formualte this task more precisely, then propose an approximation algorithm for solving this problem and provide experimental results that demonstrate the excellent performance of our approach.

We further discuss related work in clinical research, budgeted learning and other areas in Section II. We then give a formal definition of the Bayesian budgeted design problem for developing personalized treatment and describe a look-ahead algorithm in Section III. Finally, Section IV describes various properties of the proposed algorithm and demonstrates its effectiveness using both simulated data and data modeled after a specific clinical trial.

II. Related Work

In clinical research, currently the most common approach for personalizing treatment is to use a conventional multi-arm randomized trial with post-hoc, subgroup analyses [3]. Typically, this type of trial recruits uniformly from the whole population of patients who meet a basic "inclusion" criterion, and then randomly assigns each patient to one of the several treatment arms. Only after the entire trial is completed do researchers search for patient subgroups for which the treatment effect seems to be significant. Unfortunately this approach suffers from severe drawbacks. Findings in these post-hoc analyses are notoriously hard to replicate: biomarker profiles discovered retrospectively in post-hoc analyses are likely to be spurious due to chance. This is due (at least partially) to the use of many statistical tests without controlling the overall error rate. It is also difficult to obtain a good estimate of treatment effects for rare biomarker profiles in these trials, as those profiles are most-likely under-represented. A more severe problem is that the conventional multi-armed randomized trial does not make efficient use of the trial resources, as it does not take advantage of the information accumulated during the trial.

The basic problem is that these multi-arm randomized trials were designed to compare treatments, but not for our task of informing the development of personalized treatment. (This type of standard trials could help develop personalized treatment *indirectly*—by proposing candidate biomarker profiles for further study.)

In clinical research literatures, other methods have been proposed to determine if there is any benefit of a single binary biomarker [2]. These designs use hypothesis-testing based approaches to test if a treatment is beneficial in neither biomarker profiles (null hypothesis) versus the treatment benefits one of the profiles (alternative hypothesis). These methods have been largely studied in a two-stage enrichment design [4], [5] over two profiles. Typically, in the first stage, the clinical trial recruits patients from both profiles, then in the second stage, the clinical trial recruits from one or both of the profiles, depending on the evidence collected from the first stage. We consider a generalization of this approach to situations with multiple biomarker profiles in which patients are recruited continuously without requiring an interim stopping of the trial.

Our formulation (Section III) also bears formal similarity with targeted response adaptive trials [6], recently popular in cancer research, which also divide patients into subgroups. The cancer patients are usually so sick that these trial designs naturally aim to place more patients on what appears to be the better treatment in each subgroup, which echoes the objective of "maximizing cumulative reward" in standard bandit problems [7]. Similarly in machine learning and statistics, bandit problems that aim to explore and exploit contextual information (i.e. subgroups or biomarker profiles) have been studied under the name of contexual bandits or bandit with covariates, [8], [9]. These works also aim to maximize cumulative reward, which is quite different from the goal of our work.

Our goal is to produce a system that works effectively, at "performance time" (i.e. treating patients outside the trial). As such, it is similar to the standard budgeted (probe) learning [10] in machine learning, which seeks to learn the most accurate classifier subject to a fixed data collection budget. The main differences are that budgeted probe learning (1) at training time, sequentially selects a feature on an instance to probe, and (2) produces a classifier; whereas we (1) apply a specified treatment to a patient from a specified biomarker profile and (2) produce treatment decision rules that map a biomarker profile to a treatment.

A budgeted version of the Bernoulli bandit problem was studied in [11], in which the goal is to minimize the difference between the success probability of the chosen arm and that of the best arm. In this setting the regret corresponds to the hinge loss. Our formulation in Section III is also a budgeted bandit problem, in which the action of "arm pull" corresponds to recruiting from a biomarker profile. Our formulation also uses the hinge loss. The main difference lies in the choice of the objective function, as our goal is to learn the best treatment for *all* biomarker profiles, which differs from their goal of selecting the *single best* arm. Other budgeted bandit problems that focus only on the "end results" have also been studied [12]–[15]. These works differ from ours mathematically either due to the choice of loss functions (sum of variances or 0-1 loss vs hinge loss) or the choice of risk functions (Bayesian vs frequentist risk).

Budgeted bandit problems are also connected to traditional sequential analysis in statistics [16], [17]. Often the question there is when one should stop data collection, where the goal is minimizing the combined loss of decision making and sampling cost, usually there is no pre-specified hard sample size constraint. Our work utilizes the Bayesian formulation [17], which is appropriate as it provides a natural way to utilize the information collected during the experiment. Finally, there is a large literature in operation research on "ranking and selection", in which the goal is usually to select the best arm meeting a prespecified probability guarantee. Our work uses some of their algorithmic ideas, especially from [18].

III. Methods

We consider the following budgeted learning problem.

1. There are 2 treatments $S = \{A, B\}$ for a fixed disease.

2. There are K distinct biomarker profiles of people with the disease.

3. For each $i \in \{1, \cdots, K\}$, and $s \in S$, the clinical outcome of treating patient from profile i with treatment s is $Y_{i,s}$. The outcomes $Y_{i,s} \sim Norm(\mu_{i,s}, \sigma_\epsilon^2)$ are independent conditioned on $\mu = (\mu_{1,A}, \mu_{1,B}, \cdots, \mu_{K,A}, \mu_{K,B})$, which is unknown. σ_ϵ^2 is assumed both known and homogeneous over profiles and treatments. The assumption of known σ reduces the clutter of notations, but does not fundamentally alter the problem.

4. A budgeted learning trial design b (i.e., an MDP policy [19]) for a sampling size (budget) $2T$, is a tuple $b = (e, d)$, with $e = (e_0, \cdots, e_{T-1})$ being the exploration rule and d being the final decision rule. For computational convenience and as an initial investigation, we consider randomization of treatments within each profile in the following way: at decision time point $t = 0, \cdots, T-1$, the exploration rule e_t suggests from which biomarker profile to recruit the next *pair* of patients. When a profile is chosen for experimentation, *a pair of patients* from that profile will be selected, one of which is assigned to treatment A and the other to treatment B. As the sample size is $2T$, a budgeted learning trial design b will need to make T exploration decisions.

Let $X^t \in (\{1, ..., K\} \times \mathbb{R} \times \mathbb{R})^t$ denote the sequence of recruitment and outcomes observed up to time t, with action $\in \{1, ..., K\}$ and real-valued outcomes for treatment A and B at each time. $e_t(X^t)$ is the next exploration action taken at time $t + 1$, and $X_{t+1} \equiv (e_t(X^t), Y_{e_t(X^t),A}, Y_{e_t(X^t),B})$,

$X^{t+1} \equiv [X^t, X_{t+1}]$, and so on. At time $t = 0$, $X^0 = \{\}$. We assume for now there is no delay in observing the outcome: before the next decision time $t+1$ arrives, X_t has already been observed. (But see Section IV-D)

At decision time $t = T$, a final decision rule d specifies how to map profile information to a treatment. That is, $d = (d_1, d_2, ..., d_K)$, where $d_i(X^T) \in \{A, B\}$ is the treatment deemed best for patients belonging to profile i.

5. We define the *loss function* for a decision action $a = (a_1, \cdots, a_k) = d(X^T) \in \{A, B\}^K$, as the sum of regrets of not selecting the better treatments:

$$L(\mu, a) = \sum_i [\mu_{i,\bar{a}_i} - \mu_{i,a_i}]_+ \tag{1}$$

where μ is the unobserved effect vector, a_i is the chosen treatment for profile i, $\bar{a}_i = \{A, B\} \backslash a_i$ is the "other" treatment, and $[x]_+ = \max(x, 0)$ is the hinge loss. While it is tempting to consider the 0-1 loss: $L_{0/1}(\mu, a) = \sum_i 1\{\mu_{i,\bar{a}_i} > \mu_{i,a_i}\}$, corresponding to the (probability of) failure to identify the best treatment, this loss can be too harsh when $\mu_{i,A}$ and $\mu_{i,B}$ are very close to each other, which might lead to excessive sampling of such profiles. In classification problems, this motivates algorithms (e.g. SVM) that choose to optimize margin loss rather than 0-1 classification error. In Section IV, we report the performance of our algorithm using both loss functions. Finally, while we could extend our model to handle weighted loss (say, by subpopulation sizes), we intentionally view all profiles equally: if the rarest profile has the largest effect, we want to discover this.

6. We take a Bayesian view by assuming the true treatment effect vector μ follows a prior multivariate Gaussian distribution $\rho = N(\mu, v)$: $\mu = (\zeta, \cdots, \zeta)$, and $v = Diag[\tau^2, \cdots, \tau^2]$; below we set $\zeta = 0$ and $\tau = 100$ as an uninformative prior. It would be trivial to extend our model to accommodate correlated profiles. We did not pursue this as in our applications, different profiles are often based on very different biological pathways, so there is no apriori reason to believe that the treatment effects should be correlated across profiles.

The *Bayes decision risk* of the budgeted trial design $b = (e, d)$ is

$$r(\rho, b) = \mathbb{E}\, L(\mu, d(X^T)) \tag{2}$$

where the expectation is taken with respect to both the prior distribution of the μ as well as the conditional distribution of observed outcome history X^T given μ. Note X^T implicitly depends on b. Our goal is to find a trial design with decision risk close to $\inf_b r(\rho, b)$.

A. Bayes Decision Rule d^ρ and Posterior Decision Analysis

Given our framework, the posterior distribution of $\mu | X^t$ for $t = 1, \cdots, T$ is a multivariate Gaussian distribution, denoted as $\rho^t = N(\mu^{(t)}, v^{(t)})$, where $\mu^{(t)}$ is the posterior mean vector, and $v^{(t)}$ is the posterior covariance matrix. Define $\kappa(\rho^t, a_i) \equiv \mathbb{E}_{\mu \sim \rho^t}[\mu_{i,\bar{a}_i} - \mu_{i,a_i}]_+$ to be the posterior expected loss for decision a_i for the ith profile. Also define the "posterior expected loss" of a decision $a = (a_1, a_2, \cdots, a_K)$ as:

$$\kappa(\rho^t, a) = \sum_i \kappa(\rho^t, a_i) = \sum_i \mathbb{E}_{\mu \sim \rho^t}[\mu_{i,\bar{a}_i} - \mu_{i,a_i}]_+ \tag{3}$$

The "posterior Bayes action", denoted as $d^\rho(X^T) \in \arg\min_a \kappa(\rho^T, a)$, is any decision a that minimizes $\kappa(\rho^T, a)$ over $a \in \{A, B\}^K$. Berger [17] shows that in our Bayesian setting, finding d^ρ (the Bayes decision rule) is equivalent to finding $d^\rho(X^T)$ for any outcome X^T.

The Bayes decision rule is

$$d_i^\rho(X^T) = \begin{cases} A & \text{if } \mu_{i,B}^{(T)} \le \mu_{i,A}^{(T)} \\ B & \text{otherwise} \end{cases} \tag{4}$$

where $\mu_{i,B}^{(T)}$ and $\mu_{i,A}^{(T)}$ are the posterior means. To see this, note that by linearity, Equation 3 is minimized when each term $\mathbb{E}_{u \sim \rho^T}[\mu_{i,\bar{a}_i} - \mu_{i,a_i}]_+$ is minimized. Now we focus on one of the profiles (say, profile 1) and drop the subscript i. To compute the Bayes action $\arg\min_{a \in \{A,B\}} \kappa(\rho^T, a)$, we first compute $\kappa(\rho^T, A)$. Here

$$\begin{aligned} \kappa(\rho^T, A) &= \sigma\phi(\delta/\sigma) + \delta\Phi(\delta/\sigma) & (5) \\ \sigma^2 &\equiv v_A^{(T)} + v_B^{(T)} & (6) \\ \delta &\equiv \mu_B^{(T)} - \mu_A^{(T)} & (7) \end{aligned}$$

where $v_A^{(T)}$ is the posterior variance parameter in ρ^T for treatment A, and $v_B^{(T)}$ for treatment B; and $\Phi(\cdot)$ (resp. $\phi(\cdot)$) is the CDF (resp. PDF) of $N(0, 1)$. We can derive a similar result for $\kappa(\rho^T, B)$, differing only by reversing the sign of δ. Thus we should choose treatment A if $\kappa(\rho^T, A) \le \kappa(\rho^T, B)$,

$$\begin{aligned} \Leftrightarrow \quad & \sigma\phi(\delta/\sigma) + \delta\Phi(\delta/\sigma) & \le \quad & \sigma\phi(\delta/\sigma) - \delta\Phi(-\delta/\sigma) \\ \Leftrightarrow \quad & \delta(\Phi(\delta/\sigma) + \Phi(-\delta/\sigma)) = \delta & \le \quad & 0 \qquad (8) \end{aligned}$$

From now on, we will assume that this Bayes decision rule d^ρ will be used. A more general derivation shows that, as long as the loss function is a summation of hinge losses, and the posterior distributions ρ^T can be factorized as products of marginals for each profile, the decision rule in Equation 4 holds for other distributions as well.

B. Approximating the Exploration Rule e^ρ

Calculating the Bayes exploration rule e^ρ is much harder than calculating the Bayes decision rule d^ρ. Conceptually, the optimal exploration rule is the solution to a continuous space Markov Decision Process (aka "belief MDP") in which the unobserved states are the sets of all possible treatment effect vectors μ. In other words, this is a "multi-armed bandit" situation. The belief states of the belief MDP are the set of all possible posterior distributions of μ; these posterior distributions can be parameterized by the posterior mean vector and the allocations of patients to each profile. Conceptually we can then define the transition kernel given the exploration action taken, and use value iteration algorithms. However in practice, belief MDPs are often computationally intractable to solve exactly, leading to many proposals for approximate solutions [20], [21].

Here we propose a simpler algorithm based on ideas from operations research using two approximation ideas. The first idea is to approximate the dynamic allocation using a static allocation, which determines how many patients from each profile to select, before seeing any of the outcomes. Because even the static allocation is computationally hard, the second idea is to employ a further approximation to the posterior distribution to compute the static allocation.

First consider approximating the dynamic allocation with a static allocation problem: Here we start with a prior distribution $\rho = N(\mu^{(0)}, v^{(0)})$ for μ, and allocate the next m pairs of patients at one shot, for some specified number of look-ahead steps m. How should we allocate these $2m$ instances to maximally reduce the decision risk $r(\rho, (e, d))$? The general strategy is to compute $r(\rho, (e, d))$ for all possible e (i.e. combinatorial allocations of m pairs among K profiles) then choose the e that produces minimal risk r. Let $n = (n_1, n_1, n_2, n_2, \cdots, n_K, n_K)$ be the number of patients to be assigned to each profile/treatment combination, and $\rho^m = N(\mu^{(m)}, v^{(m)})$ be the posterior distribution of μ after observing the $2m$ outcomes (denoted as X^m) of these patients. Then the static allocation problem is to choose n such that

$$\underset{n}{\text{minimize}} \quad r(\rho, (e, d)) = \mathbb{E}_{X^m}\, \kappa(\rho^m, d(X^m)) \tag{9}$$

$$\text{s.t.}\ \ \kappa(\rho^m, d(X^m)) = \sum_i \sigma_i \phi(\delta_i/\sigma_i) + \delta_i \Phi(\delta_i/\sigma_i) \tag{10}$$

$$\sigma_i^2 = v_{i,A}^{(m)} + v_{i,B}^{(m)}$$

$$\delta_i = \mu_{i,\bar{d}_i(X^m)}^{(m)} - \mu_{i,d_i(X^m)}^{(m)}$$

$$\sum n_i = m$$

where the expectation in $\mathbb{E}_{X^m}\, \kappa(\rho^m, d(X^m))$ is with respect to the marginal density of $X^{(m)}$; $v_{\cdot,\cdot}^{(m)}$ are the posterior variance parameters; and $\mu_{\cdot,\cdot}^{(m)}$ are the posterior mean parameters conditioned on X^m. Note that $v^{(m)}$ is fully determined by the prior ρ and the allocation n, but does not depend on outcomes X^m, while $\mu^{(m)}$ does depend on the "future" outcome X^m.

The difficulty in solving this static allocation problem is that the objective function is hard to directly calculate. The posterior expected loss $\kappa(\rho^m, d(X^m))$ depends on data X^m, which is not yet observed. This is where the second approximation idea is useful. When m is not too big, Chen [18] proposed assuming $\mu^{(m)} \simeq \mu$, i.e. assume the posterior mean does not change too much in the next few allocations. With this assumption, $\kappa(\rho, d(X^m))$ does not depend on the actual outcomes but just on the allocation n, so there is no need to take expectation over X^m. The changes to the above static allocation are: in Equation 9, replace $r(\rho, (e, d))$ by $\kappa(\rho^m, d(X^m))$ and replace δ_i by $-|\mu_{i,B}^{(0)} - \mu_{i,A}^{(0)}|$.

After obtaining the solution to the static allocation problem, there are several ways to use the static allocation to design a sequential allocation algorithm. For example, one can select the next profile probabilistically according to *Multinomial*(n_i/m), giving priorities to profiles that have higher static allocation. For the special case when $m = 1$, this corresponds to selecting the next profile that would decrease the decision risk the most. We call this algorithm "Look-ahead (LA)", which is sketched in Algorithm 1.

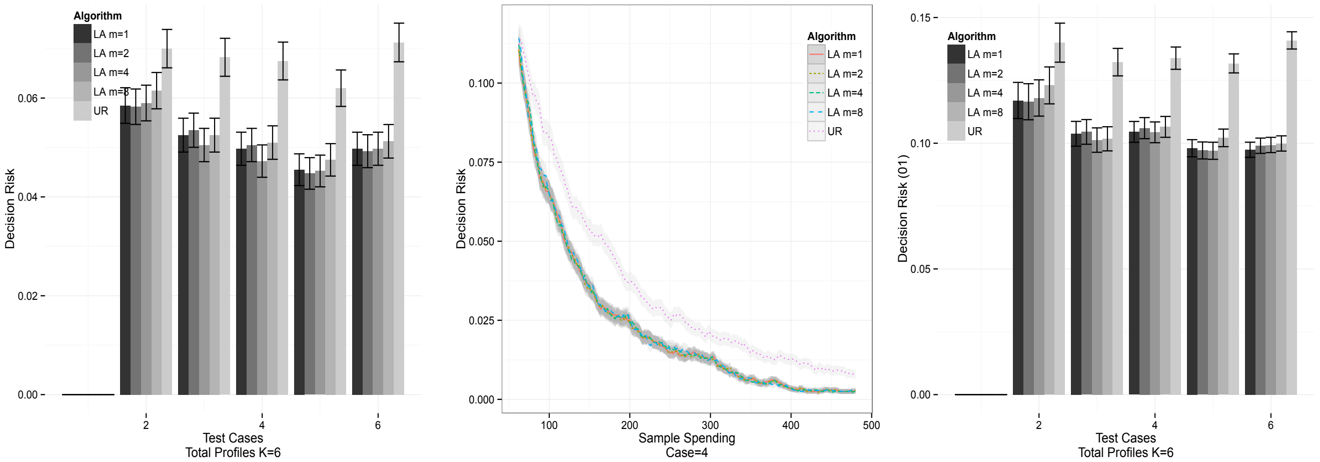
Algorithm 1 The look-ahead algorithm LA

```
Inputs. $m$: the look-ahead step; $2T$: the budget; $2w$: initial probe per profile
1. Initially, recruit $2w$ patients from each profile; compute the posterior distribution $\rho$.
2. while there is some budget left do
    a. Using $\rho$ as prior, solve the static allocation with $m$ to obtain (approximately) optimal allocations $\{n_i\}$ (note $\sum n_i = m$)
    b. Choose the next profile probabilistically according to Multinomial$(n_i/m)$.
    c. Recruit a pair of patients from the selected biomarker profile; obtain response pair, then update $\rho$.
end while
```

IV. Experiments

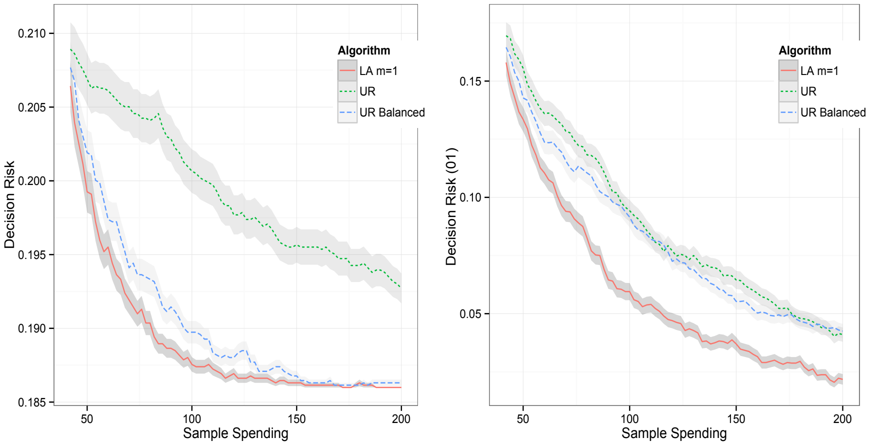
In this section, we evaluate the LA algorithm and discuss several practical issues. Although the formulation is Bayesian, the evaluation is "frequentist": for a fixed but unknown μ, we assess how well the algorithm minimizes the expected hinge loss in repeated experiments.

We generated test cases in the following way. For an experiment with K profiles, we generate a total of K test cases. E.g. for $K = 6$ and $\delta = 0.5$, test case 1 has no useful profile, with $\mu = (0, 0, 0, \cdots, 0)$. Test case 2 has 1 useful profile, with $\mu = (\delta, 0, 0, \cdots, 0)$ and so on. We choose standard error $\sigma = 1$, so that $\delta/\sigma = 0.5$; note this matches a minimal effect commonly regarded as significant by clinicians. Note our experimental setups are actually worst–case designs; the problem becomes easier if some profiles have smaller (nonzero) effects, as this permits an algorithm to focus some resources on these easier profiles. The initial prior is almost flat, with $\zeta = (0, 0, ...0)$, and $\tau^2 = 100^2$. We set the total budget to be $20 \times K$, which means on average 20 patients per profile. Finally, each profile/treatment combination is initially allocated with $w = 5$ patients, and each evaluation of a test case is an average of 2000 repetitions.



(a) (b) (c)

Figure 1. Basic properties of the Look-ahead algorithm. (a) result for the hinge loss; (b) consistency of the LA algorithm; (c) result for 0-1 loss



(a) (b)

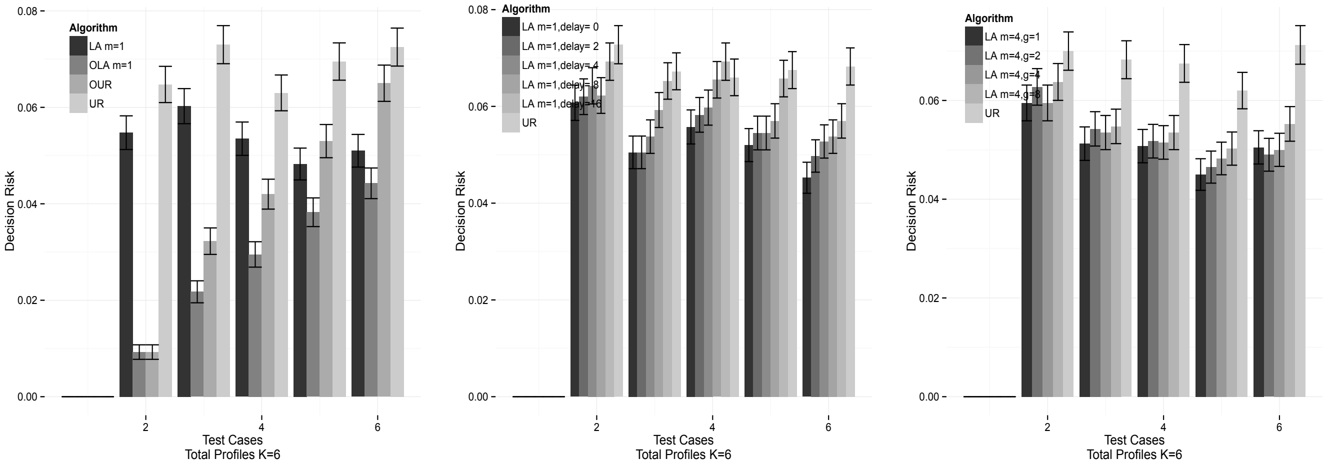
Figure 2. The performance on the Nefazodone-CBASP trial data. The algorithms are the look-ahead algorithm (LA), and the Uniform-Random algorithm (UR) and the balanced UR. (a) result for the hinge loss; (b) result for 0-1 loss

A. *Look-ahead (LA) vs. Uniform Random (UR) Algorithm*

Our experiments show that the Look-ahead algorithm is robust to the number of look-ahead steps m. Figure 1(a) plots the decision risk of the test cases for $K = 6$, where the x-axis indexes the test cases and the y-axis is the decision risk. The algorithms shown are the Look-ahead algorithms (denoted as LA) with different look-ahead steps m and the Uniform-Random (UR) algorithm, which recruits the next patient pair uniformly at random from the whole population. In these sets of experiments, we assume that each of the biomarker profiles has equal presence in the whole population.

Second, as the budget increases, the decision risk (expected hinge loss) converges to 0. This is true for all the test cases in our experiments. Figure 1(b) shows an example of diminishing decision risk (y-axis) as the sample spending increases for one of the test cases (case 5). The color-shaded regions show the variability of the decision risk over repeated runs. The total budget has been increased to 480 only to illustrate the trend. Figure 1(b) also shows significant reduction of total recruitments with the LA: to drive the decision risk below 0.025, LA needs about 200 participants while UR needs about 25% more participants.

Third, we evaluated 0-1 loss: In Figure 1(c), the y-axis is the expected **0-1** decision risk among only **the useful profiles**. For example, in the test case 4, the truth is that profiles 1,2,3 are useful profiles, while profiles 4,5,6 are the futile profiles, which respond equally well to either



(a) (b) (c)

Figure 3. (a) the oracle effect, (b) the delay effects and (c) the staged allocations.

treatment (i.e., the treatment does not matter). If a decision is $(A, \mathbf{B}, A, *, *, *)$, then the loss is $1/3 \approx 0.33$ for any values of "$*$"s. We notice that the Look-ahead algorithm performs very well in minimizing 0-1 loss. Also our results suggest that there is smaller variation across all algorithms (including the UR algorithm), which is likely due to the binary nature of the 0-1 loss function.

B. Case Study

To demonstrate the new algorithm, we use the data collected from the Nefazodone-CBASP trial [22], [23] dealing with depression, which randomized 681 patients to Nefazodone (NFZ), cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments (COMBO). For the 440 patients assigned to either NFZ or COMBO, Gunter et al. [23] suggested 2 potential variables–Obsessive Compulsiveness (OC) and Alcohol–for making treatment decisions. The outcome in their analysis was based on the 24-item Hamilton Rating Scale for Depression (HAMD) score [24].

We use the trial data to form a patient population, and draw patients with replacement from the 440 patients. We followed [23] by excluding the mono CBASP treatment in our simulation. We created 4 disjoint profiles based on the above two variables: profile 1: OC=NO, Alcohol=Absent; profile 2: OC=NO, Alcohol=Abuse; profile 3: OC=NO, Alcohol=Dependence; profile 4: OC=YES. Table I shows that in this population, there is a very small differential treatment effect $(2.81 - 2.73 = 0.08)$ in profile 1, and larger treatment effects in the remaining profiles. Note very few patients belong to profile 4. We assume that the differential treatment effect is significant enough to pursue for profile 2, 3, 4.

The simulation involves 2000 trials, each with a total budget of 200, which is much smaller than the original sample size of 440. Figure 2(a) shows that the Uniform Random (UR) performed fairly poorly, as the useful profiles in this dataset are rare. The algorithm UR Balanced forces equal randomization among these 4 profiles. Figure 2(b) shows that LA was able to drive the 0-1 decision risk down much more quickly than the other two algorithms.

Table I
SUMMARY OF THE SIMULATED POPULATION, CHARACTERIZED BY OBSESSIVE COMPULSIVENESS (OC), ALCOHOL, AND THE RANDOMIZED TREATMENT(1 FOR NFZ, 0 FOR COMBO). THE OUTCOMES ARE NORMALIZED TO HAVE UNIT VARIANCE.

OC	Alcohol	Treatment	Percentage of patients	Mean Outcome
No	Absent	0	33%	2.81
	(Profile 1)	1	34%	2.73
	Abuse	0	9.1%	2.50
	(Profile 2)	1	8.6%	2.69
	Dependence	0	5.2%	2.89
	(Profile 3)	1	5.5%	3.26
Yes		0	1.5%	2.30
	(Profile 4)	1	1.5%	1.10

C. Oracle Effect

To better understand our algorithms, we consider unrealistically powerful algorithms that have access to an oracle that identifies which profiles have differential effects. In particular, the "Oracle Uniform Random" (OUR) algorithm excludes the profiles with no differential effect, and proceeds as uniform random sampling over the remaining profiles. The "Oracle Look-ahead" (OLA) algorithm excludes profiles with no effect and otherwise proceeds as the LA algorithm. Figure 3(a) shows the decision risks of these algorithms for different test cases. Clearly, the use of prior knowledge to avoid recruiting from futile profiles should improve both the Random and Look-ahead algorithms. The Oracle Look-ahead (OLA) algorithm fared best among all four algorithms. Notice that the regular LA algorithm **can** sometimes outperform the OUR algorithm; this is a little surprising given that the look-ahead algorithm has little knowledge about the futile profiles at the beginning of the trial. But notice this only happened for test case 5 (with 4 useful profiles) and test case 6 (with 5 useful profiles): when the prior knowledge of which profiles are futile does not give too much advantage to an oracle algorithm (i.e., eliminating only 1 or 2 profiles). The LA algorithm, despite "wasting" draws by recruiting in the "futile profiles", was able to outperform OUR, because it took advantage of the *realized* trajectory of responses (i.e., signals in the data)

Nonetheless, the performance of the OLA algorithm indicates potential improvement by incorporating early-stopping rules to terminate sampling from futile profiles.

D. Delayed Observation of Outcomes and Solutions

In real life, there is a delay between a patient treatment assignment and the observation of the outcome. An exploration policy may therefore be forced to make decisions without observing the outcomes of patients who are still in treatment. The queueing characteristics of a clinical trial can be quite complex in real life, and to gain some understanding, we study (for simplicity), a deterministic, constant delay model. This model assumes, at any time, q pairs of patients are currently in treatment, whose outcomes are unavailable for the lookahead algorithm at that time. This means when a new pair is assigned, the earliest of the q pair patients just finishes treatment, thus the queue size is kept constant.

Figure 3(b) suggests that the advantage of the look-ahead algorithm will be hampered by delays. To alleviate this problem, one can run the trial in semi-staged fashion, i.e. waiting for a group of patients to finish the treatment before assigning new patients to treatment. This strategy has the effect of increasing the departure rate as compared to the arriving rate. In Figure 3(c), the lookahead algorithm is modified to assign g pairs of patients at the same time (for various values of g), whose outcomes are simultaneously observed before making decisions for the next batch of patients. We see that the advantage of the look-ahead algorithm is restored.

E. Comparison with Frequentist Algorithms: Preliminary Results and Discussion

Several frequentist's approaches with a similar objective have been studied in [13]–[15]. Except for some differences in modeling assumptions, they all aim to minimize the maximal probabilities of picking the wrong treatment over all profiles [14]. The "minmaxpcs" algorithm in [14], under the experimental setup described in Section IV, favors profiles with small $\hat{\delta}_i/\hat{\sigma_i}$, where $\hat{\delta}_i$ is the sample mean of treatment effect, and $\hat{\sigma_i}$ is the sample standard deviation of the treatment effect. This is simply the t-statistic. Under our experimental setup, "ucbe" [13], [15] (an upper confidence bound (UCB) based algorithm) selects a profile with the maximal $-\hat{\delta}_i + \sqrt{\eta G H^{-1} n_i^{-1}}$, in which G is the total budget, H is an intrinsic parameter characterizing the difficulty of a problem, n_i is the number of times a profile has been selected, η is a tuning parameter. That is, a profile is preferred if it seems hard to distinguish the two treatments or it has not been explored enough. For our experimental setup however, the value of H is ∞, because we assume that some profiles have no differential treatment effects. In our experiments, we used instead $-\hat{\delta}_i + \sqrt{\eta \log(G) n_i^{-1}}$ where $\eta = K^{-1}$ and K is the number of profiles, and we use $\log(G)$ instead of G, following the formulation in the original UCB algorithm. Examining equation (3) or the equation (10), the LA algorithm proposed in this paper is also a confidence bound based algorithm *in spirit*, albeit developed in a Bayesian framework.

Figure 4(a) shows the 0-1 decision risks of these algorithms for different test cases under a small (but more realistic) budget. The performances of the two frequentist's algorithms are on par with the LA algorithm. However, the "ucbe" algorithm converges quite slowly when we increase the budget to 480 as shown Figure 4(b).

V. Future Work and Contributions

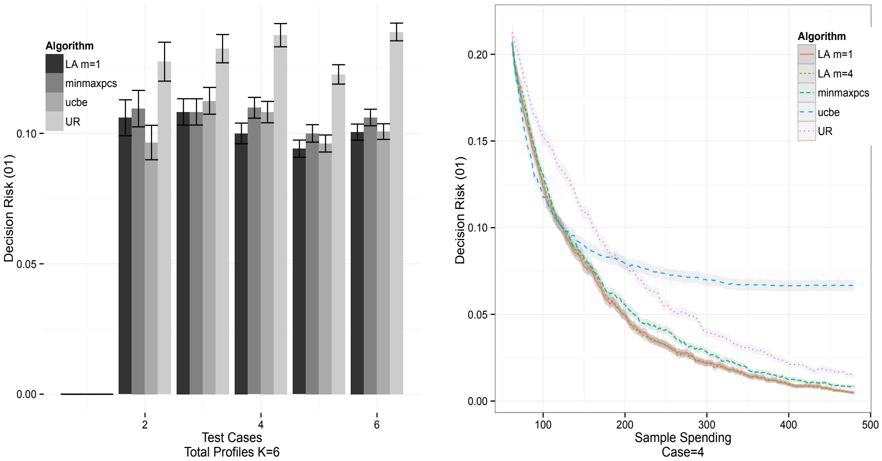
We have identified several interesting new directions to pursue: *a*) The current approach will devote resources to find a winner among the treatments for a profile, even when there is none. Are there ways to incorporate effective stopping rules, allowing the system to give up on such profiles. Note this is especially challenging when the effect size or the budget is small. *b*) We should relax the assumptions of *paired* outcomes. *c*) Currently, our system does the best it can with the total budget, expressed as the parameter T. It would be useful to estimate the budget required to ensure a certain quality of performance – to determine whether the given budget is sufficient to produce a high quality decision rule. *d*) We plan to extend our preliminary results in Section IV-D perhaps by staged allocations.

In summary, we identified and formulated a clinical trial problem as the budgeted learning problem, providing a framework and experimentally showing that a simple heuristic work very effectively for realistic experimental conditions.

References

[1] M. Trusheim, E. Berndt, and F. Douglas, "Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers," *Nat Rev Drug Discov*, vol. 6, no. 4, pp. 287–293, 2007.

[2] S. J. Mandrekar and D. J. Sargent, "Predictive biomarker validation in practice: lessons from real trials," *Clinical Trials*, vol. 7, no. 5, pp. 567–573, 2010.



(a) (b)

Figure 4. comparison with “minmaxpcs” and “ucbe” (a)small budget, (b) bigger budget

[3] S. Assmann, S. Pocock, L. Enos, and L. Kasten, “Subgroup analysis and other (mis) uses of baseline data in clinical trials,” *The Lancet*, vol. 355, no. 9209, pp. 1064–1069, 2000.

[4] Y. Song and G. Chi, “A method for testing a prespecified subgroup in clinical trials,” *Statistics in Medicine*, vol. 26, no. 19, 2007.

[5] M. Alosh and M. Huque, “A flexible strategy for testing subgroups and overall population,” *Statistics in Medicine*, vol. 28, no. 1, pp. 3–23, 2008.

[6] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee, “Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine,” *Clinical Trials (London, England)*, vol. 5, no. 3, pp. 181–193, 2008.

[7] D. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. Springer, Oct. 1985.

[8] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire, “Contextual bandit algorithms with supervised learning guarantees,” *arXiv:1002.4058*, Feb. 2010.

[9] P. Rigollet and A. Zeevi, “Nonparametric bandits with covariates,” *arXiv:1003.1630*, Mar. 2010.

[10] D. Lizotte, O. Madani, and R. Greiner, “Budgeted learning of naive-bayes classifiers,” in *UAI03*, 2003, pp. 378–38.

[11] O. Madani, D. Lizotte, and R. Greiner, “Active model selection,” in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 357–365.

[12] A. Antos, V. Grover, and C. Szepesvári, “Active learning in multi-armed bandits,” in *ALT*, 2008, pp. 287–302.

[13] J. Y. Audibert, S. Bubeck, and R. Munos, “Best arm identification in multi-armed bandits,” in *COLT*, 2010.

[14] K. Deng, J. Pineau, and S. Murphy, “Active learning for developing personalized treatment.” in *UAI*, 2011, pp. 161–168.

[15] V. Gabillon, M. Ghavamzadeh, A. Lazaric, S. Bubeck *et al.*, “Multi-bandit best arm identification,” *NIPS*, 2011.

[16] A. Wald, *Sequential analysis*. Dover Publications, 2004.

[17] J. Berger, *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[18] C. Chen, J. Lin, E. Yucesan, and S. Chick, “Simulation budget allocation for further enhancing the efficiency of ordinal optimization,” *Discrete Event Dynamic Systems*, vol. 10, no. 3, pp. 251–270, 2000.

[19] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998.

[20] M. Hauskrecht, “Value-function approximations for partially observable markov decision processes,” *JAIR*, 2000.

[21] W. Lovejoy, “A survey of algorithmic methods for partially observed markov decision processes,” *Annals of Operations Research*, vol. 28, 1991.

[22] M. B. Keller and J. e. a. McCullough, “A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression,” *New England Journal of Medicine*, vol. 342, no. 20, p. 1462, 2000.

[23] L. Gunter, J. Zhu, and S. Murphy, “Variable selection for qualitative interactions,” *Statistical Methodology*, vol. 8, no. 1, pp. 42–55, 2011.

[24] M. Hamilton, “Development of a rating scale for primary depressive illness,” *British Journal of Social and Clinical Psychology*, vol. 6, no. 4, pp. 278–296, 2011.