Learning Personalized Treatment Strategies with Predictive and Prognostic Covariates in Adaptive Clinical Trials

Andres Alban
Harvard Medical School, aalban@mgh.harvard.edu

Stephen E. Chick
INSEAD, stephen.chick@insead.edu

Spyros I. Zoumpoulis
INSEAD, spyros.zoumpoulis@insead.edu

We consider the problem of sequentially allocating sample observations to learn personalized treatment strategies, motivated by the design of adaptive clinical trials that aim to learn the best treatment as a function of patient covariates. In such settings there may be clinical knowledge of which covariates are predictive (they may interact with the treatment choice) and which are prognostic (they may influence the outcome independent of treatment choice). We extend the expected value of information (EVI)/knowledge gradient framework to develop useful heuristics for a context with predictive and prognostic covariates and a delay in observing outcomes. We also propose and analyze closely related Monte Carlo-based allocation policies to enhance our proposal's computational efficiency and applicability for adaptive contextual learning. We show that several of our proposed allocation policies are asymptotically optimal in learning treatment strategies. We run simulation experiments motivated by an application for clinical trial design to assess potential treatments of sepsis. We illustrate that the proposed EVI-based allocation policies, with knowledge about which covariates are predictive and prognostic, can improve the rate of inference relative to some existing approaches to adaptive contextual learning.

Keywords: Adaptive Clinical Trial Design; Contextual Bandit; Sequential Learning; Bayesian Optimization; Expected Value of Information; Sepsis

History : Working paper as of July 8, 2022.

Electronic copy available at: http://ssrn.com/abstract=4160045

Learning Personalized Treatment Strategies with Predictive and Prognostic Covariates in Adaptive Clinical Trials

Andres Alban
MGH Institute for Technology Assessment, Harvard Medical School, Boston, MA, USA 02114, aalban@mgh.harvard.edu

Stephen E. Chick
Technology and Operations Management, INSEAD, Fontainebleau, FRANCE 77300, stephen.chick@insead.edu

Spyros I. Zoumpoulis
Decision Sciences, INSEAD, Fontainebleau, FRANCE 77300, spyros.zoumpoulis@insead.edu

We consider the problem of sequentially allocating sample observations to learn personalized treatment strategies, motivated by the design of adaptive clinical trials that aim to learn the best treatment as a function of patient covariates. In such settings there may be clinical knowledge of which covariates are predictive (they may interact with the treatment choice) and which are prognostic (they may influence the outcome independent of treatment choice). We extend the expected value of information (EVI)/knowledge gradient framework to develop useful heuristics for a context with predictive and prognostic covariates and a delay in observing outcomes. We also propose and analyze closely related Monte Carlo-based allocation policies to enhance our proposal's computational efficiency and applicability for adaptive contextual learning. We show that several of our proposed allocation policies are asymptotically optimal in learning treatment strategies. We run simulation experiments motivated by an application for clinical trial design to assess potential treatments of sepsis. We illustrate that the proposed EVI-based allocation policies, with knowledge about which covariates are predictive and prognostic, can improve the rate of inference relative to some existing approaches to adaptive contextual learning.

Key words: adaptive clinical trial design, contextual bandit, sequential learning, Bayesian optimization, expected value of information, sepsis

History: Working paper as of July 8, 2022.

Precision medicine tailors the treatment to patient characteristics and has received considerable attention in recent years (NIH 2015, US FDA 2021). It has provided significant health improvements in cancer treatment (Tsimberidou et al. 2020) and is expected to lead to advancements in other areas, such as cardiovascular disease (Lee et al. 2012), neurology (Cutter and Liu 2012), and sepsis (Rello et al. 2018, Seymour et al. 2019). Due to the increasing number of medical tests and interventions, a key aspect of developing personalized treatments is designing clinical trials that efficiently find the best interventions for each type of patient (Berry 2011, Opal et al. 2014, Schork 2018). This paper is motivated by the design of sequential clinical trials to identify the best personalized interventions.

We study the sequential allocation of subjects to a finite set of alternative treatments to identify the best *treatment strategy*. A treatment strategy assigns a treatment based on the subject's characteristics, and "best" is defined as maximizing the expected mean reward. Because the best strategy is not initially known, the trial manager enrolls subjects

sequentially in the trial to learn the best strategy. While the trial manager can observe many characteristics, she may have insights into which are the relevant characteristics that interact with one or more treatments and can lead to personalized treatments. Using terminology from the medical literature, we refer to such characteristics as *predictive* (sometimes called moderators, e.g., Wu and Zumbo 2008). Other characteristics do not interact with treatments but may affect the outcome. We use the medical term *prognostic* to refer to these characteristics (e.g., Oldenhuis et al. 2008). Predictive covariates can help identify groups of patients that benefit from personalized treatment, and prognostic covariates improve the precision of estimating the benefits of personalized treatment (US FDA 2021).

To illustrate predictive and prognostic covariates, consider treatments for cancer patients that target the presence or overexpression of specific proteins in the tumor (Tsimberidou et al. 2020). In this scenario, the presence of such proteins is a predictive characteristic because it interacts with some treatments. Other characteristics, such as age and comorbidities, are prognostic because they influence the patient's outcome regardless of the treatment. This paper considers the problem of designing a fully sequential clinical trial with a fixed budget in order to find the best treatment strategy in the presence of prognostic and potentially predictive covariates.

Our model can be thought of as a contextual bandit. It is contextual in that subjects sequentially arrive with observable covariates, and treatments can be selected based on their values. After enrolling a finite number of subjects and observing their outcomes, a treatment strategy is selected for implementation in the patient population. We seek to optimize the selected treatment strategy, i.e., a function that maps observable covariates to treatments, instead of identifying a single overall best alternative, as is more typical in the ranking and selection (R&S) literature (Kim and Nelson 2006).

Four distinguishing features of the problem we consider are the following:

1. We seek to exploit structural knowledge, as may be available in precision medicine trials, regarding the potential for covariates to interact with treatments (i.e., to be predictive) or to influence the outcome independent of treatment (i.e., prognostic).
2. The trial manager does not fully control the subject's characteristics: patients arrive at the trial sites with random covariates from a known distribution.
3. Rewards are obtained at the conclusion of the trial, when treating post-trial patients. Patients enrolled in the trial are treated to learn the best treatment strategy.

4. Outcomes may be observed with a fixed *delay* after treatment has begun.

While different communities consider these features, to the best of our knowledge we are the first to take all of them into account. Biostatisticians have studied Feature 1 to reduce the variance of estimators of the mean effectiveness of treatments and sharpen comparisons between treatments, but have tended not to deal with Feature 3. Feature 2 is standard in contextual bandits, which tends not to consider Features 1 and 3. R&S maximizes the rewards from Feature 3 but tends not to model Features 1 and 2. Most sequential learning work in these streams, discussed further in Section 1, does not account for delay (Feature 4) in observing outcomes. Work on batch allocation or parallel Bayesian optimization can be adapted to account for delay, but does not account for Feature 1.

This paper assumes that the patient outcomes are continuous-valued, modeled with a normal distribution with an unknown mean, where the mean is given by a linear response function that depends on the treatment and the subject's covariates. We learn the unknown mean using a Bayesian conjugate inference model that infers the values of the parameters of the linear response model. The response model allows for the mean rewards to be correlated across treatments for a given set of covariates (to model that different treatments may have similar features) or across covariates for a given treatment (to model that a treatment may work similarly well for multiple patient types).

Although we focus here on clinical trial applications, nonclinical applications exist, such as (i) online advertising, where the alternative ads target specific characteristics of the users, while some (prognostic) user characteristics may affect the outcome (e.g., click or not on the ad) independent of the ad, and (ii) engineering testing, where the uncertain environment represents the subjects and the alternative system configurations represent the treatments. Despite broader potential applicability, the sequential learning literature tends to assume that covariates are all predictive (involving more parameters to estimate), all prognostic (missing the chance to learn which treatment is best for which subpopulation), or not relevant, as discussed in Section 1.

The paper makes the following contributions:

- **Modeling contribution.** We present what appears to be the first fully sequential learning model that accounts for both predictive and prognostic covariates with either Bayesian techniques or offline rewards (Section 2).

- **Algorithmic contributions.** Although the optimal policy can be characterized by Bellman's equation (Section 3), its computation suffers from the curse of dimensionality, and we are motivated to provide pragmatic heuristics (Section 4). To the best of our knowledge, this is the first work to extend the so-called expected value of information (EVI) or knowledge gradient (KG) framework to develop useful heuristics for a context with predictive and prognostic covariates and a delay in observing outcomes after treatment decisions. We call the main proposed heuristic fEVI because it maximizes the EVI of myopic lookahead allocations of patients to treatments to learn the best treatment strategy, a function (f) from covariates to treatments. We also propose and analyze closely related Monte Carlo-based allocation policies to enhance our proposal's computational efficiency and applicability for adaptive contextual learning.
- **Analytical contributions.** We show that the proposed EVI-based heuristics are asymptotically optimal. The justification of the theoretical guarantees involves some novel variations on standard EVI/KG proof structures (Section 5 and Appendix EC.3).
- **Pragmatic contributions.** Simulation experiments motivated by an application to sepsis treatment (Section 7) show that knowledge that a covariate is prognostic, rather than predictive, can improve the speed of inference of the best treatment strategy. Our numerical results suggest that there is promise for using our proposed fEVI family of allocation policies in the design of a clinical trial to assess the potential of precision medicine for treating sepsis (Singer et al. 2016, van Mourik et al. 2022), a leading driver of global mortality (WHO 2020) and hospital costs (Paoli et al. 2018).

We delineate some areas that lie outside of our scope in this work. We are exploring the design of a trial to assess the potential for precision medicine for sepsis treatment with medical collaborators. We model adaptive contextual learning problems (multiarm trials) that can be highly sequential and where the goal is to identify the treatment with the highest mean for real-valued outcomes. There are trials where the heuristics, as proposed, are less suitable. Bernoulli trials are not formally analyzed. More work is required to assess the identification of covariates as predictive or prognostic as the trial proceeds. We do not handle multiple longitudinal measures through time (as do Anderer et al. 2022). The delays considered here are short enough relative to the enrollment period to allow adaptations in treatment allocation. Multiyear survival studies are less suited to the model as proposed.

We do not consider covariates of interest that are observable with a delay or with measurement error. We discuss adaptations of the model to handle practical issues that may arise during trials, or that may provide paths for future research in Appendix EC.6.

1. Related Literature and Contributions

This work is linked to several streams of literature. This includes work on using covariates in the clinical trial context, both for analyzing clinical trial data and designing clinical trials. There are also related methods for learning treatment strategies, including contextual bandits and R&S. We also discuss the structural assumptions about the functional form of the mean response of patients to treatments and its interaction with covariates.

Covariates and Clinical Trials. The use of clinical trial data for identifying subgroups of patients that have an exceptional response to treatment has been studied under the subgroup selection area of biostatistics (Foster et al. 2011, Lipkovich et al. 2017). Interesting approaches include information theory (Sechidis et al. 2018), which ranks covariates by their predictive and prognostic value, and optimization (Bertsimas et al. 2019b). That literature focuses on analyzing already collected clinical trial data to identify subgroups, while we aim at designing trials that choose how to allocate patients to treatments to more efficiently learn predictive and prognostic effects.

The use of prognostic covariates for clinical trials focuses on balancing covariates across the treatment arms, most notably using the "biased coin" approach (Pocock and Simon 1975), which adjusts the randomization probabilities to favor the balancing of covariates. Optimization approaches have been proposed (Bertsimas et al. 2019a, Bhat et al. 2020). That line of research assumes that all covariates are prognostic, aiming to learn average treatment effects in a population, and does not tailor treatment to patient covariates.

The literature on trial design for precision medicine is expanding. Prominent examples include the BATTLE trial (Zhou et al. 2008) for lung cancer and the I-SPY 2 breast cancer platform trial (Barker et al. 2009, Berry 2011, Wang and Yee 2019). Initial work focused on predictive covariates, but increasing attention is given to both predictive and prognostic effects (Symmans et al. 2018, US FDA 2021), our focus here. Lai et al. (2013) provide a trial design using multi-armed bandit policies, generalized likelihood ratio tests, and ranking and selection ideas to design a trial that accomplishes three goals: a) give trial patients good treatment, b) identify an effective treatment strategy, and c) demonstrate its effectiveness. We model prognostic covariates and use them to help speed up learning, while Lai et al. (2013) rely on randomization to balance them across arms.

Sampling to learn the best alternative. The ranking and selection (R&S) literature (reviews include Kim and Nelson 2006, Chick 2006) focuses on finding the best alternative among a small set focusing on offline rewards. This rich literature primarily focuses on selecting the best overall treatment rather than a treatment strategy. More recently, R&S has been extended to account for covariates (Pearce and Branke 2018, Gao et al. 2019, Xiong 2020, Li et al. 2020, Ding et al. 2021, Shen et al. 2021). These papers assume control over both covariates and treatments for various simulation optimization problems. In contrast, our work assumes that the covariates are random arrivals, as is typical in clinical trials.

Another related stream of literature studies the classic bandit problem (e.g., Auer 2002). In particular, the contextual bandit literature focuses on designing policies that minimize the cumulative regret for enrolled patients, and several algorithms have been shown to achieve asymptotically optimal regret (Goldenshluger and Zeevi 2013, Russo and Van Roy 2014, Villar and Rosenberger 2018, Bastani and Bayati 2020, Bastani et al. 2021). Regret in that context is typically relative to online rewards (for patients in trial), whereas we focus on offline rewards obtained upon committing to a treatment strategy for future patients.

The contextual R&S and bandit works above consider all covariates to be predictive and do not exploit the potential gains in learning quicker using prognostic biomarkers. A recent working paper (Carranza et al. 2022) decomposes rewards of an online contextual bandit into treatment effect and confounder terms (analogous to our model below of predictive and prognostic covariates) and gives asymptotic, frequentist, online regret results.

Our approach extends the EVI/KG framework of Bayesian optimization (Chick and Inoue 2001, Frazier et al. 2008, Powell and Ryzhov 2012) to allocate patients to arms in a fully sequential manner before selecting a treatment strategy. Those works do not model covariates. Some contextual R&S work noted above uses the EVI approach (Pearce and Branke 2018, Ding et al. 2021). We model correlated outcomes across treatment types, building on Frazier et al. (2009) and Chick et al. (2021). We extend the methods of Alban et al. (2021), allowing for covariates to be continuous and prognostic, and we provide theoretical results. We also make a novel use of a result of Frazier et al. (2009) in an efficiency improvement for Monte Carlo estimates of the EVI when there are delayed outcomes.

Our model is related to other KG work as well. Ryzhov and Powell (2011) study a problem where the set of implementation decisions (choice of a path in a graph) are not in

one-to-one correspondence with the set of sampling choices (measure an arc of the graph). Similarly, our model's implementation decision (choice of a function from covariates to treatment options) differs from the sampling actions (choice of treatment for a patient). We build on Negoescu et al. (2011) for learning regression coefficients by incorporating structural knowledge about the predictive and prognostic nature of covariates. Chick and Inoue (2001) account for batching with stochastic outputs but use a less precise EVI approximation and do not model covariates. Wu and Frazier (2016), Wang et al. (2020), and Astudillo et al. (2021) explore batch and parallel allocations with Bayesian optimization. Those works share a commonality with our delayed observation model: multiple allocations are made before observing outcomes. We differ in that they can choose covariate values for the batch simultaneously, whereas our covariates arrive randomly and sequentially; and our observations are stochastic rather than deterministic. Finally, Wang et al. (2016) studied prognostic (but not predictive) covariates for 0-1 (rather than real-valued) outcomes.

Other related works on clinical trial design that do not focus specifically on covariates but that relate to our sequential learning include Williamson et al. (2017), Jacko (2018), Pallmann et al. (2018), Rojas-Cordova and Bish (2018). Williamson and Villar (2020) study Bayesian response adaptive multi-arm trials with normally distributed data. Williamson et al. (2022) compare random and fixed delays for 0-1 trials.

Structural Assumptions on Mean Response. When learning treatment strategies, it is necessary to make some assumptions about the *response function*, the expected outcome of a patient with a given set of covariates and treatment. When the space of covariates and treatments is finite, each response for each covariate-treatment combination can be learned independently and effectively model-free. Several papers use this approach (Lai et al. 2013, Villar and Rosenberger 2018, Gao et al. 2019, Li et al. 2020), which is a special case of our model. Those works assume that the covariates are predictive and only consider a small set of possible covariate values.

When we have many possibly continuous covariates, we make structural assumptions about the response function. One approach assumes a Gaussian process (GP) model (e.g., Xiong 2020, Ding et al. 2021). We assume a linear response function because it is a common model in the literature and tractable (e.g., Goldenshluger and Zeevi 2013, Bhat et al. 2020, Shen et al. 2021). Instead of learning a different model for each treatment (a standard assumption in multi-armed bandits, e.g., Bastani et al. 2021), we learn a single model in

which covariates can be predictive and prognostic (e.g., Lipkovich et al. 2017). We discuss possible implementations of model selection for the response function in Appendix EC.6.

Summary. There is much relevant work in the biostatistics, bandit, R&S, and simulation optimization literature. To the best of our knowledge, the literature on adaptive contextual learning has not yet studied models with both predictive and prognostic covariates in a context with offline rewards or delayed observations. We do so here using the EVI/KG framework to develop practical heuristics for allocating patients to arms to determine the best treatment strategy as a function of predictive covariates.

2. Mathematical Model of a Sequential Trial with Covariates

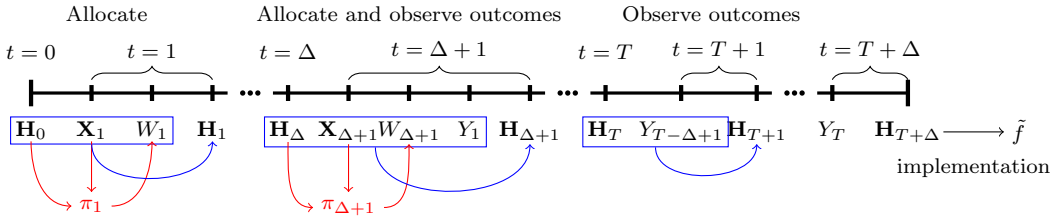
We formulate a model of a sequential clinical trial that enrolls a fixed number of patients. The trial manager assigns a treatment for each enrolled patient after observing the patient's covariates. We model the patient outcomes using a linear response function that allows for treatment effects, effects from predictive and prognostic covariates, and homoskedastic noise. The trial manager uses the patient outcomes, observable with some delay, to update sequentially in a Bayesian manner the posterior distribution for the parameters of the linear response model. The trial manager assesses all the information collected so far to decide on each subsequent patient's treatment allocation. After the trial, the trial manager has learned a treatment strategy, i.e., a mapping from patient covariates to treatments, which she then applies to a population of post-trial patients. The trial manager's objective is to learn the best treatment for each set of covariates in order to optimize expected cumulative outcomes in a post-trial patient population.

2.1. Information and decisions

Consider a trial with a budget to enroll T patients. The patients are enrolled sequentially in evenly spaced time steps, where the time steps $t = 0, 1, \ldots, T$ represent the number of patients that have been allocated to treatment. At time steps $t = 1, 2, \ldots, T$, the trial manager observes one patient with covariates $\mathbf{X}_t \in \mathscr{X} \subseteq \mathbb{R}^m$, and allocates that patient to treatment $W_t \in \mathscr{W} = \{1, 2, \ldots, n\}$. We fix T now, and discuss allowing T to be a response-adaptive stopping time in Appendix EC.6.

The patient's outcome $Y_t \in \mathscr{Y} \subseteq \mathbb{R}$ is observed after a fixed delay of $\Delta \in \{0, 1, \ldots\}$ time steps, so the trial manager observes at time $t + \Delta$ the outcome of the patient who was enrolled at time t. At time steps $t = T + 1, T + 2, \ldots, T + \Delta$, no more allocations are made

Figure 1 Timeline of decisions and information acquisition.



but patient outcomes are still observed. The response function $r_{\boldsymbol{\mu}}(\mathbf{X}_t, W_t) = \mathbb{E}[Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t]$ yields the expected outcome of the patient given an unknown vector of parameters $\boldsymbol{\mu}$, the covariates, and the treatment. Figure 1 summarizes the sequence of events. Table EC.1 in the Appendix summarizes the notation of the model.

At time $t = T + \Delta$ the trial concludes and each of P post-trial patients with covariates $\tilde{\mathbf{X}}_i \in \mathscr{X}$ receives treatment $\tilde{W}_i \in \mathscr{W}$ and obtains health outcome $\tilde{Y}_i$, where $i = 1, 2, \ldots, P$ indexes the post-trial patients. We assume here that P is a known constant. The expected outcome for post-trial patients is also assumed to be $r_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_i, \tilde{W}_i) = \mathbb{E}[\tilde{Y}_i \mid \boldsymbol{\mu}, \tilde{\mathbf{X}}_i, \tilde{W}_i]$. The distribution of the covariates may differ for patients in the trial, $\mathbf{X}_t$, and those post-trial, $\tilde{\mathbf{X}}_i$, as discussed below in Section 2.4.

The *history* $\mathbf{H}_t$ represents all the available data at time t. At time $t = 0$, the trial manager has not observed any patients, so $\mathbf{H}_0 = \emptyset$. For $1 \le t \le \Delta$ she has not observed any patient outcomes yet, so $\mathbf{H}_t$ includes only the covariates and treatments of the enrolled patients. For $t \ge \Delta + 1$, she observes delayed outcomes, so $\mathbf{H}_t$ continues to gather patient covariates and treatments and also includes delayed patient outcomes. For $t \ge T + 1$, she no longer enrolls patients, but delayed patient outcomes are included in $\mathbf{H}_t$ until time $T + \Delta$. Thus,

$$\mathbf{H}_t := \begin{cases} \emptyset, & \text{for } t = 0 \\ (\mathbf{X}_1, \ldots \mathbf{X}_t, W_1, \ldots, W_t), & \text{for } 1 \le t \le \Delta \\ (\mathbf{X}_1, \ldots \mathbf{X}_t, W_1, \ldots, W_t, Y_1, \ldots, Y_{t-\Delta}), & \text{for } \Delta + 1 \le t \le T \\ (\mathbf{X}_1, \ldots \mathbf{X}_T, W_1, \ldots, W_T, Y_1, \ldots, Y_{t-\Delta}), & \text{for } T + 1 \le t \le T + \Delta \end{cases} \tag{1}$$

An *allocation policy* $\pi = (\pi_t)_{t=1,2,\ldots,T}$ is a sequence of functions that map the available data, $\mathbf{H}_{t-1}$, to a probability distribution over the set of treatments for each of the possible covariate vectors; i.e., $\pi_t(w \mid \mathbf{h}, \mathbf{x}) = \mathbb{P}(W_t = w \mid \mathbf{H}_{t-1} = \mathbf{h}, \mathbf{X}_t = \mathbf{x})$ is the probability that allocation policy π assigns treatment w to a patient with covariates $\mathbf{x}$ at time t given data $\mathbf{h}$. Thus, an *action* is not a treatment but rather a probability distribution over the available treatments. The allocation policy assigns a treatment after observing the covariate vector,

and is specified for all possible patient covariate vectors. Because the allocation policy only uses past information, it is a member of the set of non-anticipatory policies Π.

A *treatment strategy* is a mapping from the space of covariates to the set of treatments. We denote the space of all treatment strategies by $\mathbf{f} = \{f : \mathscr{X} \to \mathscr{W}\}$. The *implementation decision* $\tilde{f} \in \mathbf{f}$ is the treatment strategy implemented for future patients: $\tilde{W}_i = \tilde{f}(\tilde{\mathbf{X}}_i)$.

We denote the probability measure induced by an allocation policy π by $\mathbb{P}^\pi$, and the expectation with respect to that measure by $\mathbb{E}^\pi$. When the allocation policy is irrelevant for the event under consideration, we omit it from the superscript; e.g., $\mathbb{E}[Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t]$ is conditioned on the choice of treatment, so the policy is irrelevant, while $\mathbb{E}^\pi[Y_t]$ depends on allocation policy through the assigned treatment.

2.2. Trial value

In our main model, we define the *trial value* with allocation policy π as the expected cumulative outcomes of the post-trial population:

$$V^\pi = \mathbb{E}^\pi\Big[\sum_{i=1}^{P} r_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_i, \tilde{f}(\tilde{\mathbf{X}}_i))\Big] = P\mathbb{E}^\pi\left[r_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_1, \tilde{f}(\tilde{\mathbf{X}}_1))\right]. \tag{2}$$

The expectation is over the prior distribution F_μ of $\boldsymbol{\mu}$ (Section 2.5), the distributions of covariates (Section 2.4), and the treatments allocated during the trial, which may be randomized by allocation policy π.

We assume the implementation decision is the treatment strategy that maximizes the value for future patients given the information collected from the trial:

$$\tilde{f}(\mathbf{x}) = \arg\max_{\tilde{w} \in \mathscr{W}} \mathbb{E}[r_{\boldsymbol{\mu}}(\mathbf{x}, \tilde{w}) \mid \mathbf{H}_{T+\Delta}] \; \forall \mathbf{x} \in \mathscr{X}.$$

Our objective is to find an allocation policy that maximizes trial value: $\sup_\pi V^\pi$. Because P is assumed fixed for now, it is sufficient to solve for the case $P = 1$ and we do so here. Sampling to find the best treatment strategy at the end of information collection optimizes so-called offline rewards. In Appendix EC.6 we discuss useful ways to extend this model.

2.3. Linear response function and labelings of active regressors

In general, a response function $r_{\boldsymbol{\mu}}(\mathbf{x}, w)$ maps a vector of covariates $\mathbf{x} \in \mathscr{X}$ and a treatment $w \in \mathscr{W}$ to the expected outcome. Here, we assume a linear model for $r_{\boldsymbol{\mu}}(\mathbf{x}, w)$ to describe the effects of covariates and their potential interactions with treatments.

Preliminary model with all regressors. Let $\mu_{i,l}$ for $i \in \{0,1,\ldots,n\}$ and $l \in \{0,1,\ldots,m\}$ be the linear coefficient associated with treatment i and covariate l. Here, $\mu_{0,l}$ for $l = 1,2,\ldots,m$ are the coefficients that represent effects of covariates regardless of treatment; $\mu_{i,0}$ for $i = 1,2,\ldots,n$ are the coefficients for treatment effects that influence mean outcomes independent of covariates. This motivates our preliminary linear response function:

$$r_{\boldsymbol{\mu}}(\mathbf{x}, w) = \underbrace{\mu_{0,0}}_{\text{intercept term}} + \sum_{i=1}^{n} \underbrace{\mathbb{1}_{w=i}\mu_{i,0}}_{\text{treatment effect}} + \sum_{l=1}^{m} \underbrace{x_l \mu_{0,l}}_{\text{prognostic term}} + \sum_{i=1}^{n}\sum_{l=1}^{m} \underbrace{\mathbb{1}_{w=i} x_l \mu_{i,l}}_{\text{predictive term}}, \tag{3}$$

where $\mathbb{1}$ is the indicator function. If the coefficient $\mu_{0,l}$ is non-zero, we say that covariate l is *prognostic*. If the coefficient $\mu_{i,l}$ is non-zero, we say that covariate l is *predictive* with respect to treatment i. If $\mu_{i,l} = 0$ for all $i = 0,1,\ldots,n$, then covariate l does not change the mean, and we say that covariate l is *idle*. Similarly, if $\mu_{i,l} = 0$ for all $l = 0,1,\ldots,m$, we say that treatment i is idle.

Active regressors. In applications, contextual knowledge may suggest that some of these coefficients are zero. We therefore consider response functions where some coefficients are constrained to be zero. Let the *labels* $\xi_{i,l}$ represent whether each coefficient is allowed to be nonzero: if $\xi_{i,l} = 0$, then $\mu_{i,l} = 0$; if $\xi_{i,l} = 1$, then $\mu_{i,l}$ is free, needs to be estimated, and is referred to as *potentially active*. We define the labeling $\boldsymbol{\xi}$ to be the matrix of labels. The set of indices of potentially active coefficients is denoted $\Xi := \{(i,l) : \xi_{i,l} = 1\}$.

We assume the trial manager knows the labels based on expert knowledge. Our *base model* for the response function for a fixed $\boldsymbol{\xi}$ is

$$r_{\boldsymbol{\mu}}(\mathbf{x}, w) = \mu_{0,0}\xi_{0,0} + \sum_{i=1}^{n} \mathbb{1}_{w=i}\mu_{i,0}\xi_{i,0} + \sum_{l=1}^{m} x_l \mu_{0,l}\xi_{0,l} + \sum_{i=1}^{n}\sum_{l=1}^{m} \mathbb{1}_{w=i} x_l \mu_{i,l}\xi_{i,l}. \tag{4}$$

To simplify these sums, we now introduce operators that will be useful to simplify the expression (4) in the sequel using matrix multiplication. We let $\boldsymbol{\mu}$ be a $(n+1)(m+1)$-dimensional column vector of the regression coefficients as follows:

$$\boldsymbol{\mu} = \Big(\underbrace{\mu_{0,0}, \mu_{0,1}, \ldots, \mu_{0,m}}_{\text{associated with no treatment}}, \underbrace{\mu_{1,0}, \ldots, \mu_{1,m}}_{\text{associated with treatment 1}}, \ldots, \underbrace{\mu_{n,0}, \ldots, \mu_{n,m}}_{\text{associated with treatment } n} \Big)^{\top}.$$

Let $\boldsymbol{\mu}_{\boldsymbol{\xi}} \in \mathbb{R}^{|\Xi|}$ be the vector containing only potentially active coefficients in $\boldsymbol{\mu}$, as indicated by the labeling $\boldsymbol{\xi}$, and let $(w \otimes_{\boldsymbol{\xi}} \mathbf{x}) \in \mathbb{R}^{1 \times |\Xi|}$ be the row vector of active regressors such that

$$r_{\boldsymbol{\mu}}(\mathbf{x}, w) = (w \otimes_{\boldsymbol{\xi}} \mathbf{x})\boldsymbol{\mu}_{\boldsymbol{\xi}}. \tag{5}$$

When we fix a known labeling and it is clear that we refer to it, we drop the $\boldsymbol{\xi}$ subscripts and use notation $r_{\boldsymbol{\mu}}(\mathbf{x}, w) = (w \otimes \mathbf{x})\boldsymbol{\mu}$ instead of $r_{\boldsymbol{\mu}}(\mathbf{x}, w) = (w \otimes_{\boldsymbol{\xi}} \mathbf{x})\boldsymbol{\mu}_{\boldsymbol{\xi}}$, where, with a slight abuse of notation, $\boldsymbol{\mu} \in \mathbb{R}^{|\Xi|}$ only includes potentially active coefficients.

EXAMPLE 1 (TWO PATIENT TYPES, TWO TREATMENTS). Consider a population with two types of patients, A and B, such that $\mathbf{X}_t = 0$ for type A and $\mathbf{X}_t = 1$ for type B. Consider two treatments, i.e., $\mathscr{W} = \{1, 2\}$. Let $\xi_{i,l} = 1$ for $i = 1, 2$ and $l = 0, 1$; and $\xi_{i,l} = 0$ for $i = 0$ and $l = 0, 1$; i.e., the treatment effect and predictive terms are potentially active, whereas the intercept and the prognostic terms are not. Then

$$(1 \otimes 0)\boldsymbol{\mu} = (1, 0, 0, 0)\boldsymbol{\mu} = \mu_{1,0}, \qquad (1 \otimes 1)\boldsymbol{\mu} = (1, 1, 0, 0)\boldsymbol{\mu} = \mu_{1,0} + \mu_{1,1},$$
$$(2 \otimes 0)\boldsymbol{\mu} = (0, 0, 1, 0)\boldsymbol{\mu} = \mu_{2,0}, \qquad (2 \otimes 1)\boldsymbol{\mu} = (0, 0, 1, 1)\boldsymbol{\mu} = \mu_{2,0} + \mu_{2,1}. \quad \square$$

2.4. Distribution of covariates

We assume the covariates of patients enrolled in the trial are exogenous, independent, and identically distributed (*i.i.d.*) random variables from a known distribution F_x: $\mathbf{X}_t \overset{i.i.d.}{\sim} F_x$. The covariates of post-trial patients are exogenous with a known distribution, $\tilde{\mathbf{X}}_i \overset{i.i.d.}{\sim} F_{\tilde{x}}$, which is not necessarily the same distribution as that of the trial patients. Allowing the distribution of covariates for patients enrolled in the trial to differ from that for patients treated post-trial provides useful flexibility, e.g., to account for trial inclusion-exclusion criteria, different disease mix near the trial centers versus in the general population, etc.

We impose two assumptions on the distributions of covariates F_x and $F_{\tilde{x}}$. Assumption 1 requires that $\mathbb{E}[\mathbf{X}_t^\top \mathbf{X}_t]$ and $\mathbb{E}[\tilde{\mathbf{X}}_t^\top \tilde{\mathbf{X}}_t]$ are bounded to ensure integrability:

ASSUMPTION 1 **(Integrability)**. *The distributions F_x of $\mathbf{X}_t$ and $F_{\tilde{x}}$ of $\tilde{\mathbf{X}}_i$ are such that $\mathbb{E}[\mathbf{X}_t^\top \mathbf{X}_t] < \infty$ and $\mathbb{E}[\tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i] < \infty$.*

This is similar to the standard assumption in regression analysis that $\mathbf{X}^\top \mathbf{X}$ is invertible, which ensures estimation can be done. Assumption 2 requires that patients to be treated post-trial be represented in the trial with nonzero probability. It will help ensure asymptotic convergence of proposed allocation policies to the optimal treatment strategy.

ASSUMPTION 2. *$F_{\tilde{x}}$ is* **absolutely continuous** *with respect to F_x.*

2.5. Bayesian modeling and updating

For a known labeling $\boldsymbol{\xi}$, we assume a conjugate normal model for patient outcomes and the parameters of the linear response function. We use the operator $\otimes$ (Section 2.3) to describe it.

ASSUMPTION 3 **(Conjugate normal model)**. *We assume the following model for patient outcomes and for the parameters of the response function:*

$$Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}((W_t \otimes \mathbf{X}_t)\boldsymbol{\mu}, \sigma^2)$$
$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0)$$

The parameters $\sigma^2 \in \mathbb{R}_{>0}$, $\boldsymbol{\theta}_0 \in \mathbb{R}^{|\Xi|}$, *and* $\Sigma_0 \in \mathbb{R}^{|\Xi|\times|\Xi|}$ *(symmetric positive semi-definite) are known and bounded, where* $\Xi := \{(i,l) : \xi_{i,l} = 1\}$.

With this model, the distribution of the unknown coefficients remains normally distributed after updating with Bayes' rule (Gelman et al. 2013, Chap. 14), with

$$\boldsymbol{\mu} \mid \mathbf{H}_t \sim \mathcal{N}(\boldsymbol{\theta}_t, \Sigma_t),$$

where $\boldsymbol{\theta}_t$ and Σ_t are updated after each observation. For $t \le \Delta$, we do not observe any outcomes and do not update, so $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0$ and $\Sigma_t = \Sigma_0$. For $t \ge \Delta + 1$, we use Bayes' rule and update with the following recursive equations:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \frac{Y_{t-\Delta} - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\boldsymbol{\theta}_{t-1}}{\sigma^2 + (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_{t-1}(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top} \Sigma_{t-1}(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top \tag{6a}$$

$$\Sigma_t = \Sigma_{t-1} - \frac{\Sigma_{t-1}(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_{t-1}}{\sigma^2 + (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_{t-1}(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}. \tag{6b}$$

We discuss specifying the prior distribution in Section 6 and provide more details on the Bayesian updating equations in Appendix EC.1.

We define a *pipeline state* $\mathbf{J}_t$ to contain the covariates and treatments of patients that have been assigned to treatment, but whose outcomes have not been observed at time t:

$$\mathbf{J}_t := \begin{cases} \emptyset, & \text{for } \Delta = 0 \text{ or } t = 0 \text{ or } t = T + \Delta \\ (\mathbf{X}_1, W_1, \ldots, \mathbf{X}_t, W_t), & \text{for } 1 \le t \le \Delta \\ (\mathbf{X}_{t-\Delta+1}, W_{t-\Delta+1}, \ldots, \mathbf{X}_t, W_t), & \text{for } \Delta + 1 \le t \le T \\ (\mathbf{X}_{t-\Delta+1}, W_{t-\Delta+1}, \ldots, \mathbf{X}_T, W_T), & \text{for } T + 1 \le t \le T + \Delta - 1. \end{cases}$$

We denote the knowledge state by $\mathbf{K}_t := (\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t)$. Note that $\mathbf{K}_t$ is a sufficient statistic for the unknown model parameters. Let τ_t be the transition function that updates the knowledge state recursively, with $\mathbf{K}_t = \tau_t(\mathbf{K}_{t-1}, \mathbf{X}_t, W_t, Y_{t-\Delta})$ for $t = 1, 2, \ldots, T + \Delta$, respecting Bayes' rule in (6) and the definition of the pipeline state above. The function τ_t has a subscript t to handle the cases $t \le \Delta$ and $t \ge T + 1$ where $\mathbf{X}_t$, W_t, or $Y_{t-\Delta}$ may not exist.

3. The Optimal Policy

In this section we describe the optimal policy. We first describe the implementation decision that maximizes trial value given the knowledge state after all data from the trial has been observed, at time $T+\Delta$. We then present a dynamic program for solving for the optimal allocation policy of patients to arms as a function of their covariates and show the existence of an optimal solution. The dynamic program is generally computationally infeasible to solve, yet provides the foundation for the development of expected value of information heuristics (Section 4) and their theoretical properties (Section 5).

Implementation decision. Recall our linear model, $r_{\boldsymbol{\mu}}(\mathbf{x}, w) = (w \otimes \mathbf{x})\boldsymbol{\mu}$, which has conditional mean $\mathbb{E}[r_{\boldsymbol{\mu}}(\mathbf{x}, w) \mid \mathbf{K}_t] = (w \otimes \mathbf{x})\boldsymbol{\theta}_t$ given knowledge at time t. The treatment that optimizes outcomes for a new patient with covariates $\mathbf{x}$ given knowledge at time t is

$$\tilde{f}_{\boldsymbol{\theta}_t}(\mathbf{x}) = \arg\max_{w \in \mathscr{W}} (w \otimes \mathbf{x})\boldsymbol{\theta}_t, \tag{7}$$

where the notation $\tilde{f}_{\boldsymbol{\theta}_t}$ emphasizes the dependence of the treatment strategy on $\boldsymbol{\theta}_t$, and not all of $\mathbf{H}_t$ or $\mathbf{K}_t$. The implementation decision given all trial data $\mathbf{H}_{T+\Delta}$ is therefore

$$\tilde{f}(\mathbf{x}) = \tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\mathbf{x}) = \arg\max_{w \in \mathscr{W}} (w \otimes \mathbf{x})\boldsymbol{\theta}_{T+\Delta}. \tag{8}$$

Dynamic program. We now write $\sup_{\pi \in \Pi} V^\pi$ as a dynamic program. Let $\mathscr{K}_t$ be the set of feasible knowledge states at time $t \in \{0, 1, \ldots, T\}$, and let $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j}) \in \mathscr{K}_t$ be a knowledge state at time t with posterior mean $\boldsymbol{\theta}$ and covariance matrix Σ for $\boldsymbol{\mu}$ and pipeline state $\mathbf{j}$.

By construction, the terminal reward $G(\mathbf{k})$ for knowledge states $\mathbf{k} \in \mathscr{K}_T$ that are feasible at time T, using the implementation decision in (8), is

$$G(\mathbf{k}) = \mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_T = \mathbf{k}] = \mathbb{E}[\max_{\tilde{w}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_T = \mathbf{k}], \tag{9}$$

with the expectation taken over the outcomes of the pipeline patients that are to be observed by time $T+\Delta$ and that help determine the optimal treatment strategy. Also by construction, the value-to-go $V_t^\pi(\mathbf{k})$ for an allocation policy π (not necessarily the optimal allocation policy), evaluated at time $t \leq T-1$ with state $\mathbf{k} \in \mathscr{K}_t$, is

$$V_t^\pi(\mathbf{k}) = \mathbb{E}^{\pi_{t+1}}[V_{t+1}^\pi(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}], \text{ for } t = 0, \ldots, T-1, \tag{10}$$

where π_t determines the treatment W_t, and τ_t updates the knowledge state as in Section 2.5. Thus, the optimal value-to-go upon treating patient t is given by the dynamic program:

$$\begin{aligned} V_t(\mathbf{k}) &= \sup_{\pi_{t+1}} \mathbb{E}^{\pi_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}], \text{ for } t = 0, 1, \ldots, T-1 \\ V_T(\mathbf{k}) &= G(\mathbf{k}). \end{aligned} \tag{11}$$

With this formulation, we can prove the existence of an optimal allocation policy π^* that solves $\max_{\pi \in \Pi} V^\pi$. The optimal allocation policy is deterministic (does not randomize treatments) and Markov (only depends on the knowledge state). Our proof of this result in Appendix EC.2 maps (11) to the model in Bertsekas and Shreve (1978, Chap. 8).

PROP. 1. *The dynamic program in (11) satisfies that* $V_0(\mathbf{K}_0) = \sup_{\pi \in \Pi} V^\pi$ *and there exists a deterministic Markov allocation policy* π^**, which can be computed using Bellman's recursion in (11), such that* $V_t(\mathbf{k}) = V_t^{\pi^*}(\mathbf{k})$ *for all* $\mathbf{k} \in \mathscr{K}_t$ *and all* $t = 0, \ldots, T$.

Because an optimal allocation policy is attainable, we use $V^* = \max_{\pi \in \Pi} V^\pi$ to denote the trial value attained by an optimal allocation policy π^*.

4. Heuristics for Learning with Predictive and Prognostic Covariates

The computation of the optimal allocation policy suffers from the curse of dimensionality. Thus, we develop heuristics to assign patients to arms that are computationally tractable, perform well, and have useful analytic properties. In this section, we extend the expected value of information approach of Bayesian sequential optimization to account for predictive and prognostic covariates. We will assess the theoretical properties in Section 5 and assess the performance numerically in Section 7.

4.1. Functional Expected Value of Information (fEVI)

We call this heuristic fEVI because it uses the EVI approach to learn a treatment strategy, a *function* from covariates to treatments. The fEVI allocation policy makes at each step the sampling decision that maximizes the expected value of information of one more sample.

In particular, the fEVI-index, $\nu_t(\mathbf{x}, w)$, is the value gained from allocating one additional patient with covariates $\mathbf{x}$ to treatment w, and selecting a treatment strategy after waiting to observe the outcomes of that patient and the patients in the pipeline, over selecting a treatment strategy using only the outcomes of the patients that have been allocated already, including the ones in the pipeline (all conditional on $\mathbf{K}_t$). For $0 \le t \le T-1$,

$$\nu_t(\mathbf{x}, w) = \mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}_{t+\Delta+1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_t, \mathbf{X}_{t+1} = \mathbf{x}, W_{t+1} = w] - \mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}_{t+\Delta}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_t]$$

$$= \underbrace{\mathbb{E}\Big[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta+1} \mid \mathbf{K}_t, \mathbf{X}_{t+1}=\mathbf{x}, W_{t+1}=w\Big]}_{\text{implement after one more patient and pipeline clears}} - \underbrace{\mathbb{E}\Big[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta} \mid \mathbf{K}_t\Big]}_{\text{implement after pipeline clears}}. \tag{12}$$

The fEVI allocation policy assigns the treatment with the highest index, $W_{t+1} = \arg\max_{w\in\mathscr{W}} \nu_t(\mathbf{X}_{t+1}, w)$, $0 \le t \le T-1$, breaking ties by sampling uniformly at random between treatments with the largest index.

4.2. Monte Carlo (MC) estimates of fEVI and randomization of treatments

The computation of the fEVI-index requires integration over a potentially high-dimensional space, and therefore quadrature can present challenges, particularly if $\Delta > 0$ or if there are multiple continuous-valued covariates. We thus propose an allocation policy that approximates the fEVI index using MC simulation.

MC simulation helps address the challenge of high-dimensional integration and also introduces some randomization in allocating treatments. Clinical trial practice may use randomization to arms to help address sampling biases. We therefore also present allocation policies that allow a clinical trial manager to further control randomization to arms. Section 7.4 presents experiment results for these allocation policies.

Monte Carlo fEVI allocation policy (fEVI-MC). The expectations in (12) that determine the fEVI-index can be estimated by simulating the unknown regression coefficients and the outcomes of pipeline patients; optimizing treatments conditional on that data; and estimating rewards for a simulated post-trial population. An advantage of this basic approach is that models that do not satisfy the conjugate normality of Assumption 3 can be handled. A potential disadvantage is computational speed.

If Assumption 3 is satisfied, as we assume here, we can improve computational efficiency in the simulation estimates of the fEVI-indices of (12). The fEVI-MC function in Algorithm 1 implements a MC estimator of the indices $\nu_t(\mathbf{x}, w)$ in (12) for each w and presents several efficiency improvements relative to the basic approach above. There are three sources of randomness to simulate: (i) the outcomes of Δ'_t pipeline patients that are observed through time $t+\Delta$, where $\Delta'_t = \min\{t, \Delta\}$ is the size of the pipeline at time t, (ii) the covariates of post-trial patients to estimate mean rewards, and (iii) outcomes of patient $t+1$ with covariate $\mathbf{X}_{t+1}$ for each treatment. If Assumption 3 is satisfied, then for (i) we propose to simulate η^{on} realizations of $\boldsymbol{\theta}_{t+\Delta}$ given $\boldsymbol{\theta}_t, \mathbf{K}_t$ in step 7 using standard

Algorithm 1 fEVI-MC: Estimate fEVI indices with common random numbers across pipeline and post-trial patients and with conditional Monte Carlo

1: **function** fEVI-MC($\mathbf{K}_t, \mathbf{X}_{t+1}; \eta^{\text{on}}, \eta^{\text{off}}$)
2: Let $\Delta'_t = \min\{t, \Delta\}$ ▷ Compute number of patients in pipeline.
3: $\hat{\boldsymbol{\sigma}} \leftarrow \tilde{\boldsymbol{\sigma}}(\Sigma_t, \mathbf{W}_{(t-\Delta'_t+1):(t)}, \mathbf{X}_{(t-\Delta'_t+1):(t)})$ ▷ Prepost std dev in (EC.5) for pipeline
4: Compute $\Sigma_{t+\Delta}$ ▷ Compute post var, after pipeline clears (6b)
5: **for** j in $\{1, \ldots, \eta^{\text{on}}\}$ **do** ▷ Compute offline rewards for η^{on} replications
6: $\hat{\mathbf{Z}} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{\Delta'_t})$ ▷ Noise vector of length of pipeline Δ'_t
7: $\hat{\boldsymbol{\theta}}_j^{(\cdot)} \leftarrow \boldsymbol{\theta}_t + \hat{\boldsymbol{\sigma}}\hat{\mathbf{Z}}$ ▷ Simulate a posterior mean $\boldsymbol{\theta}_{t+\Delta}$ when pipeline clears
8: $\hat{\mathbf{X}}_{1:\eta^{\text{off}}} \overset{i.i.d.}{\sim} F_{\tilde{x}}$ ▷ Sample η^{off} post-trial covariates
9: **for all** $w \in \mathscr{W}$ **do** ▷ For each treatment that could be given to patient type $\mathbf{X}_{t+1}$...
10: **for all** $i = 1, 2, \ldots, \eta^{\text{off}}$ **do** ▷ For each post-trial patient...
11: $\mathbf{a} \leftarrow ((1, 2, \ldots, n) \otimes \hat{\mathbf{X}}_i)\hat{\boldsymbol{\theta}}_j^{(\cdot)}$ ▷ vector of means for n treatments, type $\hat{\mathbf{X}}_i$ patients
12: $\mathbf{b} \leftarrow ((1, 2, \ldots, n) \otimes \hat{\mathbf{X}}_i)\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, w, \mathbf{X}_{t+1})$ ▷ Prepost std dev (EC.5) if patient $t+1$ has covariates $\mathbf{X}_{t+1}$ with treatment $W_{t+1} = w$
13: $\log(\nu_{j,w,i}) \leftarrow \log(h(\mathbf{a}, \mathbf{b}))$ ▷ conditional EVI, where $\log(h(\mathbf{a}, \mathbf{b}))$ is the ν ...
14: **end for** ▷ ... for cKG computed in Algorithm 2 in Frazier et al. (2009).
15: $\log(\hat{\nu}_{j,w}) \leftarrow \log((1/\eta^{\text{off}}) \sum_{i=1}^{\eta^{\text{off}}} \nu_{j,w,i})$ ▷ Estimate conditional EVI for w, given $\hat{\boldsymbol{\theta}}_j^{(\cdot)}$
16: **end for**
17: **end for**
18: **for all** $w \in \mathscr{W}$ **do**
19: $\log(\hat{\nu}_w) \leftarrow \log((1/\eta^{\text{on}}) \sum_{j=1}^{\eta^{\text{on}}} \hat{\nu}_{j,w})$ ▷ Est. index for w, ave. over parameter uncertainty
20: **end for**
21: **return** $W_{t+1} = \arg\max_{w \in \mathscr{W}} \log(\hat{\nu}_w)$ ▷ Pick largest estimated index (break ties randomly)
22: **end function**

Bayesian results for the preposterior distribution discussed shortly. For (ii), we simulate η^{off} post-trial patients in step 8. We use common random numbers (CRN) across treatment choices for patient $t+1$ for the simulation of (i) and (ii) to reduce the noise when selecting the largest index. For (iii), we observe that for a given posterior distribution at time $t+\Delta$ and individual post-trial patient with covariates $\hat{\mathbf{X}}_i$, the *conditional* EVI of computing the reward for a patient with covariates $\mathbf{X}_{t+1}$ can be computed using a subroutine of the correlated knowledge gradient (cKG, Frazier et al. 2009). This is computed in steps 11-13 and leads to an important further variance reduction.

Appendix EC.1 derives the so-called preposterior standard deviation used in fEVI-MC, $\tilde{\boldsymbol{\sigma}}(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})$, a $|\Xi| \times ((t_2 - t_1) - (\Delta - \Delta'_{t_1}))$ matrix such that $\tilde{\boldsymbol{\sigma}}(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\tilde{\boldsymbol{\sigma}}^\top(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})$ is the covariance of the mean $\boldsymbol{\theta}_{t_2}$, given Σ_{t_1}, and the outcomes of pipeline patients with indices $t_1 - \Delta'_{t_1} + 1, t_1 - \Delta'_{t_1} + 2, \ldots, t_2 - \Delta$ to be observed by time t_2. Here, $\Delta'_{t_1} = \min\{t_1, \Delta\}$. Step 3 uses $t_1 = t$ and $t_2 = t + \Delta$. Step 12 uses $t_1 = t + \Delta$ and $t_2 = t + \Delta + 1$.

Randomized fEVI *allocation policies:* fEVI-rand *and* fEVI-MC-rand. The fEVI allocation policy is deterministic. In clinical trials, randomization in the allocation of treatments helps mitigate potential biases (Piantadosi 1997). Although fEVI-MC allows for some randomization, the degree of randomness in fEVI-MC is not under explicit control of the trial manager. To accommodate greater control over randomization, allocation policy fEVI-rand(ϵ) samples a treatment uniformly at random with probability $\epsilon > 0$, and chooses the treatment with the largest fEVI-index with probability $1-\epsilon$ (an approach followed by Cheng and Berry 2007, Lai et al. 2013, Williamson et al. 2017). When the fEVI-indices are estimated using the fEVI-MC algorithm, we call the policy fEVI-MC-rand(ϵ).

5. Theoretical Properties of Expected Value of Information Heuristics

This section discusses the theoretical properties of the expected value of information heuristics of Section 4. EVI heuristics are designed to be one-step look-ahead optimal, at least when computed exactly. While they are myopic, EVI heuristics may also be optimal as the sample size goes to infinity. These properties, in the small and large sample size regimes, were already proved for settings where there are no covariates (Frazier et al. 2009, Chick et al. 2021) or where covariates are selected by the allocation policy (Ding et al. 2021). Here, we extend these properties for the setting where the covariates are random arrivals, satisfying Assumptions 1, 2 and 3 and for a positive fixed delay in observing outcomes. Proofs of claims in this section are found in Appendix EC.3.

If the sample size is $T=1$, the fEVI allocation policy is optimal for (11).

PROP. 2. *If* $T=1$, *then* $V^{f\text{EVI}}(\mathbf{K}_0) = V^*(\mathbf{K}_0)$.

The fEVI-MC, fEVI-rand, and fEVI-MC-rand allocation policies are not necessarily optimal for $T=1$, due to their reliance on Monte Carlo samples, randomization, or both.

For asymptotically large T, each of fEVI, fEVI-rand, and fEVI-MC-rand can be shown to be asymptotically optimal. To show this, we first show an upper bound on the value function and then show that the value approaches this upper bound as the sample size grows unbounded. Because the following results characterize V^π as a function of the sample size, we use $V^\pi(\mathbf{K}_0;T)$ to make the dependence on $\mathbf{K}_0$ and T explicit.

Let $U(\boldsymbol{\theta}_0,\Sigma_0) := \mathbb{E}[\max_{w\in\mathscr{W}}(w\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0,\Sigma_0]$ be the expected rewards obtained by an *oracle* that knows the parameters exactly ex ante. We first provide a characterization of the estimated parameters and the value function:

THEOREM 1. *For a given allocation policy π, there exists a random vector $\boldsymbol{\theta}_\infty^\pi$ and a random matrix Σ_∞^π such that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_\infty^\pi$ almost surely and $\Sigma_t \to \Sigma_\infty^\pi$ almost surely. Moreover, $V^\pi(\mathbf{K}_0; T)$ is bounded above by $U(\boldsymbol{\theta}_0, \Sigma_0)$, and $V^\pi(\mathbf{K}_0; \infty) := \lim_{T\to\infty} V^\pi(\mathbf{K}_0; T)$ exists.*

Given the upper bound on the trial value, we define an asymptotically optimal allocation policy to be one that achieves the upper bound as the sample size goes to infinity:

DEFINITION 1. If π is such that $V^\pi(\mathbf{K}_0; \infty) = U(\boldsymbol{\theta}_0, \Sigma_0)$, then we say that π is an *asymptotically optimal allocation policy.*

Theorem 2 states that allocation policies that have a positive probability of sampling the treatment with the largest expected value of information are asymptotically optimal.

THEOREM 2. *Let allocation policy π be such that, for $0 \le t \le T-1$, if $\nu_t(\mathbf{x}, w) \ge \nu_t(\mathbf{x}, v)\ \forall v \in \mathscr{W}$, then there exists a $\delta > 0$ so that $\mathbb{P}^\pi(W_{t+1} = w \mid \mathbf{K}_t, \mathbf{X}_{t+1} = \mathbf{x}) \ge \delta$. Then, π is asymptotically optimal.*

The proof of Theorem 2 shows that policies are asymptotically optimal if they always sample, with probability bounded away from zero, covariate-treatment combinations with positive value of information. Previous proofs of asymptotic optimality in settings without covariates (as in Frazier et al. 2009, Xie et al. 2016), show that policies are asymptotically optimal if they sample all treatments infinitely often (excluding treatments fully correlated with other treatments). The proof of Ding et al. (2021), which shows asymptotic optimality in the presence of covariates when the covariates are a choice of the allocation policy, also relies on showing that each treatment is sampled infinitely often.[1] An alternative approach is necessary in the presence of randomly arriving covariates because sampling each alternative infinitely often is not enough to achieve asymptotic optimality.[2] Using Theorem 2, we show that fEVI, fEVI-rand, and fEVI-MC-rand are asymptotically optimal.

THEOREM 3. *Allocation policies f*EVI*, f*EVI-rand*, and f*EVI-MC-rand *are asymptotically optimal.*

In the analysis for the proof of Theorem 1 (Corollary EC.3 in Appendix EC.3.3.1) we show that $V^\pi(\mathbf{K}_0; \infty) = \mathbb{E}^\pi[(\tilde{f}_{\boldsymbol{\theta}_\infty^\pi}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_0]$. The definitions of U and $\tilde{f}_{\boldsymbol{\mu}}$ imply that

[1] Compare Lemma EC.3 in the Appendix to Lemma A.7. of Frazier et al. (2009), to Lemma 7 of Xie et al. (2016), and to Proposition 3 of Ding et al. (2021).

[2] Recall Example 1. If all type A patients get treatment 1 and all type B patients get treatment 2, then both treatments are sampled infinitely often but one cannot learn the mean outcome for the two combinations never sampled (A-2,B-1).

$U(\boldsymbol{\theta}_0, \Sigma_0) = \mathbb{E}[(\tilde{f}_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0, \Sigma_0]$. For an asymptotically optimal policy π we thus have

$$\mathbb{E}^{\pi}[(\tilde{f}_{\boldsymbol{\theta}^{\pi}_{\infty}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_0] = \mathbb{E}[(\tilde{f}_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_0].$$

By the definition of $\tilde{f}_{\boldsymbol{\mu}}$, we know that $(\tilde{f}_{\boldsymbol{\theta}^{\pi}_{\infty}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \leq (\tilde{f}_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$, which implies

$$(\tilde{f}_{\boldsymbol{\theta}^{\pi}_{\infty}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} = (\tilde{f}_{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \text{ a.s.}$$

Thus, asymptotically optimal allocation policies (including, by Theorem 3, fEVI, fEVI-rand, and fEVI-MC-rand) result in convergence to an optimal treatment strategy. Although a uniform random allocation of treatments for each patient is also asymptotically optimal, numerical results in Section 7 suggest it is not as effective with small sample sizes.

We do not prove fEVI-MC to be asymptotically optimal using the same approach. However, note that $\hat{\nu}_w$ (in step 19 of Algorithm 1) is a consistent estimator of $\nu_t(\mathbf{X}_{t+1}, w)$, so that fEVI-MC$(\eta^{\text{on}} = \infty, \eta^{\text{off}} = \infty) = f$EVI. Moreover, fEVI-MC-indices are averages of modified cKG-indices, conditional on the outcomes of pipeline patients and post-trial patient covariates, and cKG indices are asymptotically optimal (Frazier et al. 2009) for subproblems that condition on those values. This explains the usefulness of fEVI-MC-indices in numerical examples (Section 7).

6. Modeler's Prior Distribution and Distribution of Problem Instances

The modeler's beliefs about the distribution of problem instances is assumed to be specified by the conjugate normal prior distribution $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0)$ in Assumption 3. It is common to use a Gaussian process model to specify realizations of means (Frazier et al. 2009).

We distinguish between a modeler's prior distribution of problem instances, specified by $\boldsymbol{\theta}_0$ and Σ_0, and the distribution of problem instances that nature presents to the modeler. For the latter, we allow nature to generate problem instances $\boldsymbol{\mu}$ with a $\mathcal{N}(\boldsymbol{\theta}_0^{nat}, \Sigma_0^{nat})$ distribution, which we refer to as *nature's distribution*.

In general, a modeler may wish to elicit a probability distribution for problem instances using known elicitation techniques (O'Hagan et al. 2006). If the elicitation is done well and Assumption 3 holds, then one expects that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{nat}$ and $\Sigma_0 = \Sigma_0^{nat}$. Alternatively, one might use empirical Bayes methods and use some pilot data together with a non-informative distribution to determine $\boldsymbol{\theta}_0, \Sigma_0$ (Gelman et al. 2013). To reduce the dimensionality of the

prior covariance matrix, we assume a Gaussian kernel (Frazier et al. 2009), though other kernels may be used (Rasmussen and Williams 2006, Chen et al. 2013):

$$\Sigma_{0,\ell_1,\ell_2} = \begin{cases} \sigma_0^2 & \text{if } \ell_1 = \ell_2 \text{ (same treatment and covariate)} \\ 0 & \text{if } \kappa_x(\ell_1) \neq \kappa_x(\ell_2) \text{ (different covariate)} \\ \sigma_0^2 e^{-\psi d(\kappa_w(\ell_1),\kappa_w(\ell_2))} & \text{otherwise (same covariate and different treatment),} \end{cases} \tag{13}$$

where σ_0^2 is the prior variance for the unknown value of parameters, $\kappa_x(\ell)$ is the covariate index associated with the regression coefficient indexed by $\ell = 1, \ldots, |\Xi|$ (0 for intercept or treatment effect terms), $\kappa_w(\ell)$ is the treatment associated with the regression coefficient indexed by $\ell = 1, \ldots, |\Xi|$ (0 for intercept or prognostic terms), and d is a distance function between treatments that is smaller for treatments that act with similar underlying mechanisms and larger if the treatments act with different mechanisms. If either argument of d is 0, corresponding to intercept or prognostic terms, we set d to be infinity (no correlation). Larger values of $\psi > 0$ model a lower correlation between similar treatments.

It may be useful to allow the modeler's prior distribution to differ from nature's distribution. For example, we may wish to assess the performance of allocation policies when nature chooses a specific problem structure (e.g., a *slippage configuration* where the "best" treatment is precisely δ units better than any alternative treatment, independent of the covariates, Shen et al. 2021). We call this a *fixed instance* setting because we let $\boldsymbol{\theta}_0^{nat}$ denote a fixed value of the true means $\boldsymbol{\mu}$ (and let Σ_0^{nat} be the zero matrix). We may also wish to allow that problem instances be randomly sampled to assess the average performance of an allocation policy (Branke et al. 2007). We call this a *random instance* setting because we allow nature to generate random draws of $\boldsymbol{\mu}$ from nature's distribution $\mathcal{N}(\boldsymbol{\theta}_0^{nat}, \Sigma_0^{nat})$.

Related work for optimal sequential learning in optimization applications without covariates suggests that it may be useful to consider manipulations of the modeler's prior distribution (Powell and Ryzhov 2012, Chick et al. 2021). With this in mind, we define the *robust prior distribution* as follows:

- for coefficients with index ℓ associated with prognostic or intercept terms, we set $\theta_{0,\ell}^{rob} = 0$, to push for evidence collection to confirm that the coefficients are non-zero;
- for coefficients with index ℓ associated with predictive or treatment effect terms, we set $\theta_{0,\ell}^{rob}$ artificially high, to push sampling to confirm whether such effects are active or not, with a "fudge" factor z_α (2 unless otherwise specified) times a standard error:

$$\theta_{0,\ell}^{rob} = \max_{\ell'}(\theta_{0,\ell'}) + z_\alpha \cdot \max\left(1, \sqrt{\max(\text{diag}(\Sigma_0))}\right); \tag{14}$$

- we set $\Sigma_0^{rob} = c \cdot \Sigma_0$, for some $c \geq 1$, to reduce the effect of the prior distribution on the inference and push for further exploration. We let $c = 4$ unless otherwise specified.

7. Empirical Performance of Expected Value of Information Heuristics

The analytical results in Section 5 provide optimality guarantees for asymptotic sample sizes or a sample size of 1. We now present numerical results on the learning efficiency of the EVI heuristics for intermediate sample sizes, relative to adaptations of techniques from the BATTLE clinical trial (Zhou et al. 2008), the biased coin approach (Pocock and Simon 1975), the Thompson sampling algorithm for multi-armed bandits (Thompson 1933), and an adaptation of Thompson sampling (Russo 2020).

Section 7.1 describes a motivating example from the field of sepsis management for the experimental setup, performance metrics, and comparator allocation policies. Section 7.2 studies the efficiency of the EVI-based allocation policies and the comparator algorithms under the assumption that the labeling of the underlying model is known to the trial manager. Section 7.3 explores the penalty incurred by a trial manager who does not know the correct labeling and mislabels the covariates, such as labeling a predictive covariate as prognostic or idle, or vice versa. Section 7.4 explores how randomization (fEVI-MC and fEVI-rand) and discretization of covariates affect the efficiency of learning. Section 7.5 explores the role of the delay on the rate of inference.

7.1. Experimental setup, performance metrics, and comparator allocation policies

We use a configuration of model coefficients motivated by work on sepsis treatment to illustrate our proposed model and allocation policies. We do not intend to provide medical recommendations here.

7.1.1. Motivating example: Sepsis treatment. Sepsis can be defined a life-threatening organ dysfunction caused by dysregulated host response to infection, and has been operationalized with an acute increase in the sequential (sepsis-related) organ failure assessment (SOFA) score of 2 points or more (Singer et al. 2016). Higher SOFA scores reflect greater organ dysfunction. Consider a clinical trial where the primary outcome is the change in SOFA score (Lambden et al. 2019) over a seven day period following the diagnosis of sepsis. We assume initially that $\Delta = 0$ (the patient inter-arrival time is larger than the delay in observing outcomes) and approximate the change in SOFA score as a continuous quantity. We observe the following patient covariates:

1. A unique type among four (the Mars1, Mars2, Mars3, and Mars4 endotypes of Scicluna et al. 2017) based on blood transcriptomic data. The four types are coded as three binary covariates. Mars4 is the baseline. Indicators for the Mars1, Mars2, and Mars3 endotypes are provided by covariates with indices 1, 2, and 3, respectively.
2. A real-valued covariate representing disease severity (with index 4), motivated by the APACHE score (Zimmerman et al. 2006) for critically ill patients.
3. A real-valued covariate that is idle (with index 5). In the context of sepsis, many covariates are idle; we use one here for illustration purposes.

Recent work on sepsis (van Mourik et al. 2022) explores three aspects of treatment that may interact with one or more Mars types. We consider these aspects of treatment: fluid management (restrictive or liberal), tighter glucose control (no or yes), and hydrocortisone dose (low or high). This gives $n = 2^3 = 8$ treatment alternatives.[3]

As an example of a fixed instance, we assume that type Mars3 is predictive with respect to treatments 5-8, that is, it interacts with one aspect of treatment (fluid management), and in particular that treatment 5 has an exceptionally good response with Mars3 patients ($\mu_{5,3} = 1$, $\mu_{6,3} = \mu_{7,3} = \mu_{8,3} = 0.5$). The other types all respond equally to each treatment and respond better to treatment 4 ($\mu_{4,0} = 0.5$). All types are prognostic, with type Mars3 having better prognosis, consistent with Scicluna et al. (2017) ($\mu_{0,1} = \mu_{0,2} = -1$, $\mu_{0,3} = 1$), and the intercept term is non-zero ($\mu_{0,0} = 1$). We assume that the APACHE score is prognostic (with $\mu_{0,4} = -3$, assuming the APACHE score is re-scaled to a [0,1] range) and is not predictive with respect to any treatment. The last covariate is assumed idle. All coefficients which are not specified above are assumed to be zero. We assume that $\sigma^2 = 1$.

As an example of a random instance structure, we assume the same labeling as the fixed instance, i.e., only the non-zero coefficients of the fixed instance are potentially active in the random instance. We further assume all active parameters have a zero mean, i.e. $\boldsymbol{\theta}_0^{nat} = \mathbf{0}$, and the standard deviation is two, i.e., the diagonal elements of Σ_0^{nat} are all the same and equal to four. Because $\mu_{5,3}$, $\mu_{6,3}$, $\mu_{7,3}$, and $\mu_{8,3}$ are the predictive coefficients of the same covariate for different treatments, we assume a positive correlation between these coefficients, with

$$\mathrm{Cov}(\mu_{5,3}, \mu_{6,3}) = \mathrm{Cov}(\mu_{5,3}, \mu_{7,3}) = \mathrm{Cov}(\mu_{6,3}, \mu_{8,3}) = \mathrm{Cov}(\mu_{7,3}, \mu_{8,3}) = 1$$

[3] We index them as 1: fluid restr., gluc. no, hydrocort. low; 2: fluid restr., gluc. no, hydrocort. high; 3: fluid restr., gluc. yes, hydrocort. low; 4: fluid restr., gluc. yes, hydrocort. high; 5: fluid lib., gluc. no, hydrocort. low; 6: fluid lib., gluc. no, hydrocort. high; 7: fluid lib., gluc. yes, hydrocort. low; 8: fluid lib., gluc. yes, hydrocort. high.

(correlation= 1/4) because these pairs share two aspects of treatment. The remaining combinations of these four coefficients have a covariance of one half (correlation= 1/8) because they share one aspect of treatment. This corresponds to the structure of (13) with $\sigma_0^2 = 4$, $\psi = \log(4)$, and $d(5,6) = d(5,7) = d(6,8) = d(7,8) = 1$ and $d(5,8) = d(6,7) = 2$.

We initially assume that the modeler correctly elicited nature's distribution ($\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{nat}$ and $\Sigma_0 = \Sigma_0^{nat}$) of the random instance and uses the robust prior defined in Section 6 with $z_\alpha = 2$ and $c = 4$. We assume the same robust prior for the fixed instance setting, too.

Because we can efficiently compute the fEVI allocation policy when $\Delta = 0$ and the covariates are discrete and finite (using the method in Alban et al. 2021), we first assume that the support of the real-valued covariates is three possible values, corresponding to a low, medium, and high level. Also, for purposes of illustration, we assume that the APACHE and idle covariates are sampled i.i.d. from the discrete distribution that takes the value 0 (low) with probability 1/4, 0.5 (medium) with probability 1/2, and 1 (high) with probability 1/4. An experiment in Section 7.4 assesses the effect of this discretization. The probabilities for each of the types Mars1, Mars2, Mars3, and Mars4 are, respectively, proportional to 150, 184, 129, and 59 (these are counts for the respective types as reported by Scicluna et al. 2017). We assume that the covariates are independent of each other and are equally distributed for patients enrolled in the trial and post-trial.

7.1.2. Performance. We measure the ability of allocation policies to learn the best treatment strategy with two performance metrics. The *Expected Opportunity Cost* (EOC) is the expected difference between the value obtained by an oracle and by an allocation policy π for a specific instance of the coefficients $\boldsymbol{\mu}$:

$$\mathrm{EOC}^{\pi}(\boldsymbol{\mu}; T) := \mathbb{E}^{\pi}\left[\max_{w \in \mathscr{W}} (w \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\mu}\right]. \tag{15}$$

The *Probability of Incorrect Selection* (PICS) is the probability that a post-trial patient does not get a treatment that gives the best mean outcome:

$$\mathrm{PICS}^{\pi}(\boldsymbol{\mu}; T) := \mathbb{P}^{\pi}\left[\max_{w \in \mathscr{W}} (w \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} > (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\mu}\right]. \tag{16}$$

These performance metrics average over trial outcomes through π and over randomly selected post-trial patients $\tilde{\mathbf{X}}_1$ after the treatment strategy $\tilde{f}$ is adopted. Their values are conditional on the true means, $\boldsymbol{\mu}$, of a given problem instance. For a random instance setting, we average realizations of random problem instances, e.g., $\mathrm{EOC}^{\pi}(T) = \mathbb{E}[\mathrm{EOC}^{\pi}(\boldsymbol{\mu}; T) \mid \boldsymbol{\theta}_0^{nat}, \Sigma_0^{nat}]$. Better allocation policies obtain lower values of EOC and PICS.

7.1.3. Comparators. We will compare the EVI-based allocation policies against several competitor allocation policies. The *Thompson Sampling* (TS) allocation policy draws a sample of $\boldsymbol{\mu}$ from the posterior distribution and assigns the best treatment for the patient given the sampled $\boldsymbol{\mu}$ (Thompson 1933). *Top-Two Thompson Sampling* (TTTS) is an adaptation of Thompson sampling: with probability β, the policy samples the treatment assigned by the TS allocation policy, and with probability $1-\beta$, the policy resamples $\boldsymbol{\mu}$ from the posterior distribution until a sample is drawn such that the best treatment differs from that recommended by TS (Russo 2020). We use $\beta = 0.5$ here. The *BATTLE** policy is derived from the allocation policy used in the BATTLE trial (Zhou et al. 2008) for 0-1 outcomes, but adapted to our setting. It samples each treatment with probability proportional to its expected outcome. The *Biased Coin* (BC*) allocation policy balances the values of prognostic covariates across arms for a given set of predictive covariate values, in the spirit of Pocock and Simon (1975). Appendix EC.4 gives further details. The *Random* policy assigns treatments uniformly at random, regardless of covariates.

7.2. Performance of fEVI allocation policy

We compare the performance of the allocation policies when the labeling is known to assess how the policies perform with finite sample size. We aim to demonstrate the performance of fEVI compared to alternative policies in the literature. Figure 2 shows the EOC against the sample size for the fEVI policy and the comparator policies defined in Section 7.1.3. Lower curves correspond to higher learning efficiency. In a challenging fixed instance setting (Figure 2a), we observe that the fEVI, TS, and TTTS allocation policies obtain the lowest EOC, with no significant difference. The BATTLE* allocation policy obtains the highest EOC in the fixed instance setting, performing worse than the Random and BC* allocation policies; these three are significantly worse than the rest.

Average performance is also of interest. In the random instance setting (Figure 2b), the fEVI policy obtains the lowest EOC. TTTS is almost as good, statistically, for sample sizes above 280 in this experiment ($\text{EOC}^{f\text{EVI}}(400) = 0.0032 \pm 0.0002$, $\text{EOC}^{\text{TTTS}}(400) = 0.0038 \pm 0.0003$[4]). TS is the next best ($\text{EOC}^{\text{TS}}(400) = 0.0052 \pm 0.0003$) but is statistically significantly worse than fEVI. BATTLE*, BC*, and Random perform worse than those three allocation policies. BATTLE* performs better than Random in the random instance

[4] We report average $\pm$ standard error of simulation results.

setting, suggesting that the fixed setting we selected for our experiments is more difficult to learn for this policy. Random and BC* have the highest EOC, at an average of $\text{EOC}^{BC*} \approx \text{EOC}^{\text{Random}}(400) = 0.0076$. To obtain the same level of average EOC, BATTLE* requires 361 samples, TS 283, TTTS 207, and fEVI 172.

In the remainder of this section we focus on the random instance setting.

7.3. Mislabeling predictive, prognostic, and idle covariates

We assumed previously that the trial manager knew the true labeling of the covariates. This is valid in some clinical settings but may not be realistic in other settings. In this subsection, we explore how the trial manager's lack of knowledge about the labeling of the covariates may impact the inference made through the trial. We present results on some practically relevant mislabelings and summarize additional results reported in an appendix. Throughout, we still assume that nature's model is as described in Section 7.1.

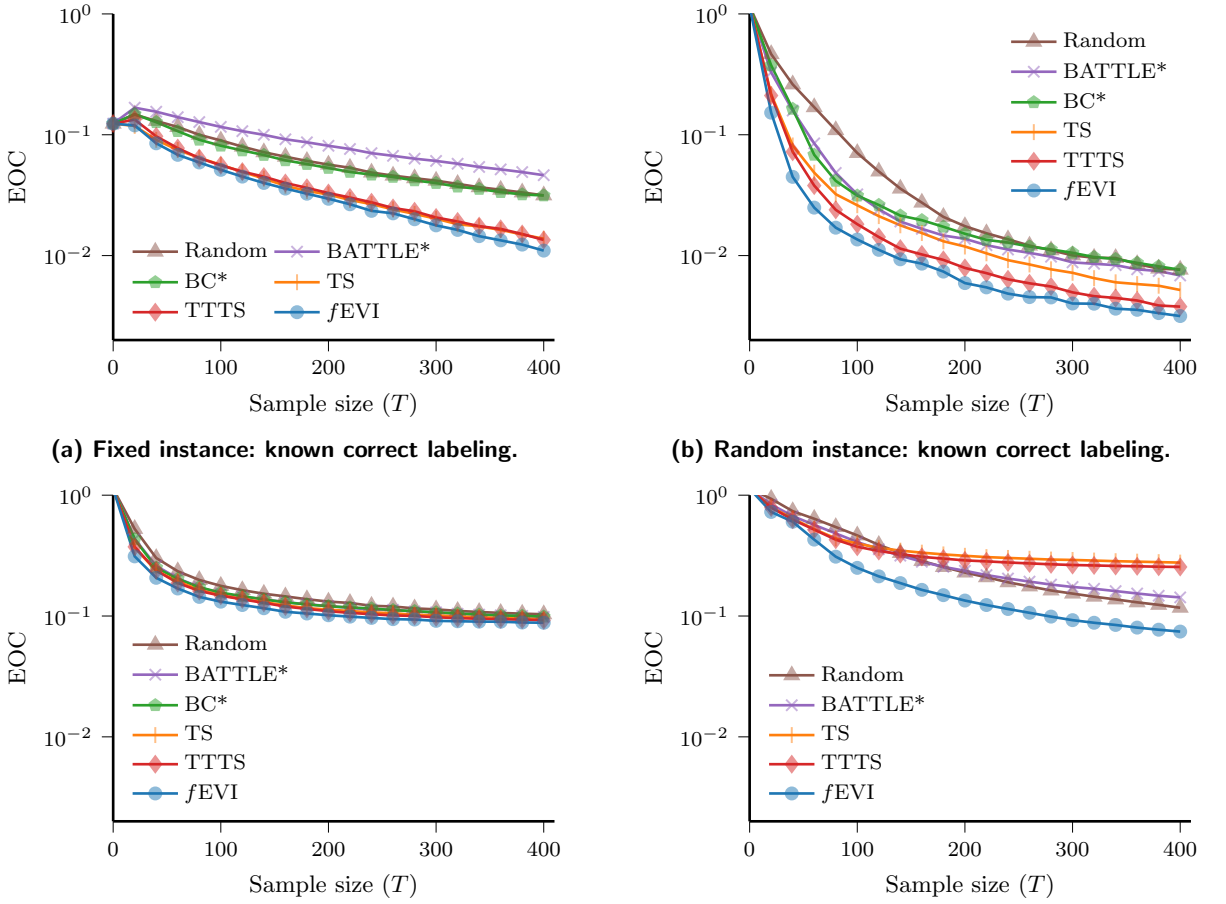
Unknown personalization potential. Consider that the trial manager is unaware that Mars3 is a predictive covariate and labels it as only prognostic. Figure 2c shows that this mislabeling prevents the potential benefit of precision medicine. The EOC of all policies approaches an asymptote: the best EOC that can be achieved without personalization.

Taking all covariates to be potentially predictive. Consider that the trial manager labels all Mars endotypes, APACHE, and the idle covariate as potentially predictive with respect to all treatments, with no prognostic covariates (48 instead of 11 coefficients). Labeling all covariates as predictive is a conservative model often adopted in the literature (e.g., Bastani and Bayati 2020). Figure 2d shows that this mislabeling slows inference. The EOC for any given allocation policy is higher compared to Figure 2b because more coefficients are estimated. Figure 2d does not show BC* because there are no prognostic covariates.

In both settings of Figures 2c and 2d, the relative performance of the different allocation policies is the same as for the known labeling setting in Section 7.2: fEVI does best.

Summary of other insights. Appendix EC.5 provides some additional numerical experiments exploring the effect of other types of mislabelings of covariates. We summarize some of those insights, all applicable to the specific scope of our model and example. (i) Knowing the right labeling is more important than employing a smart allocation policy: the random allocation policy assuming the correct labeling outperforms the adaptive allocation policies on highly erroneous labelings. (ii) The general insights of Sections 7.2–7.3 with respect to the EOC performance metric are largely in line with those for PICS in these experiments

Figure 2 Expected opportunity cost (EOC) for policies with different labelings.



(a) Fixed instance: known correct labeling.

(b) Random instance: known correct labeling.

(c) Random instance: unknown personalization potential.

(d) Random instance: all covariates potentially predictive.

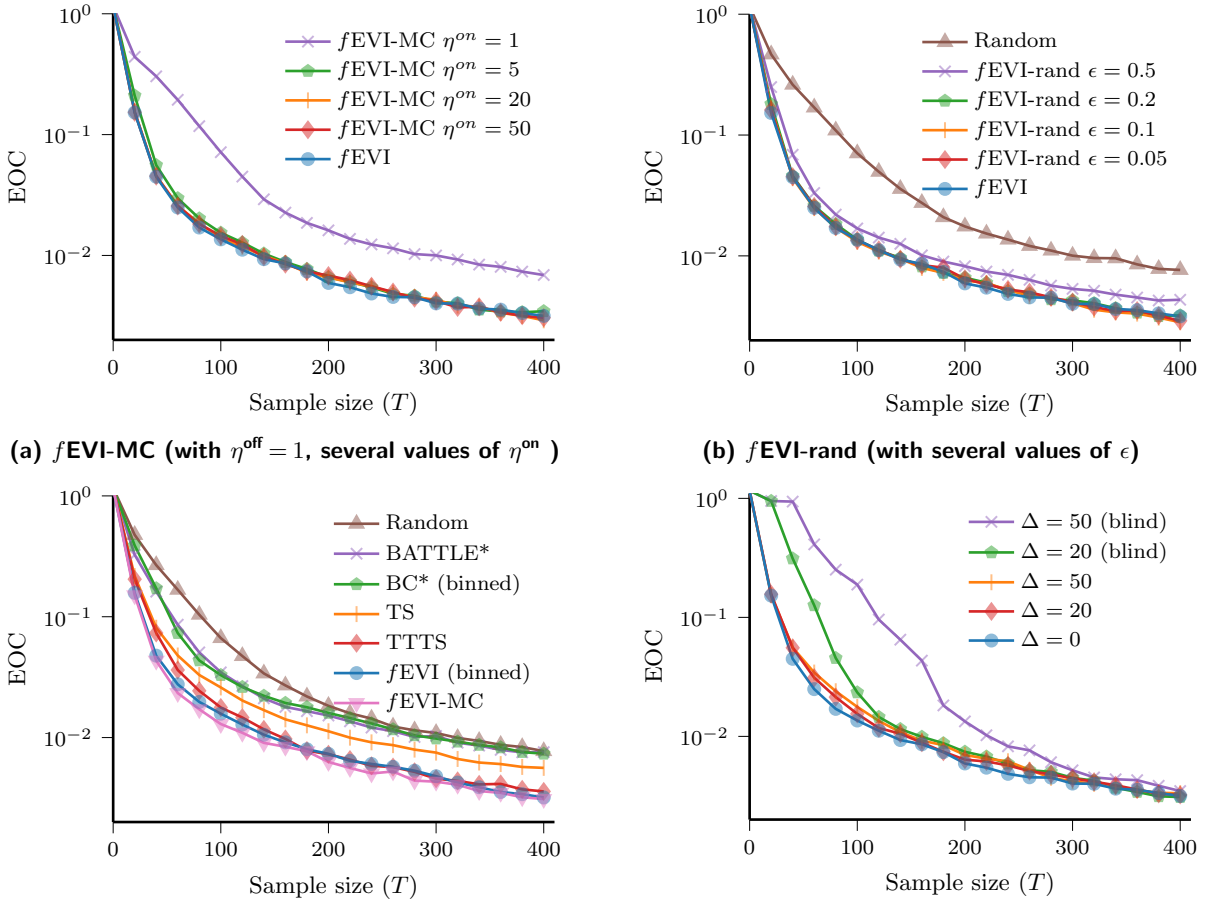
(Appendix EC.5.1). (iii) Comparisons across fixed instance to random instance settings for other labelings tested resemble those for Figures 2a –2b (data not shown).

7.4. Simulation to estimate fEVI or to randomize patients

We assess the performance of the allocation policies that use simulation or randomization, compared to that of fEVI, on the base case with random instances described in Section 7.1.

Firstly, the fEVI-MC allocation policy uses simulation to estimate fEVI indices. Figure 3a shows the EOC for the fEVI-MC allocation policy with several values of η^{on}. In this experiment, we fixed $\eta^{\text{off}} = 1$ because, when $\Delta = 0$, the fEVI-MC policy makes $\eta^{\text{on}}\eta^{\text{off}}$ i.i.d. simulations regardless of the specific values of η^{on} and η^{off}. The fEVI-MC policy performs very closely to fEVI even with a few replications (small $\eta^{\text{off}}\eta^{\text{on}}$). Importantly, the variance reduction in fEVI-MC (see Section 4.2) that computes an expected information gain for all potential outcomes for the extra patient, not just for one realized outcome, is effective.

Figure 3 Effect of practical considerations on the performance of fEVI. The plots show the EOC of the random instance setting.



(a) fEVI-MC (with $\eta^{\text{off}} = 1$, several values of η^{on})

(b) fEVI-rand (with several values of ϵ)

(c) EOC when APACHE score is continuous-valued. Allocation policies fEVI and BC* bin the APACHE score into three levels. fEVI-MC uses $\eta^{\text{on}} = 20$, $\eta^{\text{off}} = 1$.

(d) Effect of delay Δ on the EOC when using the fEVI-MC and fEVIblind allocation policies.

Appendix EC.5.3 shows that an implementation of Monte Carlo estimation without this variance reduction uses at least two orders of magnitude more replications.

Secondly, fEVI-rand provides additional control to randomize patients. Figure 3b compares the random instance EOC of the fEVI-rand policy for several values of ϵ to the EOC obtained by the Random policy (equivalent to fEVI-rand with $\epsilon = 1$) and fEVI (equivalent to fEVI-rand with $\epsilon \to 0$). The EOC increases with ϵ. For $\epsilon \leq 0.2$, we do not observe a practically significant difference compared to fEVI. For $\epsilon = 0.5$, we observe a loss in efficiency compared to fEVI but obtain a significant improvement compared to Random. We observe similar results for the fEVI-MC-rand allocation policy presented in Appendix EC.5.4. In line with observations of Williamson et al. (2017) for a related context, some randomization can be obtained without significantly degrading performance.

Lastly, to further assess the practical value of fEVI-MC, we compare which of the two approximations is better for handling continuous-valued covariates: binning the continuous covariates' values in buckets and employing the exact fEVI versus leaving the covariates continuous-valued and approximating the fEVI-index computation with fEVI-MC. Figure 3c considers the random instance setting when the APACHE score is a continuous-valued covariate, as described in Section 7.1. fEVI-MC estimates the fEVI-indices with $\eta^{\text{on}} = 20$ and $\eta^{\text{off}} = 1$, while fEVI estimates the indices by binning the APACHE score into three levels. The fEVI-MC allocation policy obtains the lowest EOC, although not practically significantly lower than that of fEVI. We also plot the comparator policies as a reference. We observe a performance very similar to that observed in Figure 2b. Thus, both approximation approaches are viable in this example.

7.5. Effect of delay and pipeline information on the EOC of the fEVI-MC policy

We quantify the effect of delay on the performance of fEVI-MC. We use fEVI-MC because it is easier to compute than fEVI when delays are not zero. We also explore the benefit of accounting for the allocations to pipeline patients before their outcomes are observed. We use an allocation policy that only uses information on patients in hand without using information about patients still in the pipeline and the treatments assigned to them. We call this latter allocation policy fEVIblind. Both allocation policies use information from all patients when making an implementation decision.

Figure 3d shows three performance curves for EOC with delay $\Delta = 0, 20, 50$ for fEVI-MC. It shows a very modest degradation in performance, but not strongly significant, for delays over the tested range. Figure 3d also shows two performance curves for fEVIblind with $\Delta = 20, 50$. Knowledge of the pipeline, even before seeing the outcome, is beneficial for arm allocation, particularly in the early stages of sampling (up to sample sizes of 4-5 times the Δ). For smaller trials, accounting for the pipeline as done in fEVI-MC is very important. Ignoring the pipeline has a lower effect if the sample size exceeds several multiples of Δ, after many samples have been collected.

8. Discussion

This paper presented modeling and algorithmic contributions to support the efficient learning of personalized treatment strategies through clinical trials designed to identify the best treatment as a function of patient covariates in the context of precision medicine.

We proposed allocation policies that maximize the expected value of information from one additional sample and that leverage knowledge of predictive and prognostic covariates to learn the best treatment strategy, a *function* from covariates to treatments. Several of our proposed fEVI-based allocation policies were shown to be asymptotically optimal in learning such treatment strategies. All proposed fEVI-based allocation policies performed well in intermediate-sample regimes in numerical experiments in comparison with a biased coin approach to balancing prognostic covariates in clinical trials and in comparison with a model that assumes all covariates are predictive with respect to all treatments, as is sometimes assumed in applications in and outside of medicine.

Our numerical experiments for adaptive contextual learning were motivated by sepsis management. In that example, the APACHE score is a known prognostic covariate, and preliminary findings in the literature suggest that certain blood transcriptomic endotypes may be predictive covariates with respect to certain sepsis treatments. Our numerical results suggest a promise for using our proposed fEVI family of allocation policies in the design of a clinical trial to validate the potential for precision medicine for sepsis on the basis of blood transcriptomic typing, in comparison with other allocation policies tested here (adaptations of the BATTLE trial, the biased coin, and Thompson sampling).

Thus, our main hypothesis is supported on theoretical and computational grounds: there is a benefit to adaptive contextual learning with predictive and prognostic covariates (if their identity is known), particularly with our proposed fEVI approach. The variance reductions embedded in fEVI-MC apply to contextual learning applications beyond clinical trials when the predictive distribution for posterior mean, for when the pending observations have been observed, is available. There are several related issues to consider towards extending the base model to allow for a broader set of clinical trials and other applications. Appendix EC.6 discusses a number of them, showing how some can be incorporated into our framework, while some can provide areas for further research.

Acknowledgement. We acknowledge the European Union's support through the MSCA-ESA-ITN project (676129), and discussions with Drs. A.P.J. Vlaar, W.J. Wiersinga, F. Uhel, B. Scicluna, and N. van Mourik (Amsterdam University Medical Center). Chick's work was also supported by the Novartis Chair of Healthcare Management at INSEAD.

References

Alban A, Chick SE, Zoumpoulis SI (2021) Expected value of information methods for contextual ranking and selection: clinical trials and simulation optimization. Kim S, Feng B, Smith K, Masoud S, Zheng Z, Szabo C, Loper M, eds., *Proc. 2021 Winter Simulation Conference*, to appear (IEEE, Inc.).

Anderer A, Bastani H, Silberholz J (2022) Adaptive clinical trial designs with surrogates: When should we bother? *Management Science* 68(3):1982–2002.

Astudillo R, Jiang DR, Balandat M, Frazier PI, Bakshy E (2021) Multi-step budgeted Bayesian optimization with unknown evaluation costs. *Advances in Neural Information Processing Systems*.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning* 3:397–422.

Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L (2009) I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharm. & Therapeutics* 86(1):97–100.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates online decision making with high-dimensional covariates. *Operations Research* 68(1):276–294.

Bastani H, Bayati M, Khosravi K (2021) Mostly exploration-free algorithms for contextual bandits. *Management Science* 67(3):1329–1349.

Berry DA (2011) Adaptive clinical trials in oncology. *Nat Rev Clinical Oncology* 9(4):199–207.

Bertsekas DP, Shreve SE (1978) *Stochastic optimal control: The discrete time case* (Belmont, Massachusetts: Athena Scientific).

Bertsimas D, Korolko N, Weinstein AM (2019a) Covariate-adaptive optimization in online clinical trials. *Operations Research* 67(4):1150–1161.

Bertsimas D, Korolko N, Weinstein AM (2019b) Identifying exceptional responders in randomized trials: An optimization approach. *INFORMS Journal on Optimization* 1(3):187–199.

Bhat N, Farias VF, Moallemi CC, Sinha D (2020) Near-optimal A-B testing. *Management Science* 66(10):4477–4495.

Branke J, Chick S, Schmidt C (2007) Selecting a selection procedure. *Management Science* 53(12):1916–1932.

Carranza AG, Krishnamurthy SK, Athey S (2022) Flexible and efficient contextual bandits with heterogeneous treatment effect oracle. Available at https://arxiv.org/abs/2203.16668.

Chen X, Ankenman BE, Nelson BL (2013) Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research* 61(2):512–528.

Cheng Y, Berry DA (2007) Optimal adaptive randomized designs for clinical trials. *Biometrika* 94(3):673–689.

Chick SE (2006) Subjective probability and Bayesian methodology. Henderson S, Nelson B, eds., *Handbooks in Operations Research and Management Science: Simulation*, chapter 9 (Elsevier).

Chick SE, Gans N, Yapar O (2021) Bayesian sequential learning for clinical trials of multiple correlated medical interventions. *Management Science* accepted to appear:doi.org/10.1287/mnsc.2021.4137.

Chick SE, Inoue K (2001) New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49(5):732–743.

Cutter GR, Liu Y (2012) Personalized medicine: The return of the house call? *Neurology Clinical Practice* 2(4):343–351.

Ding L, Hong LJ, Shen H, Zhang X (2021) Technical note – Knowledge gradient for selection with covariates: Consistency and computation. *Naval Research Logistics* 69(3):496–507.

Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30(24):2867–2880.

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Frazier PI, Powell WB, Dayanik S (2009) The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* 21(4):599–613.

Gao S, Du J, Chen CH (2019) Selecting the optimal system design under covariates. *IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 547–552 (IEEE).

Gelman A, Carlin JB, Stern HS, et al. (2013) *Bayesian data analysis* (CRC press).

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Jacko P (2018) Mitigating the curse of dimensionality of the Bayesian Beta-Bernoulli bandit problem. *Workshop on Multi-Armed Bandits and Learning Algorithms* (Rotterdam, Netherlands), accessed May 7, 2020, https://www.lancaster.ac.uk/staff/jacko/goal/Jacko2018_05_MAB_slides.pdf.

Kim SH, Nelson BL (2006) Selecting the best system. Henderson S, Nelson B, eds., *Handbooks in Operations Research and Management Science*, chapter 17 (Elsevier).

Lai TL, Liao OYW, Kim DW (2013) Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemporary Clinical Trials* 36:651–663.

Lambden S, Laterre PF, Levy MM, Francois B (2019) The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care* 23(374):1–9.

Lee MS, Flammer AJ, Lerman LO, Lerman A (2012) Personalized medicine in cardiovascular diseases. *Korean Circulation Journal* 42(9):583–591.

Li H, Lam H, Liang Z, Peng Y (2020) Context-dependent ranking and selection under a Bayesian framework. Bae KH, Feng B, Kim S, Lazarova-Molnar S, Zheng Z, Roeder T, Thiesing R, eds., *Proc. 2020 Winter Simulation Conference*, 2060–2070 (IEEE, Inc.).

Lipkovich I, Dmitrienko A, D'Agostino Sr RB (2017) Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine* 36(1):136–196.

Negoescu D, Frazier P, Powell W (2011) The knowledge gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing* 23(3):331–492.

NIH (2015) Personalized medicine. US National Institutes of Health, https://www.nih.gov/about-nih/what-we-do/nih-turning-discovery-into-health/personalized-medicine.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgements: Eliciting Experts' Probabilities* (John Wiley & Sons).

Oldenhuis C, Oosting S, Gietema J, De Vries E (2008) Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer* 44(7):946–953.

Opal S, Dellinger R, Vincent J, et al. (2014) The next generation of sepsis clinical trial designs: What is next after the demise of recombinant human activated protein C?*. *Crit Care Med* 2014:42.

Pallmann P, et al. (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine* 16(29), Accessed May 7, 2018, https://doi.org/10.1186/s12916-018-1017-7.

Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E (2018) Epidemiology and costs of sepsis in the united states-an analysis based on timing of diagnosis and severity level. *Crit Care Med* 46(12):1889–1897.

Pearce M, Branke J (2018) Continuous multi-task Bayesian optimisation with correlation. *European Journal of Operational Research* 270(3):1074–1085.

Piantadosi S (1997) *Clinical Trials: A Methodologic Perspective* (Wiley).

Pocock SJ, Simon R (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31:103–115.

Powell WB, Ryzhov IO (2012) *Optimal learning* (John Wiley & Sons).

Rasmussen CE, Williams CK (2006) *Gaussian Processes for Machine Learning* (MIT Press Cambridge).

Rello J, et al. (2018) Towards precision medicine in sepsis: a position paper from the European society of clinical microbiology and infectious diseases. *Clinical Microbiology and Infection* 24(12):1264–1272.

Rojas-Cordova A, Bish EK (2018) Optimal patient enrollment in sequential adaptive clinical trials with binary response. Available at SSRN, https://ssrn.com/abstract=3234590.

Russo D (2020) Simple Bayesian algorithms for best arm identification. *Operations Research* 68(6):1625–1931.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 68(4):1221–1243.

Ryzhov IO, Powell WB (2011) Information collection on a graph. *Operations Research* 59(1):188–201.

Schork NJ (2018) Randomized clinical trials and personalized medicine. *Social Science & Medicine* 210:71.

Scicluna BP, et al. (2017) Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *The Lancet Respiratory Medicine* 5(10):816–826.

Sechidis K, Papangelou K, Metcalfe PD, Svensson D, Weatherall J, Brown G (2018) Distinguishing prognostic and predictive biomarkers : an information theoretic approach. *Bioinformatics* 34(19):3365–3376.

Seymour CW, et al. (2019) Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 321(20):2003–2017, URL http://dx.doi.org/10.1001/jama.2019.5791.

Shen H, Hong LJ, Zhang X (2021) Ranking and selection with covariates for personalized decision making. *INFORMS Journal on Computing* 33(3):1500–1519.

Singer M, et al. (2016) The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 315(8):801–810, URL doi:10.1001/jama.2016.0287.

Symmans WF, et al. (2018) Residual cancer burden (RCB) as prognostic in the I-SPY 2 trial. *J. Clinical Oncology* 36(15_suppl):520–520, URL http://dx.doi.org/10.1200/JCO.2018.36.15_suppl.520.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294, URL http://www.jstor.org/stable/2332286.

Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R (2020) Review of precision cancer medicine: evolution of the treatment paradigm. *Cancer Treatment Reviews* 86:102019.

US FDA (2021) Adjusting for covariates in randomized clinical trials for drugs and biological products. US Food and Drug Administration, https://www.fda.gov/media/148910/download.

van Mourik N, MARS Consortium, et al. (2022) Blood transcriptomic endotypes and the response to treatment modalities in sepsis: A prospective cohort study. Abstract, submission, 2022 European Society of Intensive Care Medicine Conf.

Villar SS, Rosenberger WF (2018) Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics* 74(1):49–57.

Wang H, Yee D (2019) I-SPY 2: A neoadjuvant adaptive clinical trial designed to improve outcomes in high-risk breast cancer. *Curr Breast Cancer Rep.* 11(4):303–310.

Wang J, Clark SC, Liu E, Frazier PI (2020) Parallel Bayesian global optimization of expensive functions. *Operations Research* 68(6):1850—1865.

Wang Y, Wang C, Powell W (2016) The knowledge gradient for sequential decision making with stochastic binary feedbacks. *Proc. 33rd International Conference on Machine Learning*, PMLR 48:1138–1147.

WHO (2020) Global report on the epidemiology and burden of sepsis: Current evidence, identifying gaps and future directions. World Health Organization.

Williamson SF, Jacko P, Jaki T (2022) Generalisations of a Bayesian decision-theoretic randomisation procedure and the impact of delayed responses. *Computational Statistics and Data Analysis* 174:107407, ISSN 0167-9473, URL http://dx.doi.org/https://doi.org/10.1016/j.csda.2021.107407.

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

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

Wu AD, Zumbo BD (2008) Understanding and using mediators and moderators. *Social Indicators Research* 87(3):367–392.

Wu J, Frazier P (2016) The parallel knowledge gradient method for batch Bayesian optimization. *Advances in Neural Information Processing Systems*, 3126–3134.

Xie J, Frazier PI, Chick SE (2016) Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research* 64(2):542–559.

Xiong S (2020) Personalized optimization and its implementation in computer experiments. *IISE Transactions* 52(5):528–536.

Zhou X, Liu S, Kim ES, Herbst RS, Lee JJJ (2008) Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials* 5(3):181–193, ISSN 17407745.

Zimmerman JE, et al. (2006) Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical Care Medicine* 34(5):1297–1310.

E-Companion

Table EC.1 summarizes the principal notation of the model. Appendix EC.1 provides additional details regarding Bayes' rule used in the inference process as trial data accumulates. Appendix EC.2 justifies the claim of Proposition 1 that Bellman's equation gives an optimal policy. Appendix EC.3 proves the main results of Section 5, including the asymptotic optimality of our heuristic allocation policies, and gives a number of useful supporting results. Appendix EC.4 describes in detail the comparator alloca-

Table EC.1 Table of notation

	Parameters	Description
Time	$T \in \{0,1,2,\ldots\}$	Sample size (deterministic for main model)
	$\Delta \in \{0,1,2,\ldots\}$	Delay in observing the outcomes of patients (in number of patients enrolled)
	$t \in \{0,\ldots,T+\Delta\}$	Discrete time steps; for $t \le T$, the number of patients enrolled so far
	$\Delta'_t = \min\{\Delta, t\}$	Number of patients in pipeline at time $t \le T$
Treatments	w	Generic treatment
	$n \in \mathbb{N}_{>0}$	Number of available treatments
	$\mathscr{W} = \{1,\ldots,n\}$	Set of treatments
	W_t	Treatment assigned to patient at time $t = 1,2,\ldots,T$
Covariates	$\mathbf{x}$	Generic vector of covariates
	$m \in \mathbb{N}_{>0}$	Dimension of covariate vector
	$\mathscr{X} \subseteq \mathbb{R}^m$	Set of feasible covariate values
	$\mathbf{X}_t$	Vector of covariates of patient at time t
Outcomes	$\mathscr{Y} \subseteq \mathbb{R}$	Set of feasible outcomes
	Y_t	Outcome of patient arriving at time $t = 1,2,\ldots,T$
	$\mathbf{H}_t$	History of covariates, treatments, and observed outcomes until time $t = 0,1,\ldots,T+\Delta$
Response function	$r_{\boldsymbol{\mu}}(\mathbf{x}, w)$	Expected outcome given covariates $\mathbf{x}$, treatment w, and coefficients $\boldsymbol{\mu}$
	$\boldsymbol{\mu}$	Unknown coefficients
	$\boldsymbol{\xi}$	Coefficients that are potentially active (labeling)
	$\otimes$	Operator from Section 2.3 that identifies potentially active coefficients
Bayes inference	σ^2	Variance in patient outcomes
	$\boldsymbol{\theta}_t$	Prior/Posterior mean of $\boldsymbol{\mu}$ at time $t = 0,1,\ldots,T+\Delta$
	Σ_t	Prior/Posterior covariance matrix of $\boldsymbol{\mu}$ at time $t = 0,1,\ldots,T+\Delta$
	$\mathbf{J}_t$	Pipeline state at time $t = 0,1,\ldots,T+\Delta$
	$\mathbf{K}_t = (\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t)$	Knowledge state at time $t = 0,1,\ldots,T+\Delta$
	$\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})$	Generic knowledge state
	$\mathscr{K}_t$	Domain of knowledge states at time $t = 0,1,\ldots,T+\Delta$
	τ_t	Updating equation of the knowledge state at time t
Decisions	$\pi = (\pi_t)_{t=1,\ldots,T}$	Allocation policy, where π_t maps $(\mathbf{H}_t, \mathbf{x})$ to a distribution over treatments
	Π	Set of non-anticipatory allocation policies
	$\tilde{f}$	Optimal implementation decision (map from $\mathscr{X} \to \mathscr{W}$) after all samples observed
	$\tilde{f}_{\boldsymbol{\theta}}$	Optimal implementation decision given posterior mean $\boldsymbol{\theta}$
	$\mathbf{f}$	Set of treatment strategies
Expected rewards	P	Size of post-trial population to be treated
	V^{π}	Trial value for policy π, prior to sampling
	V^*	Optimal trial value
	$V_t^{\pi}(\mathbf{k})$	Value-to-go with policy π from time t
	$V_t(\mathbf{k})$	Optimal value-to-go from time t
	$q_t(\mathbf{k}, \mathbf{x}, w)$	Value-to-go if $\mathbf{X}_{t+1} = \mathbf{x}$ and $W_{t+1} = w$
	$Q_t(\mathbf{k}, f)$	Value-to-go if $W_{t+1} = f(\mathbf{X}_{t+1})$
	$U(\boldsymbol{\theta}, \Sigma)$	Oracle's value, assuming perfect information about μ
	$G(\mathbf{k})$	Expected terminal reward before observing outcomes of pipeline patients
	$\tilde{G}(\boldsymbol{\theta})$	Terminal reward after observing all patients and their outcomes
fEVI-MC	η^{on}	Number of replications for unknown regression parameters for EVI estimation
	η^{off}	Number of replications of post-trial patients per sampled regression parameter

tion policies against which we compare our EVI-based allocation policies in the numerical experiments. Appendix EC.5 presents additional numerical experiments to supplement the numerical results of the main paper. Computer code (in the Julia programming language) that implements the numerical results of this manuscript can be found at `https://github.com/andres-alban/EVI-covariates`. Appendix EC.6 discusses additional conceptual and practical considerations for extending the base model. Some are dealt with directly here, while others identify areas for potential further research.

EC.1. Bayesian Inference for Unknown Regression Coefficients

This appendix provides additional discussion for the steps that lead to the Bayesian updating for the unknown regression coefficients in (6) in Section 2.5 of the main paper. It also describes the derivation of the so-called preposterior[5] distribution, the distribution of the posterior mean to be realized after outcomes of pipeline patients are observed, given the current knowledge state.

Bayesian updating. For times $t_1 \leq t_2$, let $\mathbf{X}_{(t_1):(t_2)}$ be the $m \times (t_2 - t_1 + 1)$ matrix where each column corresponds to $\mathbf{X}_{t'}$ for $t' = t_1, t_1+1, \ldots, t_2$, $\mathbf{W}_{(t_1):(t_2)}$ be the $(t_2 - t_1 + 1)$-dimensional column vector where each entry corresponds to $W_{t'}$ for $t' = t_1, t_1+1, \ldots, t_2$, and $\mathbf{Y}_{(t_1):(t_2)}$ be the $(t_2 - t_1 + 1)$-dimensional column vector where each entry corresponds to $Y_{t'}$ for $t' = t_1, t_1+1, \ldots, t_2$. We extend the definition of $\otimes$ to allow for matrix operations as follows:

$$\mathbf{W}_{(t_1):(t_2)} \otimes \mathbf{X}_{(t_1):(t_2)} = \begin{pmatrix} W_{t_1} \otimes \mathbf{X}_{t_1} \\ W_{t_1+1} \otimes \mathbf{X}_{t_1+1} \\ \cdots \\ W_{t_2} \otimes \mathbf{X}_{t_2} \end{pmatrix}, \tag{EC.1}$$

where the result is a $(t_2 - t_1 + 1) \times |\Xi|$ matrix, where $\Xi := \{(i,l) : \xi_{i,l} = 1\}$ is the set of indices of potentially active coefficients in the linear model for a given labeling $\boldsymbol{\xi}$.

Suppose that we want to update the knowledge state at time t_1 to the knowledge state at time $t_2 > t_1$. If $t_2 \leq \Delta$, then no patient outcomes have been observed at time t_1 or t_2, so $\boldsymbol{\theta}_{t_2} = \boldsymbol{\theta}_{t_1} = \boldsymbol{\theta}_0$ and $\Sigma_{t_2} = \Sigma_{t_1} = \Sigma_0$. If $t_2 \geq \Delta + 1$, then the patient outcomes $\mathbf{Y}_{(t_1 - \Delta'_{t_1} + 1):(t_2 - \Delta)}$ are observed in the time window that starts at $t_1 + 1$ and ends at t_2 (inclusive), where $\Delta'_{t_1} = \min\{t_1, \Delta\}$ represents the number of patients in the pipeline at time t_1. If Σ_{t_1} is

[5] We use the term "preposterior" consistent with preposterior analysis used in decision theory, particularly in expected value of information (see Raiffa and Schlaifer 1961).

positive definite, we can use the following result from Bayesian updating (Gelman et al. 2013, Chapter 14):

$$\boldsymbol{\theta}_{t_2} = \left(\Sigma_{t_1}^{-1} + \left(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}\right)^\top \left(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}\right)/\sigma^2\right)^{-1} \left(\left(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}\right)^\top \mathbf{Y}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}/\sigma^2 + \Sigma_{t_1}^{-1}\boldsymbol{\theta}_{t_1}\right) \quad \text{(EC.2a)}$$

$$\Sigma_{t_2} = \left(\Sigma_{t_1}^{-1} + \left(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}\right)^\top \left(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}\right)/\sigma^2\right)^{-1} \quad \text{(EC.2b)}$$

An alternative computation using the Sherman-Morrison-Woodbury matrix identity is valid also when Σ_0 is positive semi-definite (e.g., Frazier et al. 2009):

$$\begin{aligned}\boldsymbol{\theta}_{t_2} = \boldsymbol{\theta}_{t_1} + \Sigma_{t_1}&(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})^\top \\ &\left(\sigma^2 I_{(t_2-t_1)-(\Delta-\Delta'_{t_1})} + (\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\Sigma_{t_1}\right. \\ &\left.(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})^\top\right)^{-1} \\ &\left(\mathbf{Y}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} - (\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\boldsymbol{\theta}_{t_1}\right),\end{aligned} \quad \text{(EC.3a)}$$

$$\begin{aligned}\Sigma_{t_2} = \Sigma_{t_1} - \Sigma_{t_1}&(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})^\top \\ &\left(\sigma^2 I_{(t_2-t_1)-(\Delta-\Delta'_{t_1})} + (\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\Sigma_{t_1}\right. \\ &\left.(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})^\top\right)^{-1} \\ &(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\Sigma_{t_1},\end{aligned} \quad \text{(EC.3b)}$$

where I_t is the identity matrix of size t. This formula requires the inversion of a $((t_2-t_1)-(\Delta-\Delta'_{t_1}))\times((t_2-t_1)-(\Delta-\Delta'_{t_1}))$ matrix, which is not computationally efficient if $(t_2-t_1)-(\Delta-\Delta'_{t_1})$ is much larger than $|\Xi|$. However, when $t_2-t_1=1$ and $t_2\geq\Delta+1$, it simplifies to the recursive equations in (6) in Section 2.5, which require scalar division instead of matrix inversion. We have thus characterized the prior and posterior distributions.

Preposterior distribution. We can also characterize the preposterior distribution, i.e., the distribution of the posterior mean of the coefficients that will be obtained at $t_2\leq t_1+\Delta$ after observing all or some of the patients in the pipeline given the knowledge state at time t_1: $\boldsymbol{\theta}_{t_2} \mid \mathbf{K}_{t_1}$. We can derive the distribution from (EC.3a). For purposes of updating the posterior mean at time t_2, given $\mathbf{K}_{t_1}$, we note that the only source of variation is the unobserved outcomes, $\mathbf{Y}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}$, which standard Bayesian results (Gelman et al. 2013, Chapter 14) show to be normally distributed:

$$\begin{aligned}\mathbf{Y}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \mid \mathbf{K}_{t_1} \sim \mathcal{N}\Big((&\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\boldsymbol{\theta}_{t_1}, \\ &\sigma^2 I_{(t_2-t_1)-(\Delta-\Delta'_{t_1})} + (\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\Sigma_{t_1}(\mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \otimes \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})^\top\Big).\end{aligned} \quad \text{(EC.4)}$$

For a vector of treatments $\mathbf{W}$ of dimension t assigned to t patients with covariates given by matrix $\mathbf{X}$ of dimension $m \times t$, we define the *preposterior standard deviation* as the following $|\Xi| \times t$ matrix:[6]

$$\tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}, \mathbf{X}) = \Sigma(\mathbf{W} \otimes \mathbf{X})^\top (\sigma^2 I_t + (\mathbf{W} \otimes \mathbf{X})\Sigma(\mathbf{W} \otimes \mathbf{X})^\top)^{-1/2}. \tag{EC.5}$$

This term is used in the fEVI-MC algorithm and in mathematical proofs below. We call it the preposterior standard deviation because, combining (EC.4) with (EC.3a), we obtain

$$\boldsymbol{\theta}_{t_2} \mid \mathbf{K}_{t_1} \sim \mathcal{N}\left(\boldsymbol{\theta}_{t_1}, \tilde{\boldsymbol{\sigma}}(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}) \tilde{\boldsymbol{\sigma}}^\top (\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)})\right).$$

Other representations of the inference for proofs. For $t \geq \Delta+1$, we define the random variable

$$Z_{t-\Delta} = \frac{Y_{t-\Delta} - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\boldsymbol{\theta}_{t-1}}{\sqrt{\sigma^2 + (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_{t-1}(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}}, \tag{EC.6}$$

such that $Z_{t-\Delta} \mid \mathbf{K}_{t-1}, \mathbf{X}_t, W_t \sim \mathcal{N}(0,1)$[7]. With this definition, we obtain another representation of (6a) and (6b) in the main paper:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \tilde{\boldsymbol{\sigma}}(\Sigma_{t-1}, W_{t-\Delta}, \mathbf{X}_{t-\Delta}) Z_{t-\Delta} \tag{EC.7}$$

$$\Sigma_t = \Sigma_{t-1} - \tilde{\boldsymbol{\sigma}}(\Sigma_{t-1}, W_{t-\Delta}, \mathbf{X}_{t-\Delta}) \tilde{\boldsymbol{\sigma}}^\top (\Sigma_{t-1}, W_{t-\Delta}, \mathbf{X}_{t-\Delta}). \tag{EC.8}$$

We also define $\mathbf{Z}_{(t_1):(t_2)} = (Z_{t_1}, \ldots, Z_{t_2})^\top$ as the $(t_2 - t_1 + 1)$-dimensional column vector to write the updating equations for multiple observations (EC.3a) and (EC.3b). For $t_2 \geq \Delta+1$, we obtain

$$\boldsymbol{\theta}_{t_2} = \boldsymbol{\theta}_{t_1} + \tilde{\boldsymbol{\sigma}}(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}) \mathbf{Z}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)} \tag{EC.9}$$

$$\Sigma_{t_2} = \Sigma_{t_1} - \tilde{\boldsymbol{\sigma}}(\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}) \tilde{\boldsymbol{\sigma}}^\top (\Sigma_{t_1}, \mathbf{W}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}, \mathbf{X}_{(t_1-\Delta'_{t_1}+1):(t_2-\Delta)}). \tag{EC.10}$$

EC.2. Proof of Proposition 1

We first rewrite the focal dynamic program (DP) in (11) into an alternative DP model, whose state space is augmented to include the covariates of the next patient to be observed. Thus, the alternative DP shifts the time of the decision from before observing the next patient's covariates to right after. The focal and alternative DPs will be shown to have

[6] If $\mathbf{W} = \mathbf{1}_t$, where $\mathbf{1}_t$ denotes the vector of all 1's of dimension t, and $n = 1$ (only one treatment is available), $\mathbf{X}$ is empty, and $\boldsymbol{\xi} = (1, 0)$ (only the intercept term is active), then $(\mathbf{W} \otimes \mathbf{X}) = \mathbf{1}_t$ is a vector of all ones, $(\mathbf{W} \otimes \mathbf{X})^\top (\mathbf{W} \otimes \mathbf{X}) = t$, $|\Xi| = 1$, Σ is a scalar, and we recover the formula for the preposterior variance before obtaining t observations without covariates: $\tilde{\boldsymbol{\sigma}}^2(\Sigma, \mathbf{W}, \mathbf{X}) = \tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}, \mathbf{X}) \tilde{\boldsymbol{\sigma}}^\top (\Sigma, \mathbf{W}, \mathbf{X}) = \Sigma^2 t / (\sigma^2 + \Sigma t)$.

[7] We need to condition on $\mathbf{X}_t$ and W_t to make the distribution valid when $\Delta = 0$.

equivalent rewards and solutions. We then construct a mapping of this alternative model to the model in Bertsekas and Shreve (1978, Sec 8.1). This will allow us to use their results to justify our Proposition 1.

More formally, the alternative DP has state $(\mathbf{K}_t, \mathbf{X}_{t+1})$, where $\mathbf{X}_{t+1} \in \mathscr{X}$ is the covariate of the next patient, for time $t = 0, 1, \ldots, T-1$. For $t = T$, the alternative DP has the same terminal reward as does our focal DP (the term $\mathbf{X}_{T+1}$ will be referenced but its value will be ignored by the updating equation τ_T). The control space of the alternative DP is the set of treatments $\mathscr{W}$ and an allocation policy, π', that maps the history and covariates to a distribution over the treatments: $\pi'_t(w \mid \mathbf{h}, \mathbf{x}) = \mathbb{P}(W_t = w \mid \mathbf{H}_{t-1} = \mathbf{h}, \mathbf{X}_t = \mathbf{x})$ for $t = 1, 2, \ldots, T$. The allocation policies are equivalent in both DP formulations (they have the same definition). We modestly abuse notation by distinguishing the value-to-go function V_t for the alternative DP by giving it two arguments, $(\mathbf{k}, \mathbf{x})$, to distinguish it from that of the focal DP which has one argument $(\mathbf{k})$.

$$\begin{aligned} V_t(\mathbf{k}, \mathbf{x}) &= \sup_{\pi'_{t+1}} \mathbb{E}^{\pi'_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1} = \mathbf{x}], \text{ for } t = 1, 2, \ldots, T-1 \\ V_T(\mathbf{k}, \mathbf{x}) &= G(\mathbf{k}). \end{aligned} \tag{EC.11}$$

Formulations (11) and (EC.11) are equivalent in the sense that $V_t(\mathbf{k}) = \mathbb{E}[V_t(\mathbf{k}, \mathbf{X}_{t+1})]$. This follows by the tower property of expectations, the definition of allocation policy in each formulation, and mathematical induction, as we proceed to show next. The base case is $V_T(\mathbf{k}) = \mathbb{E}[V_T(\mathbf{k}, \mathbf{x})] = G(\mathbf{k})$. The induction step shows that $V_{t+1}(\mathbf{k}) = \mathbb{E}[V_{t+1}(\mathbf{k}, \mathbf{X}_{t+2})]$ implies $V_t(\mathbf{k}) = \mathbb{E}[V_t(\mathbf{k}, \mathbf{X}_{t+1})]$ for any $t = T-1, T-2, \ldots, 0$:

$$\begin{aligned} \mathbb{E}[V_t(\mathbf{k}, \mathbf{X}_{t+1})] &= \mathbb{E}\Big[\sup_{\pi'_{t+1}} \mathbb{E}^{\pi'_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1}]\Big] \\ &= \sup_{\pi'_{t+1}} \mathbb{E}\Big[\mathbb{E}^{\pi'_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1}]\Big] \\ &= \sup_{\pi_{t+1}} \mathbb{E}\Big[\mathbb{E}^{\pi_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1}]\Big] \\ &= \sup_{\pi_{t+1}} \mathbb{E}^{\pi_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{K}_t = \mathbf{k}] \\ &= \sup_{\pi_{t+1}} \mathbb{E}^{\pi_{t+1}}[\mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2}) \mid \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}] \mid \mathbf{K}_t = \mathbf{k}] \\ &= \sup_{\pi_{t+1}} \mathbb{E}^{\pi_{t+1}}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}] \\ &= V_t(\mathbf{k}), \end{aligned}$$

where the first equality is by the definition in (EC.11), the second and third equalities follow by our definition of π'_{t+1}, which allows us to exchange the order of the supremum and the outer expectation, noting that the outer expectation is over $\mathbf{X}_{t+1}$ only, the fourth equality follows by the tower property of expectations, the fifth equality is also by the tower property of expectations, the sixth equality uses the induction hypothesis, and the last equality is by the definition in (11).

We now map our alternative DP model in (EC.11) to the model in Bertsekas and Shreve (1978, Section 8.1). The state space in (EC.11) is $\mathscr{K} \times \mathscr{X}$. The control space is the finite set of treatments $\mathscr{W}$. The disturbance space is given by $\mathscr{Y} \times \mathscr{X}$ and is handled in two cases. If $t \geq \Delta$, for a given state $(\mathbf{k}, \mathbf{x})$ and control w, the disturbance $Y_{t-\Delta+1} \in \mathscr{Y}$ has a stochastic kernel given by $Y_{t-\Delta+1} \mid \mathbf{k} \sim \mathcal{N}\left((W_{t-\Delta+1} \otimes \mathbf{X}_{t-\Delta+1})\boldsymbol{\theta}, \sigma^2 + (W_{t-\Delta+1} \otimes \mathbf{X}_{t-\Delta+1})\Sigma(W_{t-\Delta+1} \otimes \mathbf{X}_{t-\Delta+1})^\top\right)$ (recall $\mathbf{X}_{t-\Delta+1}$ and $W_{t-\Delta+1}$ are in the knowledge state), and, independently, $\mathbf{X}_{t+2} \in \mathscr{X}$ has a stochastic kernel given by F_x regardless of the state. For $t < \Delta$, we can define the disturbance $Y_{t-\Delta+1}$ arbitrarily (e.g., $Y_{t-\Delta+1} = 0$ with probability 1) because such values of Y_t are not used by the transition function when $t < \Delta$ and do not affect the model, and $\mathbf{X}_{t+2}$ has the kernel given by F_x regardless of the state. The system function (transition function) of Bertsekas and Shreve (1978), $\mathrm{f}_t(\mathbf{k}, \mathbf{x}, w, (Y_{t-\Delta+1}, \mathbf{X}_{t+2}))$, is our transition function that does the Bayesian updating, $(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1}), \mathbf{X}_{t+2})$. The discount factor is one. The one-stage cost function is $g(\mathbf{k}, \mathbf{x}, w) = 0$ for $t = 1, \ldots, T-1$ and $g(\mathbf{k}, \mathbf{x}, w) = G(\mathbf{k})$ for $t = T$. The horizon is T. Finally, all entries in Σ_0 are finite, and using notation $\Sigma_{t,i,j}$ to denote the (i,j) entry of matrix Σ_t, each diagonal entry $\Sigma_{t,i,i}$ is non-increasing in t (Xie et al. 2016, Lemma 6), and $\Sigma^2_{t,i,j} \leq \Sigma_{t,i,i}\Sigma_{t,j,j}$ by the positive semi-definiteness of Σ_t. Thus, each element of Σ_t is bounded. By Assumption 1, we know that $\mathbf{X}_t$ and $\tilde{\mathbf{X}}_i$ have at least two finite moments. Therefore, $\mathbb{E}^\pi[|g(\mathbf{K}_t, \mathbf{X}_{t+1}, W_{t+1})|]$ exists (is bounded) for all $t = 1, \ldots, T$ and any $\pi \in \Pi$, so we satisfy condition (F^+) and (F^-).

Because $\mathscr{W}$ is finite, by Corollary 8.1.1 and Proposition 8.5 of Bertsekas and Shreve (1978), there exists a deterministic, Markov allocation policy π'^* that achieves the supremum in (EC.11). Moreover, $V_t(k) = \mathbb{E}[V_t(\mathbf{k}, \mathbf{X}_{t+1})]$, so an allocation policy that solves (EC.11) gives an allocation policy that solves (11), so these results also hold for (11). □

EC.3. Proofs of Results in Section 5

We first prove optimality for the fEVI allocation policy with sample size $T = 1$ and then turn to the asymptotic results for the EVI heuristics presented in Section 4.

PROPOSITION 2 *If $T = 1$, then $V^{f\mathrm{EVI}}(\mathbf{K}_0) = V^*(\mathbf{K}_0)$.*

Proof. By Proposition 1, the optimal allocation policy π^* can be computed using Bellman's recursion. Thus, there is a π^* that maximizes

$$\mathbb{E}^\pi[G(\tau_1(\mathbf{K}_0, \mathbf{X}_1, W_1, Y_{1-\Delta})) \mid \mathbf{K}_0] = \mathbb{E}^\pi\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{\Delta+1} \mid \mathbf{K}_0\right].$$

The fEVI allocation policy, which samples the treatment with the highest fEVI-index defined in (12), maximizes the same expectation. □

In the rest of this section we assume that $T \geq \Delta + 1$ for two reasons. First, the main result we are set out to prove is Theorem 3, which is an asymptotic result as $T \to \infty$, so the assumption that $T \geq \Delta + 1$ does not affect it. Second, this allows us to avoid handling a variable number of patients in the pipeline, when $T < \Delta$, without losing any insights for the asymptotic regime.

We first recall some related results from the literature and describe how we adapt our model to be able to use them and build on them (Appendix EC.3.1). Then, in Appendix EC.3.2 we describe and prove properties of the value functions that will be useful for proving our main analytical results. Appendix EC.3.3 provides some additional intermediate results and uses them to prove the main theorems in Section 5.

EC.3.1. Preliminaries for asymptotic results

We use many elements of the proofs in Frazier et al. (2009, Online supplement A) and Xie et al. (2016), extending the statements and proof methodology to our setting with covariates where necessary. The proofs rely on the DP model in Section 3 through the value functions $V_t(\mathbf{k})$ (optimal value-to-go with knowledge state $\mathbf{k}$), $q_t(\mathbf{k}, \mathbf{x}, w)$ (value-to-go if treatment w is given to patient $t+1$ with covariates $\mathbf{x}$), $Q_t(\mathbf{k}, f)$ (value-to-go if treatment strategy f is used to treat patient $t+1$), and $V_t^\pi(\mathbf{k})$ (value-to-go of policy π):

$$
\begin{aligned}
V_T(\mathbf{k}) &= G(\mathbf{k}) = \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_T = \mathbf{k}] \\
V_t(\mathbf{k}) &= \max_{f \in \mathbf{f}} \mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1}), Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}],\ t = 0, \ldots, T-1 \\
q_t(\mathbf{k}, \mathbf{x}, w) &= \mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{x}, w, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}],\ t = 0, \ldots, T-1 \\
Q_t(\mathbf{k}, f) &= \mathbb{E}[q_t(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1})) \mid \mathbf{K}_t = \mathbf{k}] \\
&= \mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1}), Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}],\ t = 0, \ldots, T-1 \\
V_T^\pi(\mathbf{k}) &= V_T(\mathbf{k}) \\
V_t^\pi(\mathbf{k}) &= \mathbb{E}^\pi[V_{t+1}^\pi(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, W_{t+1}, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}],\ t = 0, \ldots, T-1
\end{aligned}
$$

Here, the action at each time point is a treatment strategy f that deterministically maps covariates to treatments, unlike (11) where the actions are allowed to be random. We do so because Proposition 1 guarantees the existence of an optimal deterministic allocation

policy. The value-to-go of an allocation policy is equivalently defined by expanding the recursive definition above:

$$V_t^\pi(\mathbf{k}) = \mathbb{E}^\pi[V_T(\mathbf{K}_T) \mid \mathbf{K}_t = \mathbf{k}].$$

We further define

$$\tilde{G}(\hat{\boldsymbol{\theta}}) := \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_{T+\Delta} = (\hat{\boldsymbol{\theta}}, \Sigma_{T+\Delta}, \emptyset)\right] = \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\hat{\boldsymbol{\theta}}\right] \tag{EC.12}$$

as the value after all patient outcomes have been observed (and the pipeline is empty) with posterior mean $\hat{\boldsymbol{\theta}}$ at time $T+\Delta$. This term does not depend on $\Sigma_{T+\Delta}$. Thus, the terminal reward is $G(\mathbf{k}) = \mathbb{E}[\tilde{G}(\boldsymbol{\theta}_{T+\Delta}) \mid \mathbf{K}_T = \mathbf{k}]$.

EC.3.2. Properties of the value functions

We provide several analytical results regarding the value functions that are useful for proofs, and that provide results that are analogous to results shown elsewhere when there is no delay and no predictive-prognostic structure. These results will show that the value functions are "well-behaved".

Continuity. We first show that the value functions are continuous in the posterior mean and covariance matrix for any pipeline state.

PROP. EC.1. *$V_t(\mathbf{k})$, $Q_t(\mathbf{k}, f)$, and $q_t(\mathbf{k}, \mathbf{x}, w)$ are continuous in $\boldsymbol{\theta}$ and Σ for any $\mathbf{j}$ such that $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j}) \in \mathscr{K}_t$ is a valid state at time t.*

Proof. The proof is by induction. The base case is that $V_T(\mathbf{k}) = \mathbb{E}[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_T = \mathbf{k}]$ is continuous because it is the expectation of a continuous function in both $\boldsymbol{\theta}$ and Σ (Billingsley 2008, Theorem 16.8). To see this, note that $\boldsymbol{\theta}_{T+\Delta}$ is updated with (EC.3a), which is a continuous function of $\boldsymbol{\theta}$ and Σ.

The induction step assumes that $V_{t+1}(\mathbf{k})$ is continuous in $\boldsymbol{\theta}$ and Σ for all $\mathbf{k} \in \mathscr{K}_{t+1}$. Then, $V_t(\mathbf{k}) = \max_{f\in\mathbf{f}} \mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1}), Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}] = \mathbb{E}[\max_{w\in\mathscr{W}} V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{X}_{t+1}, w, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}]$ is the expectation of the maximum of continuous functions. The maximum of a finite set of continuous functions is continuous, and the expectation of continuous functions is continuous.

Similarly, $q_t(\mathbf{k}, \mathbf{x}, w) = \mathbb{E}[V_{t+1}(\tau_{t+1}(\mathbf{k}, \mathbf{x}, w, Y_{t-\Delta+1})) \mid \mathbf{K}_t = \mathbf{k}]$ is the expectation of a continuous function for any $\mathbf{x} \in \mathscr{X}$ and $w \in \mathscr{W}$. Finally, $Q_t(\mathbf{k}, f) = \mathbb{E}[q_t(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1})) \mid \mathbf{K}_t = \mathbf{k}]$ is the expectation of a continuous function. □

Value of information. The next proposition shows that it is more valuable in expectation to have an additional observation, regardless of the covariates and treatment, than having no observation at all, if sampling is optimal thereafter.

PROP. EC.2. *$q_t(\mathbf{k}, \mathbf{x}, w) \geq V_{t+1}(\mathbf{k})$ for $t = \Delta, \ldots, T-1$, any $\mathbf{k} \in \mathscr{K}_t$, any $\mathbf{x} \in \mathscr{X}$, and any $w \in \mathscr{W}$.*

Proof. To prove the claimed inequality, we show that an allocation policy that discards the information from patient $t+1$ obtains the same value as not having the observation at all, which corresponds to the smaller side of the inequality. Moreover, such a policy is within the set of possible policies, and therefore it cannot obtain a larger value than does the optimal policy, which corresponds to the larger side of the inequality. We now formalize this argument.

First, fix a given $t = \Delta, \ldots, T-1$. The assumption $t \geq \Delta$ implies that the pipeline is full with exactly Δ treatments whose outcomes are not yet observed. Let $\pi_{[t+1]} = (\pi_{t'})_{t'=t+2,\ldots,T}$ be an allocation policy for the allocation of patients starting at $t+2$ and $\Pi_{[t+1]}$ be the set of all such allocation policies. Let $\pi^*_{[t+1]} \in \Pi_{[t+1]}$ be an optimal allocation policy, so that

$$V_{t+1}(\mathbf{k}) = \mathbb{E}^{\pi^*_{[t+1]}}\left[V_T(\mathbf{K}_T) \mid \mathbf{K}_{t+1} = \mathbf{k}\right] = \mathbb{E}^{\pi^*_{[t+1]}}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_{t+1} = \mathbf{k}\right]. \quad \text{(EC.13)}$$

Thus, $q_t(\mathbf{k}, \mathbf{x}, w)$ is the expected value from sampling a patient with covariates $\mathbf{x}$ and treatment w at $t+1$, and using $\pi^*_{[t+1]}$ for the remaining patients:

$$q_t(\mathbf{k}, \mathbf{x}, w) = \mathbb{E}^{\pi^*_{[t+1]}}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1} = \mathbf{x}, W_{t+1} = w\right]$$

Consider the set of policies $\tilde{\Pi}_{[t+1]}$ constrained to only see a subset of the history $\tilde{\mathbf{H}}_{t'} \subset \mathbf{H}_{t'}$ for $t' \geq t+1$, which we now describe. First, $\tilde{\mathbf{H}}_{t'}$ is defined so that it does not include the data from the patient arriving at $t+1$, i.e., $\mathbf{X}_{t+1}$, W_{t+1}, and Y_{t+1} are not included in the histories $\tilde{\mathbf{H}}_{t'}$ for $t' \geq t+1$, and therefore they are not available to policies in $\tilde{\Pi}_{[t+1]}$. Second, for $t'' = t-\Delta, \ldots, t$, $Y_{t''}$ is observed with one additional unit of delay, i.e., $Y_{t''}$ is included in $\tilde{\mathbf{H}}_{t''+\Delta+1}$ (and all later histories) but is missing in $\tilde{\mathbf{H}}_{t''+\Delta}$ (whereas it is present in $\mathbf{H}_{t''+\Delta}$).

Let $\tilde{\boldsymbol{\theta}}_{T+\Delta}$ be the posterior mean excluding patient $t+1$ (i.e., the terminal posterior mean when using history $\tilde{\mathbf{H}}_{T+\Delta}$) and let $\tilde{\pi}^*_{[t+1]}$ be the policy that maximizes value among all policies in $\tilde{\Pi}_{[t+1]}$:

$$\tilde{\pi}^*_{[t+1]} = \operatorname*{arg\,max}_{\tilde{\pi} \in \tilde{\Pi}_{[t+1]}} \mathbb{E}^{\tilde{\pi}}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\tilde{\boldsymbol{\theta}}_{T+\Delta} \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1} = \mathbf{x}, W_{t+1} = w\right],$$

where $\mathbf{x}$ and w do no affect the expectation because they are not included in the history available to policies in $\tilde{\Pi}_{[t+1]}$.

We define $\tilde{q}_t$ as the maximum value-to-go if a patient with covariates $\mathbf{x}$ is allocated to treatment w at $t+1$ under the additional constraint that the allocation policy belongs to $\tilde{\Pi}_{[t+1]} \subset \Pi_{[t+1]}$:

$$\tilde{q}_t(\mathbf{k}, \mathbf{x}, w) := \mathbb{E}^{\tilde{\pi}^*_{[t+1]}} \left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1) \tilde{\boldsymbol{\theta}}_{T+\Delta} \mid \mathbf{K}_t = \mathbf{k}, \mathbf{X}_{t+1} = \mathbf{x}, W_{t+1} = w \right].$$

By construction, $\tilde{\mathbf{H}}_{t'}$ is the same available history as if $\mathbf{K}_{t+1} = \mathbf{k}$ (instead of $\mathbf{K}_t = \mathbf{k}$), and therefore, $\tilde{q}_t(\mathbf{k}, \mathbf{x}, w) = V_{t+1}(\mathbf{k})$. Moreover, because $\tilde{\pi}^*_{[t+1]}$ is among the available policies $\Pi_{[t+1]}$, the optimal policy obtains higher or the same value: $q_t(\mathbf{k}, \mathbf{x}, w) \geq \tilde{q}_t(\mathbf{k}, \mathbf{x}, w)$. Putting these two observations together, we get the final result:

$$q_t(\mathbf{k}, \mathbf{x}, w) \geq \tilde{q}_t(\mathbf{k}, \mathbf{x}, w) = V_{t+1}(\mathbf{k}). \quad \square$$

The following corollaries show how Proposition EC.2 maps to value functions V_t and Q_t.

COROLLARY EC.1. *$Q_t(\mathbf{k}, f) \geq V_{t+1}(\mathbf{k})$ for $t = \Delta, \ldots, T-1$, any $\mathbf{k} \in \mathscr{K}_t$, and any $f \in \mathbf{f}$.*

Proof. By the definition of Q_t, we get $Q_t(\mathbf{k}, f) = \mathbb{E}[q_t(\mathbf{k}, \mathbf{X}_{t+1}, f(\mathbf{X}_{t+1})) \mid \mathbf{K}_t = \mathbf{k}] \geq V_{t+1}(\mathbf{k})$, where the inequality follows from Proposition EC.2. $\square$

COROLLARY EC.2. *$V_t(\mathbf{k}) \geq V_{t+1}(\mathbf{k})$ for $t = \Delta, \ldots, T-1$ and any $\mathbf{k} \in \mathscr{K}_t$.*

Proof. From the definition of Q_t, we have $V_t(\mathbf{k}) = \max_f Q_t(\mathbf{k}, f) \geq V_{t+1}(\mathbf{k})$, where the inequality follows from Corollary EC.1. $\square$

Boundedness. We now show that the value function is bounded. We show that an upper bound is the value obtained by an oracle that knows the value of $\boldsymbol{\mu}$ ex ante, which we redefine here to generalize the definition of $U(\boldsymbol{\theta}_0, \Sigma_0)$ in Section 5. This generalization allows for knowledge states, $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})$, and uses the fact that perfect knowledge of $\boldsymbol{\mu}$ does not depend on $\mathbf{j}$.

$$U(\boldsymbol{\theta}, \Sigma) := U(\mathbf{k}) = \mathbb{E}[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1) \boldsymbol{\mu} \mid \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \Sigma_0 = \Sigma]. \tag{EC.14}$$

We will call $U(\boldsymbol{\theta}, \Sigma)$ the *oracle's value*. The proof of Theorem 2 below will give some sufficient conditions on allocation policies to achieve this upper bound.

PROP. EC.3. *For $t = 0, \ldots, T$, for any $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j}) \in \mathscr{K}_t$, and any allocation policy π, $V^\pi_t(\mathbf{k})$ is bounded above by $U(\boldsymbol{\theta}, \Sigma)$ and below by $\mathbb{E}[\min_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1) \boldsymbol{\mu} \mid \mathbf{K}_t = \mathbf{k}]$.*

Proof. For any allocation policy π and knowledge state $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j}) \in \mathscr{K}_t$,

$$
\begin{aligned}
V_t^\pi(\mathbf{k}) &= \mathbb{E}^\pi[V_T^\pi(\mathbf{K}_T) \mid \mathbf{K}_t = \mathbf{k}] \\
&= \mathbb{E}^\pi\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_t = \mathbf{k}\right] \\
&= \mathbb{E}^\pi\left[\max_{\tilde{w}\in\mathscr{W}} ((\tilde{w} \otimes \tilde{\mathbf{X}}_1)\mathbb{E}[\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta}]) \mid \mathbf{K}_t = \mathbf{k}\right] \\
&\leq \mathbb{E}^\pi\left[\mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta}\right] \mid \mathbf{K}_t = \mathbf{k}\right] \qquad \text{(EC.15)} \\
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_t = \mathbf{k}\right] \qquad \text{(EC.16)} \\
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \Sigma_0 = \Sigma\right] \qquad \text{(EC.17)} \\
&= U(\boldsymbol{\theta}, \Sigma).
\end{aligned}
$$

The inequality in (EC.15) follows from Jensen's inequality and the convexity of the maximization operator. The tower property of expectations justifies (EC.16). The step in (EC.17) follows because the expectation does not depend on the time but only on the distribution of $\boldsymbol{\mu}$, which is fully given by $\boldsymbol{\theta}$ and Σ. The same argument using Jensen's inequality and the concavity of the minimization operator shows the lower bound. □

EC.3.3. Analysis for Theorems 1, 2, and 3

Equipped with the results about value functions in Appendix EC.3.2, we now turn to prove the main theorems for asymptotic optimality stated in Section 5.

Proofs of the asymptotic optimality of EVI algorithms without covariates typically rely on showing that each alternative (with an uncertain mean performance) is sampled infinitely often. Such condition is not sufficient in our setting: some allocation policies may sample infinitely often each alternative and still be inconsistent. To illustrate, consider Example 1 and an allocation policy that always allocates patients of type A to arm 1 and patients of type B to arm 2. Such an allocation policy allocates both treatments infinitely often but fails to learn the expected outcomes for type-treatment combinations A-2 and B-1. Thus, the allocation policy cannot make perfectly informed decisions as the number of observed patients grows large.

Therefore, we use a different approach for the proofs of our results. We will construct a contradiction by first showing that any non-anticipatory allocation policy $\pi \in \Pi$ has posterior parameters converging to some $\boldsymbol{\theta}_\infty^\pi$ and Σ_∞^π (Lemma EC.1). Then, we will show

that, if an allocation policy with some special properties (sampling the treatment with largest EVI with positive probability, see the premise of Theorem 2) did not reach a knowledge state with zero value of information, the posterior parameters would not converge (Lemma EC.3). Furthermore, we show that an allocation policy that reaches a state with zero value of information is asymptotically optimal (Lemma EC.2). This shows that an allocation policy with such special properties is asymptotically optimal.

Here, we characterize the trial value as a function of the sample size T. Let $V^\pi(\mathbf{K}_0; T)$ be the value of allocation policy π with prior knowledge state $\mathbf{K}_0$ and sample size T.

EC.3.3.1. Convergence of parameters and value function — Proof of Theorem 1. We now provide convergence results that are used to build the proof of Theorem 1. Lemma EC.1 guarantees that the posterior mean vector and covariance matrix converge. However, it does not guarantee that all allocation policies converge to the same random variables, i.e., the parameters $\boldsymbol{\theta}_\infty^\pi$ and Σ_∞^π that an allocation policy π converges to may depend on the allocation policy π. For instance, an allocation policy that always samples the same treatment might not converge to the same posterior parameters as a policy that randomizes uniformly over all treatments.

LEMMA EC.1. *For a given allocation policy π, there exist random variables $\boldsymbol{\theta}_\infty^\pi$ and Σ_∞^π such that $\boldsymbol{\theta}_t$ converges to $\boldsymbol{\theta}_\infty^\pi$ and Σ_t converges to Σ_∞^π almost surely.*

Proof. Let $M_t = (\boldsymbol{\theta}_t, \Sigma_t + \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top)$. Because $\boldsymbol{\theta}_t$ and Σ_t are continuous transformations of M_t, by the continuous mapping theorem (Van der Vaart 2000, Theorem 2.3) it is sufficient to show that M_t converges almost surely as $t \to \infty$.

We first show that M_t is a martingale relative to the filtration induced by $\mathbf{K}_t$. Because

$$\boldsymbol{\theta}_t = \mathbb{E}[\boldsymbol{\mu} \mid \mathbf{K}_t] \text{ and } \Sigma_t + \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top = \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^\top \mid \mathbf{K}_t]$$

are conditional expectations of integrable random variables, M_t is integrable. Moreover, using (6a) and (6b), we obtain [8]

$$\begin{aligned}
\mathbb{E}^\pi[\boldsymbol{\theta}_{t+1} \mid \mathbf{K}_t] &= \boldsymbol{\theta}_t + \mathbb{E}^\pi\left[\frac{Y_{t+1} - (W_{t+1} \otimes \mathbf{X}_{t+1})\boldsymbol{\theta}_t}{\sigma^2 + (W_{t+1} \otimes \mathbf{X}_{t+1})\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top}\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top \,\middle|\, \mathbf{K}_t\right] \\
&= \boldsymbol{\theta}_t + \mathbb{E}^\pi\left[\mathbb{E}\left[\frac{Y_{t+1} - (W_{t+1} \otimes \mathbf{X}_{t+1})\boldsymbol{\theta}_t}{\sigma^2 + (W_{t+1} \otimes \mathbf{X}_{t+1})\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top}\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top \,\middle|\, \mathbf{X}_{t+1}, W_{t+1}\right] \,\middle|\, \mathbf{K}_t\right] \\
&= \boldsymbol{\theta}_t + \mathbb{E}^\pi\left[\frac{\mathbb{E}[Y_{t+1} - (W_{t+1} \otimes \mathbf{X}_{t+1})\boldsymbol{\theta}_t \mid \mathbf{X}_{t+1}, W_{t+1}]}{\sigma^2 + (W_{t+1} \otimes \mathbf{X}_{t+1})\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top}\Sigma_t(W_{t+1} \otimes \mathbf{X}_{t+1})^\top \,\middle|\, \mathbf{K}_t\right] \\
&= \boldsymbol{\theta}_t
\end{aligned}$$

[8] Here, we assume $\Delta = 0$. If $\Delta > 0$, the only difference is that we subtract Δ in the subscripts of $\mathbf{X}_{t+1}$, W_{t+1}, and Y_{t+1}. In that case, we can skip the conditioning on $\mathbf{X}_{t+1}$ and W_{t+1} because they are given by the knowledge state.

and

$$\begin{aligned}
\mathbb{E}^\pi[\Sigma_{t+1}+\boldsymbol{\theta}_{t+1}\boldsymbol{\theta}_{t+1}^\top \mid \mathbf{K}_t] &= \Sigma_t - \mathbb{E}^\pi\left[\frac{\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t}{\sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top}\,\middle|\,\mathbf{K}_t\right]+\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top \\
&\quad + 2\mathbb{E}^\pi\left[\frac{Y_{t+1}-(W_{t+1}\otimes\mathbf{X}_{t+1})\boldsymbol{\theta}_t}{\sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top}\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top\boldsymbol{\theta}_t^\top\,\middle|\,\mathbf{K}_t\right] \\
&\quad + \mathbb{E}^\pi\left[\frac{(Y_{t+1}-(W_{t+1}\otimes\mathbf{X}_{t+1})\boldsymbol{\theta}_t)^2}{(\sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top)^2}\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top\right. \\
&\qquad\qquad \left.(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t\,\middle|\,\mathbf{K}_t\right] \\
&= \Sigma_t+\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top - \mathbb{E}^\pi\left[\frac{\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t}{\sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top}\,\middle|\,\mathbf{K}_t\right] \\
&\quad + \mathbb{E}^\pi\left[\frac{\mathbb{E}[(Y_{t+1}-(W_{t+1}\otimes\mathbf{X}_{t+1})\boldsymbol{\theta}_t)^2\mid\mathbf{X}_{t+1},W_{t+1}]}{(\sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top)^2}\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top\right. \\
&\qquad\qquad \left.(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t\,\middle|\,\mathbf{K}_t\right] \\
&= \Sigma_t+\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top,
\end{aligned}$$

where we have used $\mathbb{E}[Y_{t+1}\mid\mathbf{K}_t,\mathbf{X}_{t+1},W_{t+1}]=(W_{t+1}\otimes\mathbf{X}_{t+1})\boldsymbol{\theta}_t$, as well as the law of total variance to obtain

$$\begin{aligned}
\mathbb{E}\left[(Y_{t+1}-(W_{t+1}\otimes\mathbf{X}_{t+1})\boldsymbol{\theta}_t)^2\mid\mathbf{K}_t,\mathbf{X}_{t+1},W_{t+1}\right] &= \mathbb{V}[Y_{t+1}\mid\mathbf{K}_t,\mathbf{X}_{t+1},W_{t+1}] \\
&= \mathbb{E}\left[\mathbb{V}[Y_{t+1}\mid\boldsymbol{\mu}]\mid\mathbf{K}_t,\mathbf{X}_{t+1},W_{t+1}\right] \\
&\quad + \mathbb{V}[\mathbb{E}[Y_{t+1}\mid\boldsymbol{\mu}]\mid\mathbf{K}_t,\mathbf{X}_{t+1},W_{t+1}] \\
&= \sigma^2+(W_{t+1}\otimes\mathbf{X}_{t+1})\Sigma_t(W_{t+1}\otimes\mathbf{X}_{t+1})^\top.
\end{aligned}$$

That M_t is a martingale with finite expectation guarantees the claimed almost sure convergence (for example, Billingsley 2008, Section 35, Theorem 35.5). □

A corollary of Lemma EC.1 is that the value function converges as $T\to\infty$.

Corollary EC.3. *For any allocation policy π, $V^\pi(\mathbf{K}_0;\infty) := \lim_{T\to\infty}V^\pi(\mathbf{K}_0;T) = \mathbb{E}[\tilde{G}(\boldsymbol{\theta}_\infty^\pi)\mid\mathbf{K}_0]$, where $\tilde{G}$ is as defined in (EC.12).*

Proof. Fix an allocation policy π. Let $g_T=(\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ be the random variable capturing the terminal reward such that $\mathbb{E}^\pi[g_T\mid\mathbf{K}_0]=V^\pi(\mathbf{K}_0;T)$. g_T is bounded above by $\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ and below by $\min_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}$, both of which are integrable. By the definition of $\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}$ in (8) and Lemma EC.1, we have that $g_T\to(\tilde{f}_{\boldsymbol{\theta}_\infty^\pi}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ almost surely. By the dominated convergence theorem (e.g., Billingsley 2008, Theorem 16.4), we get that $\lim_{T\to\infty}\mathbb{E}[g_T\mid\mathbf{K}_0]=\mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}_\infty^\pi}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}\mid\mathbf{K}_0]$, or equivalently

$$\lim_{T\to\infty}V^\pi(\mathbf{K}_0;T)=\lim_{T\to\infty}\mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}_{T+\Delta}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}\mid\mathbf{K}_0]$$

$$
\begin{aligned}
&= \mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}^\pi_\infty}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_0] \\
&= \mathbb{E}[\mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}^\pi_\infty}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}^\pi_\infty] \mid \mathbf{K}_0] \\
&= \mathbb{E}[(\tilde{f}_{\boldsymbol{\theta}^\pi_\infty}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}^\pi_\infty \mid \mathbf{K}_0] \\
&= \mathbb{E}[\tilde{G}(\boldsymbol{\theta}^\pi_\infty) \mid \mathbf{K}_0]. \quad \square
\end{aligned}
$$

These results cumulatively justify Theorem 1.

THEOREM 1 *For a given allocation policy π, there exists a random vector $\boldsymbol{\theta}^\pi_\infty$ and a random matrix Σ^π_∞ such that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^\pi_\infty$ almost surely and $\Sigma_t \to \Sigma^\pi_\infty$ almost surely. Moreover, $V^\pi(\mathbf{K}_0; T)$ is bounded above by $U(\boldsymbol{\theta}_0, \Sigma_0)$, and $V^\pi(\mathbf{K}_0; \infty) := \lim_{T\to\infty} V^\pi(\mathbf{K}_0; T)$ exists.*

Proof. The claims follow directly from Lemma EC.1, Proposition EC.3 and Corollary EC.3, respectively. $\square$

EC.3.3.2. A sufficient condition to obtain the oracle's value. Theorem 1 showed the convergence of posterior means and covariances for the unknown rewards and the convergence of the trial value, but does not guarantee that the asymptotic trial value obtained by an allocation policy is optimal. We now prove two lemmas to show conditions under which the asymptotic trial value converges to the optimal value obtained by the oracle, in building to the proof of Theorem 2. We then show that the EVI heuristics in Section 4 satisfy those conditions in the proof of Theorem 3.

Lemma EC.2 shows that if there is no value in sampling any treatment on any patient, then the value of that knowledge state is equal to the value that an oracle would obtain. In the proof of Theorem 2, we will show that any allocation policy that samples the treatment with the largest one-step look-ahead value with positive probability will almost surely converge to a state where there is no value in further sampling. Lemma EC.2 will then justify that any such allocation policy (e.g., the fEVI policies) converges to the oracle's value $U(\boldsymbol{\theta}, \Sigma)$.

LEMMA EC.2. *Assume $T \geq \Delta + 1$. If $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j}) \in \mathscr{K}_T$ is such that $Q_{T-1}(\mathbf{k}, f) = G(\mathbf{k})$ for all $f \in \mathbf{f}$, then $G(\mathbf{k}) = \tilde{G}(\boldsymbol{\theta}) = U(\boldsymbol{\theta}, \Sigma)$, where $\tilde{G}$ is as in (EC.12).*

Proof. We fix the knowledge state $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})$. Using the notation introduced in Appendix EC.1, $\mathbf{X}_{(T-\Delta):(T-1)}$ and $\mathbf{W}_{(T-\Delta):(T-1)}$ are the matrix of covariates and vector of treatments of the patients arriving at times $T-\Delta, T-\Delta+1, \ldots, T-1$. At time $T-1$, the

covariates of these patients are in the knowledge state. The covariates from the patient that arrives at time T are still unobserved and hence random. Let $\mathbf{Z}_{(T-\Delta):(T-1)}$ be the vector of normalized outcomes of the patients arriving at times $T-\Delta, T-\Delta+1, \ldots, T-1$ as defined in Appendix EC.1.

We consider the difference $Q_{T-1}(\mathbf{k}, f) - G(\mathbf{k})$. Instead of using the transition function τ_T to write $Q_{T-1}(\mathbf{k}, f)$, we use an expectation conditional on allocating the next patient with treatment strategy f:

$$\begin{aligned} Q_{T-1}(\mathbf{k}, f) &= \mathbb{E}[G(\tau_T(\mathbf{k}, \mathbf{X}_T, f(\mathbf{X}_T), Y_{T-\Delta}) \mid \mathbf{K}_{T-1} = \mathbf{k}] \\ &= \mathbb{E}[\mathbb{E}[G(\mathbf{K}_T) \mid \mathbf{X}_T, W_T = f(\mathbf{X}_T)] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\ &= \mathbb{E}\left[\mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{X}_T, W_T = f(\mathbf{X}_T)\right] \mid \mathbf{K}_{T-1} = \mathbf{k}\right], \end{aligned}$$

where the second line follows because $\mathbf{K}_T = \tau_T(\mathbf{K}_{T-1}, \mathbf{X}_T, W_T, Y_{T-\Delta})$ and by the tower property of expectations, and the last step follows from plugging in (9) for $G(\mathbf{k})$ and using the tower property of expectations.

The terminal reward $G(\mathbf{k})$ is the reward of making an implementation decision after observing the outcomes of the Δ patients in the pipeline. While its definition is specific to the patients in the pipeline at $t = T-\Delta, \ldots, T$, we can shift the pipeline to any other time point that has a full pipeline. In particular, here we use

$$\begin{aligned} G(\mathbf{k}) &= \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_T = \mathbf{k}\right] \\ &= \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T-1} = \mathbf{k}\right]. \end{aligned} \tag{EC.18}$$

Thus, we obtain the following, where we let $\tilde{\boldsymbol{\sigma}}_T := \tilde{\boldsymbol{\sigma}}(\Sigma_{T+\Delta-1}, f(\mathbf{X}_T), \mathbf{X}_T)$ to simplify notation:

$$\begin{aligned} &Q_{T-1}(\mathbf{k}, f) - G(\mathbf{k}) \\ &\quad= \mathbb{E}\left[\mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{X}_T, W_T = f(\mathbf{X}_T)\right] \mid \mathbf{K}_{T-1} = \mathbf{k}\right] - \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T-1} = \mathbf{k}\right] \\ &\quad= \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} \left((\tilde{w} \otimes \tilde{\mathbf{X}}_1)\left(\boldsymbol{\theta}_{T+\Delta-1} + \tilde{\boldsymbol{\sigma}}_T Z_T\right)\right) \mid \mathbf{K}_{T-1} = \mathbf{k}\right] \\ &\qquad\quad - \mathbb{E}\left[\left(\left(\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1\right)(\boldsymbol{\theta}_{T+\Delta-1})\right) \mid \mathbf{K}_{T-1} = \mathbf{k}\right] \\ &\quad= \mathbb{E}\Bigg[\max_{\tilde{w} \in \mathscr{W}} \Bigg(\left((\tilde{w} \otimes \tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\right)\left(\boldsymbol{\theta}_{T+\Delta-1} + \tilde{\boldsymbol{\sigma}}_T Z_T\right)\Bigg) \\ &\qquad\quad + \left(\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1\right)\tilde{\boldsymbol{\sigma}}_T Z_T \mid \mathbf{K}_{T-1} = \mathbf{k}\Bigg], \end{aligned}$$

where in the second step we have used (EC.7) to update $\boldsymbol{\theta}_{T+\Delta}$ from the knowledge state $\mathbf{K}_{T+\Delta-1}$. Because Z_T is a standard normal random variable independent of $\mathbf{X}_T$ and $\tilde{\mathbf{X}}_1$, the last term in the expectation is zero, and we can write

$$Q_{T-1}(\mathbf{k}, f) - G(\mathbf{k}) = \mathbb{E}\Bigg[\max_{\tilde{w}\in\mathscr{W}}\Bigg(\Big((\tilde{w}\otimes\tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Big)\boldsymbol{\theta}_{T+\Delta-1} + \Big((\tilde{w}\otimes\tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Big)\tilde{\boldsymbol{\sigma}}_T Z_T\Bigg) \mid \mathbf{K}_{T-1} = \mathbf{k}\Bigg]. \tag{EC.19}$$

All almost sure equalities that follow in this proof are implicitly conditioned on $\mathbf{K}_{T-1} = \mathbf{k}$. Because $\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)$ is a possible value for treatment $\tilde{w}$, for any given set of realizations of the random variables the expression inside the expectation is non-negative and, because by assumption $Q_{T-1}(\mathbf{k}, f) - G(\mathbf{k}) = 0$, the expression inside the expectation needs to be almost surely zero. Because Z_T has infinite support, we obtain that

$$\Big((\tilde{w}\otimes\tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Big)\tilde{\boldsymbol{\sigma}}(\Sigma_{T+\Delta-1}, f(\mathbf{X}_T), \mathbf{X}_T) = \mathbf{0} \text{ a.s. } \forall \tilde{w}\in\mathscr{W},\ \forall f\in\mathbf{f}.$$

Using the definition of $\tilde{\boldsymbol{\sigma}}$ in (EC.5) and replacing $\tilde{w}$ with $\hat{f}(\tilde{\mathbf{X}}_1)$ for some treatment strategy $\hat{f}$, we obtain

$$\Big((\hat{f}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Big)\Sigma_{T+\Delta-1}(f(\mathbf{X}_T)\otimes\mathbf{X}_T)^\top = 0 \text{ a.s. } \forall f, \hat{f}\in\mathbf{f}.$$

By Assumption 2, we can replace $\mathbf{X}_T$ by $\tilde{\mathbf{X}}_1$ and maintain almost sure equality:

$$(\hat{f}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)^\top = (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)^\top \text{ a.s. } \forall f, \hat{f}\in\mathbf{f},$$

which can be rewritten as

$$(i\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(j\otimes\tilde{\mathbf{X}}_1)^\top = (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(j\otimes\tilde{\mathbf{X}}_1)^\top \text{ a.s. } \forall i, j\in\mathscr{W}. \tag{EC.20}$$

This implies that for any $i, j\in\mathscr{W}$,

$$\begin{aligned}\mathrm{Cov}((i\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu}, (j\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1) &= (i\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(j\otimes\tilde{\mathbf{X}}_1)^\top \\ &= (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1)\otimes\tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(j\otimes\tilde{\mathbf{X}}_1)^\top \\ &= C_{T+\Delta-1}(\tilde{\mathbf{X}}_1),\end{aligned} \tag{EC.21}$$

where the second step uses (EC.20), and where $C_{T+\Delta-1}$ is a function that does not depend on $i, j\in\mathscr{W}$. To see why this holds, notice that for any i, we obtain the same covariance.

Moreover, we know that $(i \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(j \otimes \tilde{\mathbf{X}}_1)^\top = (j \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(i \otimes \tilde{\mathbf{X}}_1)^\top$, so we can repeat the same argument to show that we also obtain the same covariance for any $j \in \mathscr{W}$.

Fix a $j \in \mathscr{W}$ and consider the random vector $\boldsymbol{s} \in \mathbb{R}^n$ with components

$$s_i = (i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} + (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} - (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1},$$

where $i = 1, \ldots, n$. Conditional on $\mathbf{K}_{T+\Delta-1}$ and $\tilde{\mathbf{X}}_1$, the vector $\boldsymbol{s}$ is normally distributed with conditional mean $(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1}$ and conditional covariance matrix with all entries equal to $\mathbb{V}((j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1) = C_{T+\Delta-1}(\tilde{\mathbf{X}}_1)$. Conditional on $\mathbf{K}_{T+\Delta-1}$ and $\tilde{\mathbf{X}}_1$, the vector $((i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu})_{i=1,\ldots,n}$ is also normally distributed with matching conditional mean and covariance matrix, and therefore $\boldsymbol{s}$ and $((i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu})_{i=1,2,\ldots,n}$ have matching distributions. So, we have

$$\begin{aligned}
U(\boldsymbol{\theta}, \Sigma) &= \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \Sigma_0 = \Sigma] \\
&= \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})] \\
&= \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{i} s_i \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&\qquad + \mathbb{E}\left[\mathbb{E}[(j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} - (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}\right] \\
&= \mathbb{E}[\mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1} \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= G(\mathbf{k}),
\end{aligned} \tag{EC.22}$$

where the first step is the definition of U; the second step follows because U does not depend on the pipeline, and because U depends on the knowledge state and not on the time index of the knowledge state; the third step is because $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})$ by assumption; the fourth step uses the tower property of expectations; the fifth step uses that $\boldsymbol{s}$ and $((i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu})_{i=1,2,\ldots,n}$ have the same distribution; the seventh step follows because the conditional expectations cancel out; the eighth step uses the tower property of expectations; the last step uses (EC.18).

We now show that $\tilde{G}(\boldsymbol{\theta}) = U(\boldsymbol{\theta}, \Sigma)$. For any $i, j \in \mathscr{W}$ and some function C_{T-1} that does not depend on $i, j \in \mathscr{W}$, we obtain

$$\begin{aligned}\mathrm{Cov}((i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}, (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1) &= (i \otimes \tilde{\mathbf{X}}_1)\Sigma_{T-1}(j \otimes \tilde{\mathbf{X}}_1)^\top \\ &= C_{T-1}(\tilde{\mathbf{X}}_1).\end{aligned}$$

We next justify the second equality. Consider the covariance matrix of the expected reward $(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ for all $i = 1, 2, \ldots, n$, given knowledge state $\mathbf{K}_t$: $\mathbf{C}_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top$, for $\Delta + 1 \le t \le T + \Delta$, where $\mathscr{W} \otimes \tilde{\mathbf{X}}_1$ is defined by (EC.1) and is a $n \times |\Xi|$-dimensional matrix. We proceed to show by induction that $\mathbf{C}_t(\tilde{\mathbf{X}}_1) = C_t(\tilde{\mathbf{X}}_1)\mathbf{L}_n$ for some function C_t, for $\Delta + 1 \le t \le T + \Delta - 1$, where $\mathbf{L}_n$ is the square matrix of dimension $n \times n$ with all entries equal to one. For the base case, by (EC.21), we know that $\mathbf{C}_{T+\Delta-1}(\tilde{\mathbf{X}}_1) = C_{T+\Delta-1}(\tilde{\mathbf{X}}_1)\mathbf{L}_n$. For the inductive step, suppose that $\mathbf{C}_t(\tilde{\mathbf{X}}_1) = C_t(\tilde{\mathbf{X}}_1)\mathbf{L}_n$ for some function C_t, for $\Delta + 2 \le t \le T + \Delta - 1$. Then, using the Bayesian updating equations[9] for Σ_t, we obtain

$$\begin{aligned}\mathbf{C}_{t-1}(\tilde{\mathbf{X}}_1) &= (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_{t-1}(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top \\ &= (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\left(\Sigma_t + \frac{\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t}{\sigma^2 - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}\right)(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top \\ &= (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top + \frac{(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top}{\sigma^2 - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}.\end{aligned}$$

Given the assumption that $\mathbf{C}_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top = C_t(\tilde{\mathbf{X}}_1)\mathbf{L}_n$ for some function C_t, the matrix $(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t^{1/2}$, where $\Sigma_t^{1/2}$ is the symmetric positive semi-definite square root of $\Sigma_t = \Sigma_t^{1/2}\Sigma_t^{1/2}$, can be shown to be a matrix in which every row is the same. Therefore, the second summand in the equation above is a matrix multiplication where the first factor $(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t^{1/2}$ is a $n \times |\Xi|$ matrix with all rows the same, and the last factor is its transpose, $\Sigma_t^{1/2}(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)^\top$, a $|\Xi| \times n$ matrix with all columns the same. As a result, the second summand is overall a matrix with all entries equal.[10] Thus, $\mathbf{C}_{t-1}(\tilde{\mathbf{X}}_1)$ is the sum of two matrices, each one of which has all entries equal. That is, $\mathbf{C}_{t-1}(\tilde{\mathbf{X}}_1) = C_{t-1}(\tilde{\mathbf{X}}_1)\mathbf{L}_n$ for some function C_{t-1}, and the induction is complete. It follows that $\mathbf{C}_{T-1}(\tilde{\mathbf{X}}_1) = C_{T-1}(\tilde{\mathbf{X}}_1)\mathbf{L}_n$ for some function C_{T-1}.

[9] (6b) solves for Σ_t as a function of Σ_{t-1}. To express Σ_{t-1} as a function of Σ_t, we solve the updating equation (EC.2b) for Σ_{t_1} and then apply the Sherman-Morrison-Woodbury matrix identity.

[10] Let $\mathbf{a} = [a_1 \ a_2 \ \ldots \ a_{|\Xi|}]^\top$ be a column vector and $\mathbf{M} = [\mathbf{m}_1 \ \mathbf{m}_2 \ \ldots \ \mathbf{m}_{|\Xi|}]$ an arbitrary $|\Xi| \times |\Xi|$ square matrix where column vectors $\mathbf{a}, \mathbf{m}_1, \ldots, \mathbf{m}_{|\Xi|}$ all have dimensions $|\Xi| \times 1$. Then, $[\underbrace{\mathbf{a}\ \mathbf{a}\ \ldots\ \mathbf{a}}_{n \text{ times}}]^\top \mathbf{M} [\underbrace{\mathbf{a}\ \mathbf{a}\ \ldots\ \mathbf{a}}_{n \text{ times}}]$ is a matrix where all entries are the same and equal to $\sum_{i=1}^{|\Xi|} \mathbf{a}^\top \mathbf{m}_i a_i$.

The remaining steps are analogous to the argument above to show that $G(\mathbf{k}) = U(\boldsymbol{\theta}, \Sigma)$. For $i = 1, 2, \ldots, n$, let

$$\tilde{s}_i = (i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta} + (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} - (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta},$$

which, conditional on $\mathbf{K}_{T-1} = \mathbf{k}$ and $\tilde{\mathbf{X}}_1$, is normally distributed with conditional mean $(j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}$ and conditional covariance matrix with all entries equal to $\mathbb{V}((j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1) = C_{T-1}(\tilde{\mathbf{X}}_1)$. Finally,

$$\begin{aligned}
U(\boldsymbol{\theta}, \Sigma) &= \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0 = \boldsymbol{\theta}, \Sigma_0 = \Sigma] \\
&= \mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{\tilde{w}}(\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{i} \tilde{s}_i \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&\qquad + \mathbb{E}\left[\mathbb{E}[(j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} - (j \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}\right] \\
&= \mathbb{E}[\mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta} \mid \mathbf{K}_{T-1} = \mathbf{k}, \tilde{\mathbf{X}}_1] \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta} \mid \mathbf{K}_{T-1} = \mathbf{k}] \\
&= \mathbb{E}[\max_{i}(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}] \\
&= \tilde{G}(\boldsymbol{\theta}),
\end{aligned}$$

where the eighth step follows because $\boldsymbol{\theta}$ is fixed and therefore the only random variable inside the expectation is $\tilde{\mathbf{X}}_1$, which does not depend on $\mathbf{K}_{T-1}$; the ninth step is the definition of $\tilde{G}$ in (EC.12); and the rest of the steps mirror the steps for deriving (EC.22). □

Remark EC.1. While the optimal allocation policy converges to the oracle's value, this does not imply that all parameters $\boldsymbol{\mu}$ are learned perfectly. If the distribution $F_{\tilde{x}}$ has zero probability of observing certain types of patients, then it will not be necessary to learn all the parameters $\boldsymbol{\mu}$ in order to attain an optimal value.[11] To guarantee an optimal value, the proof of Lemma EC.2 requires that all types of patients that will be observed post-trial be also observed in the trial, mathematically stated as $F_{\tilde{x}}$ being absolutely continuous with respect to F_x (Assumption 2).

[11] Following Example 1, if type B patients are not observed in the post-trial patients, then there is no real value in learning coefficients $\mu_{1,1}$ and $\mu_{2,1}$, which are only used in the response model for type B patients.

COROLLARY EC.4. *If for some* $\boldsymbol{\theta}, \Sigma, f$ *there is a* $\mathbf{j}$ *such that* $Q_{T-1}(\boldsymbol{\theta}, \Sigma, \mathbf{j}, f) = G(\boldsymbol{\theta}, \Sigma, \mathbf{j})$, *then* $Q_{T-1}(\boldsymbol{\theta}, \Sigma, \mathbf{j}', f) = G(\boldsymbol{\theta}, \Sigma, \mathbf{j}')$ *for any* $\mathbf{j}'$.

Proof. Notice that the premise is similar to that of Lemma EC.2, where the only difference is that the premise in this corollary is only stated for a fixed f instead of all possible treatment strategies. Thus, the steps in the proof of Lemma EC.2 leading up to (EC.20) hold with the only difference that they do not hold for any $j \in \mathscr{W}$ but only for allocation f, so we can conclude:

$$(i \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top = (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \text{ a.s. } \forall i \in \mathscr{W}. \tag{EC.23}$$

In addition, the reverse argument also holds, i.e., if (EC.23) holds then $Q_{T-1}(\boldsymbol{\theta}, \Sigma, \mathbf{j}, f) = G(\boldsymbol{\theta}, \Sigma, \mathbf{j})$. To see this, notice that, when (EC.23) holds, then

$$\Big((\tilde{w} \otimes \tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\Big)\, \tilde{\boldsymbol{\sigma}}(\Sigma_{T+\Delta-1}, f(\mathbf{X}_T), \mathbf{X}_T) = \mathbf{0} \text{ a.s. } \forall \tilde{w} \in \mathscr{W},$$

and (EC.19) becomes

$$\begin{aligned} Q_{T-1}(\mathbf{k}, f) - G(\mathbf{k}) &= \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} \Big((\tilde{w} \otimes \tilde{\mathbf{X}}_1) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\Big)\, \boldsymbol{\theta}_{T+\Delta-1}\right] \\ &= \mathbb{E}\left[\max_{\tilde{w} \in \mathscr{W}} \Big((\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1}\Big) - (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1}\right] \\ &= 0, \end{aligned}$$

where the last equality follows by the definition of $\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}$.

Using the same logic we used to arrive to (EC.21) in the proof of Lemma EC.2, we obtain that for any $i \in \mathscr{W}$,

$$\begin{aligned} \mathrm{Cov}((i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}, (f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_{T+\Delta-1}, \tilde{\mathbf{X}}_1) &= (i \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \\ &= (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\Sigma_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \\ &= \tilde{C}_{T+\Delta-1}(\tilde{\mathbf{X}}_1), \end{aligned} \tag{EC.24}$$

where the function $\tilde{C}_{T+\Delta-1}(\cdot)$ does not depend on $i \in \mathscr{W}$. (The function $\tilde{C}_{T+\Delta-1}(\cdot)$ here is analogous to the function $C_{T+\Delta-1}(\cdot)$ in the proof of Lemma EC.2.)

For any $t = \Delta+1, \ldots, T+\Delta$, let $\tilde{\mathbf{C}}_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top$ be the n-dimensional column vector that represents the covariances between $(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ and $(f(\tilde{\mathbf{X}}_1) \otimes$

$\tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ for all $i = 1, 2, \ldots, n$, given knowledge state $\mathbf{K}_t$. We proceed to show by induction that $\tilde{\mathbf{C}}_t(\tilde{\mathbf{X}}_1) = \tilde{C}_t(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}_t$, for $\Delta + 1 \le t \le T + \Delta - 1$, where $\mathbf{1}_n$ is the n-dimensional column vector with all elements equal to one. For the base case, by (EC.24), we know that $\tilde{\mathbf{C}}_{T+\Delta-1}(\tilde{\mathbf{X}}_1) = \tilde{C}_{T+\Delta-1}(\tilde{\mathbf{X}}_1)\mathbf{1}_n$. For the inductive step, suppose that $\tilde{\mathbf{C}}_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top = \tilde{C}_t(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}_t$, for $\Delta + 2 \le t \le T + \Delta - 1$. Then, using the Bayesian updating equations for Σ_t that we used in the proof of Lemma EC.2, we obtain

$$\begin{aligned}
\tilde{\mathbf{C}}_{t-1}(\tilde{\mathbf{X}}_1) =& (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_{t-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \\
=& (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\left(\Sigma_t + \frac{\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t}{\sigma^2 - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}\right)(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \\
=& (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \\
&+ \frac{(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top}{\sigma^2 - (W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})\Sigma_t(W_{t-\Delta} \otimes \mathbf{X}_{t-\Delta})^\top}.
\end{aligned}$$

Given the assumption that $\tilde{\mathbf{C}}_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top = \tilde{C}_t(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}_t$, $(\mathscr{W} \otimes \tilde{\mathbf{X}})\Sigma_t$ is a matrix in which every row is the same. Therefore, the second summand in the equation above, which is the product of matrix $(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma_t$ and a $|\Xi| \times 1$ vector, is a vector with all its elements equal to each other. Thus, $\tilde{\mathbf{C}}_{t-1}(\tilde{\mathbf{X}}_1)$ is the sum of two vectors, each one of which has all elements equal. That is, $\tilde{\mathbf{C}}_{t-1}(\tilde{\mathbf{X}}_1) = \tilde{C}_{t-1}(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}_{t-1}$, and the induction is complete. It follows that $\tilde{\mathbf{C}}_{T-1}(\tilde{\mathbf{X}}_1) = \tilde{C}_{T-1}(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}_{T-1}$.

Suppose now that we change the pipeline to be any $\mathbf{j}'$ and let $\boldsymbol{\theta}'_t$ and Σ'_t be the posterior mean and covariance matrix, respectively, for $t = T, T+1, \ldots, T+\Delta-1$. For any $t = T-1, \ldots, T+\Delta-1$, let $\tilde{\mathbf{C}}'_t(\tilde{\mathbf{X}}_1) = (\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma'_t(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top$ be the n-dimensional column vector that represents the covariances between $(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ and $(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ for all $i = 1, 2, \ldots, n$, given knowledge state $(\boldsymbol{\theta}'_t, \Sigma'_t, \mathbf{j}')$ and patient covariates $\tilde{\mathbf{X}}_1$. Using the same induction argument as above, but now forward in time, we use as a base case that $\tilde{\mathbf{C}}'_{T-1}(\tilde{\mathbf{X}}_1) = \tilde{\mathbf{C}}_{T-1}(\tilde{\mathbf{X}}_1) = \tilde{C}_{T-1}(\tilde{\mathbf{X}}_1)\mathbf{1}_n$, and can show that $\tilde{\mathbf{C}}'_t(\tilde{\mathbf{X}}_1) = \tilde{C}'_t(\tilde{\mathbf{X}}_1)\mathbf{1}_n$ for some function $\tilde{C}'_t$, for $t = T-1, T, \ldots, T+\Delta-1$ (note that $\tilde{C}'_t(\tilde{\mathbf{X}}_1)$ is not necessarily equal to $\tilde{C}_t(\tilde{\mathbf{X}}_1)$). Therefore, we obtain that $(\mathscr{W} \otimes \tilde{\mathbf{X}}_1)\Sigma'_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top$ is a vector with all elements equal, and therefore that the covariances between $(i \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$ and $(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu}$,

given knowledge state $(\boldsymbol{\theta}'_{T+\Delta-1}, \Sigma'_{T+\Delta-1}, \mathbf{j}')$, are all equal to the same constant for all treatments $i \in \mathscr{W}$. This implies that

$$(i \otimes \tilde{\mathbf{X}}_1)\Sigma'_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top = (\tilde{f}_{\boldsymbol{\theta}_{T+\Delta-1}}(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)\Sigma'_{T+\Delta-1}(f(\tilde{\mathbf{X}}_1) \otimes \tilde{\mathbf{X}}_1)^\top \text{ a.s. } \forall i \in \mathscr{W},$$

which in turn implies that $Q_{T-1}(\boldsymbol{\theta}, \Sigma, \mathbf{j}', f) = G(\boldsymbol{\theta}, \Sigma, \mathbf{j}')$, where we write $\boldsymbol{\theta}, \Sigma$ and not $\boldsymbol{\theta}', \Sigma'$ because the posterior mean and covariance matrix part of the knowledge state at time $T-1$ is the same with either pipeline $\mathbf{j}$ and $\mathbf{j}'$. $\square$

A second lemma, Lemma EC.3, will also be used in the proof of Theorem 2 to argue, for the purposes of contradiction, that if an allocation policy that satisfies certain properties were not asymptotically optimal, then the posterior parameters $\boldsymbol{\theta}_t$ and Σ_t would not converge almost surely, contradicting Theorem 1. Let w as an argument to $Q_{T-1}(\mathbf{k}, w)$ denote a treatment strategy that assigns treatment w uniformly, regardless of the covariates. Lemma EC.3 shows that if there is a positive expected value of information from sampling the next patient to arrive with treatment w, before knowing the patient's covariates, then there is a nontrivial set S of patient covariates for which the posterior mean of the coefficients will change if w is selected as the next treatment. This lemma holds regardless of the allocation policy. In stating the lemma, we let $\tilde{\Sigma}(\mathbf{k})$ represent the updated posterior covariance matrix after observing the outcomes of all the patients in the pipeline state of $\mathbf{k}$. For instance, at time t, we use (EC.10) with $t_1 = t$ and $t_2 = t + \Delta$ to obtain:

$$\tilde{\Sigma}(\mathbf{K}_t) := \Sigma_{t+\Delta} = \Sigma_t - \tilde{\boldsymbol{\sigma}}(\Sigma_t, \mathbf{W}_{(t-\Delta'_t+1):(t)}, \mathbf{X}_{(t-\Delta'_t+1):(t)})\tilde{\boldsymbol{\sigma}}^\top(\Sigma_t, \mathbf{W}_{(t-\Delta'_t+1):(t)}, \mathbf{X}_{(t-\Delta'_t+1):(t)}).$$

LEMMA EC.3. *Assume $T \geq \Delta + 1$. If $\mathbf{k} \in \mathscr{K}_T$ satisfies $Q_{T-1}(\mathbf{k}, w) > G(\mathbf{k})$ for some $w \in \mathscr{W}$, then there exists some set $S \subseteq \mathscr{X}$ (S may depend on $\mathbf{k}$ and w) such that*

1. $\mathbb{P}(\mathbf{X}_t \in S) > 0$ *(for a generic $\mathbf{X}_t$, not conditioned on $\mathbf{K}_{T-1}$),*
2. $q_{T-1}(\mathbf{k}, \mathbf{x}, w) > G(\mathbf{k})$ *for $\mathbf{x} \in S$, and*
3. $||\tilde{\boldsymbol{\sigma}}(\tilde{\Sigma}(\mathbf{k}), w, \mathbf{x})||_2 > 0$ *for $\mathbf{x} \in S$.*

Proof. Suppose $Q_{T-1}(\mathbf{k}, w) = \mathbb{E}[q_{T-1}(\mathbf{k}, \mathbf{X}_T, w) \mid \mathbf{K}_{T-1} = \mathbf{k}] > V_T(\mathbf{k})$. Recall that $q_{T-1}(\mathbf{k}, \mathbf{x}, w) \geq V_T(\mathbf{k}) = G(\mathbf{k})$ for any $\mathbf{k}$, $\mathbf{x}$, and w by Proposition EC.2. It follows that there exists a set $S \subset \mathscr{X}$ such that $q_{T-1}(\mathbf{k}, \mathbf{x}, w) > G(\mathbf{k})$ for $\mathbf{x} \in S$ and $\mathbb{P}(\mathbf{X}_T \in S) > 0$.

We let $\mathbf{k} = (\boldsymbol{\theta}, \Sigma, \mathbf{j})$, and write $q_t(\mathbf{k}, \mathbf{x}, w)$ as in the proof of Proposition EC.2 and $G(\mathbf{k})$ as in the proof of Lemma EC.2, to obtain:

$$q_{T-1}(\mathbf{k}, \mathbf{x}, w) - G(\mathbf{k})$$

$$
\begin{aligned}
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}\left((\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta}\right) \mid \mathbf{K}_{T-1}=\mathbf{k}, \mathbf{X}_T=\mathbf{x}, W_T=w\right] - \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}\left((\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta-1}\right) \mid \mathbf{K}_{T-1}=\mathbf{k}\right] \\
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}\left((\tilde{w}\otimes\tilde{\mathbf{X}}_1)\left(\boldsymbol{\theta}+\tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}_{(T-\Delta):(T)}, \mathbf{X}_{(T-\Delta):(T)})\mathbf{Z}_{(T-\Delta):(T)}\right)\right) \mid \mathbf{K}_{T-1}=\mathbf{k}, \mathbf{X}_T=\mathbf{x}, W_T=w\right] \\
&\qquad - \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}\left((\tilde{w}\otimes\tilde{\mathbf{X}}_1)\left(\boldsymbol{\theta}+\tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}_{(T-\Delta):(T-1)}, \mathbf{X}_{(T-\Delta):(T-1)})\mathbf{Z}_{(T-\Delta):(T-1)}\right)\right) \mid \mathbf{K}_{T-1}=\mathbf{k}\right] \\
&= \mathbb{E}\Bigg[\max_{\tilde{w}\in\mathscr{W}}\Bigg((\tilde{w}\otimes\tilde{\mathbf{X}}_1)(\boldsymbol{\theta}+\tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}_{(T-\Delta):(T-1)}, \mathbf{X}_{(T-\Delta):(T-1)})\mathbf{Z}_{(T-\Delta):(T-1)} \\
&\qquad + \tilde{\boldsymbol{\sigma}}(\tilde{\Sigma}(\mathbf{k}), w, \mathbf{x})Z_T)\Bigg) \mid \mathbf{K}_{T-1}=\mathbf{k}, \mathbf{X}_T=\mathbf{x}, W_T=w\Bigg] \\
&\qquad - \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}\left((\tilde{w}\otimes\tilde{\mathbf{X}}_1)\left(\boldsymbol{\theta}+\tilde{\boldsymbol{\sigma}}(\Sigma, \mathbf{W}_{(T-\Delta):(T-1)}, \mathbf{X}_{(T-\Delta):(T-1)})\mathbf{Z}_{(T-\Delta):(T-1)}\right)\right) \mid \mathbf{K}_{T-1}=\mathbf{k}\right],
\end{aligned}
$$

where we have used (EC.9) to update $\boldsymbol{\theta}_{T+\Delta-1}$ and $\boldsymbol{\theta}_{T+\Delta}$. Suppose to the contrary of claim 3 that $\tilde{\boldsymbol{\sigma}}(\tilde{\Sigma}(\mathbf{k}), w, \mathbf{x})$ is the zero vector for $\mathbf{x}\in S$. Then, $q_{T-1}(\mathbf{k}, \mathbf{x}, w) - G(\mathbf{k}) = 0$ for $\mathbf{x}\in S$. This provides the necessary contradiction to prove claim 3. □

EC.3.3.3. Combining lemmas to prove Theorems 2 and 3. We now combine Lemmas EC.1, EC.2, and EC.3 to prove the consistency results. While the preceding lemmas and propositions hold for any allocation policy, or only the optimal allocation policy, Theorem 2 constrains the set of allocation policies to those that sample the treatment with the largest one-step look-ahead value with positive probability. The allocation policies introduced in Section 4 will be shown to satisfy this condition in the proof of Theorem 3.

We define $Q(\mathbf{k}, f) = Q_{T-1}(\mathbf{k}, f)$ (value of sampling one more patient with treatment strategy f before sampling stops), and $q(\mathbf{k}, \mathbf{x}, w) = q_{T-1}(\mathbf{k}, \mathbf{x}, w)$ (value of sampling one more patient with covariates $\mathbf{x}$ and treatment w before sampling stops) to clarify that the one-step look-ahead value only depends on the knowledge state and not the time. In fact, while the mathematical definition of q_{T-1} is specific for the time $T-1$, it is easy to see that the definition is equivalent for any other time, as long as the implementation decision is made after Δ time steps:

$$
\begin{aligned}
q(\mathbf{k}, \mathbf{x}, w) = q_{T-1}(\mathbf{k}, \mathbf{x}, w) :=& \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{T+\Delta} \mid \mathbf{K}_{T-1}=\mathbf{k}, \mathbf{X}_T=\mathbf{x}, W_T=w\right] \\
=& \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta} \mid \mathbf{K}_{t-1}=\mathbf{k}, \mathbf{X}_t=\mathbf{x}, W_t=w\right],
\end{aligned}
$$

for all $t = 1, 2, \ldots, T$. Also important in the proof is that the fEVI indices in our allocation policy can be expressed in terms of q. That is, we have that $\nu_t(\mathbf{x}, w) = q(\mathbf{K}_t, \mathbf{x}, w) - G(\mathbf{K}_t)$, which follows by (12) because $q(\mathbf{K}_t, \mathbf{x}, w) = \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta+1} \mid \mathbf{K}_t, \mathbf{X}_{t+1}=\mathbf{x}, W_{t+1}=w\right]$ and $G(\mathbf{K}_t) = \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta} \mid \mathbf{K}_t\right]$.

THEOREM 2 *Let allocation policy* π *be such that, for* $0 \le t \le T-1$, *if* $\nu_t(\mathbf{x},w) \ge \nu_t(\mathbf{x},v)\ \forall v \in \mathscr{W}$, *then there exists a* $\delta > 0$ *so that* $\mathbb{P}^\pi(W_{t+1} = w \mid \mathbf{K}_t, \mathbf{X}_{t+1} = \mathbf{x}) \ge \delta$. *Then,* π *is asymptotically optimal.*

Proof. For this proof, assume the probability measure $\mathbb{P}^\pi$ induced by the allocation policy π that satisfies the premise of the theorem. Here, we may separately handle the components of the knowledge state, so the input to value functions is given explicitly by $\boldsymbol{\theta}_t$, Σ_t, and $\mathbf{J}_t$, instead of only $\mathbf{K}_t = (\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t) \in \mathscr{K}_t$. As this is an asymptotic result as $T \to \infty$, we assume that $T \ge t \ge \Delta + 1$, and thus that $\mathbf{J}_t$ is a full pipeline of Δ patients.

For a given $\mathbf{K}_t = \mathbf{k}$, we have $\nu_t(\mathbf{x},w) = q(\mathbf{k},\mathbf{x},w) - G(\mathbf{k})$. Therefore, the premise of the theorem can be rewritten as follows: if $q(\mathbf{k},\mathbf{x},w) \ge q(\mathbf{k},\mathbf{x},v)\ \forall v \in \mathscr{W}$, then $\mathbb{P}^\pi(W_{t+1} = w \mid \mathbf{K}_t, \mathbf{X}_{t+1} = \mathbf{x}) \ge \delta > 0$.

Let $A \subset \mathscr{W}$ be a given nonempty subset of treatments. Let E_A be the event that there is positive expected value, asymptotically, in treating the next patient with treatments in A (i.e., $Q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, w) > G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for at least one $\mathbf{j}$ for $w \in A$) and that there is zero expected value, asymptotically, in treating the next patient with a treatment not in A (i.e., $Q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, w) = G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for at least one $\mathbf{j}$ and for $w \notin A$), where the $\mathbf{j}$ in these statements has Δ patients. Note also that, by Corollary EC.1, $Q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, w) \ge G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for any treatment w. Therefore, by Corollary EC.4, the above definition of E_A is equivalent to:

$$E_A = \{\omega \in \Omega : \forall \mathbf{j} :\ Q(\boldsymbol{\theta}^\pi_\infty(\omega), \Sigma^\pi_\infty(\omega), \mathbf{j}, w) > G(\boldsymbol{\theta}^\pi_\infty(\omega), \Sigma^\pi_\infty(\omega), \mathbf{j}))\ \forall w \in A \\ \text{and } Q(\boldsymbol{\theta}^\pi_\infty(\omega), \Sigma^\pi_\infty(\omega), \mathbf{j}, w) = G(\boldsymbol{\theta}^\pi_\infty(\omega), \Sigma^\pi_\infty(\omega), \mathbf{j})\ \forall w \notin A\}.$$

The events E_A are exhaustive, i.e, $\mathbb{P}(\bigcup_{A \in 2^{\mathscr{W}}} E_A) = 1$. We will show that $\mathbb{P}^\pi(E_A) = 0$ for any nonempty A and use Lemma EC.2 to complete the proof.

By Theorem 1, a random vector $\boldsymbol{\theta}^\pi_\infty$ and a random matrix Σ^π_∞ exist such that $\boldsymbol{\theta}_t$ converges to $\boldsymbol{\theta}^\pi_\infty$ and Σ_t converges to Σ^π_∞ almost surely. Thus, we restrict to the sample paths that converge: let $D = \{\omega \in \Omega : \lim_{t\to\infty} \boldsymbol{\theta}_t(\omega) = \boldsymbol{\theta}^\pi_\infty(\omega) \text{ and } \lim_{t\to\infty} \Sigma_t(\omega) = \Sigma^\pi_\infty(\omega)\}$. Because $\mathbb{P}^\pi(D) = 1$, we have that $\mathbb{P}^\pi(E_A) = \mathbb{P}^\pi(E_A \cap D)$.

Frazier et al. (2009) prove asymptotic optimality of their algorithm by showing that the event $E_A \cap D$ is empty. In our case, $E_A \cap D$ is not empty because the randomly observed

covariates may lead us to observe only a subset of values of covariates.[12] So, we prove the weaker result that its probability is zero.

Suppose, for purposes of contradiction, that $\mathbb{P}^\pi(E_A \cap D) > 0$. In the remainder of this proof we focus on sample paths ω in $E_A \cap D$. By Lemma EC.3, for any $\omega \in E_A \cap D$ and any $w \in A$, there is a set $S_w(\omega) \subset \mathscr{X}$ such that $\mathbb{P}(\{\omega' \in \Omega : \mathbf{X}_t(\omega') \in S_w(\omega)\}) > 0$ and $q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, w) > G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for $\mathbf{x} \in S_w(\omega)$ and all $\mathbf{j}$. Moreover, because $Q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, v) = G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for any $\omega \in E_A \cap D$ and $v \notin A$, we have that $q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, v) = G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})$ for $\mathbf{x} \in S_w(\omega)$ and all $\mathbf{j}$. By the continuity of q in $\boldsymbol{\theta}$ and Σ (Proposition EC.1), there exists a $\tilde{t}_1(\omega)$ such that, for all $t \geq \tilde{t}_1(\omega)$,

$$|q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{j}, \mathbf{x}, w) - q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, w)| < \frac{q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, w) - G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})}{2}, \tag{EC.25}$$

where we just apply the definition of continuity with $\varepsilon = \frac{q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, w) - G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})}{2}$. Similarly, there exists a $\tilde{t}_2(\omega)$ such that, for all $t \geq \tilde{t}_2(\omega)$,

$$\begin{aligned} |q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{j}, \mathbf{x}, v) - G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})| &= |q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{j}, \mathbf{x}, v) - q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, v)| \\ &< \frac{q(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j}, \mathbf{x}, w) - G(\boldsymbol{\theta}^\pi_\infty, \Sigma^\pi_\infty, \mathbf{j})}{2}. \end{aligned} \tag{EC.26}$$

It follows that there is a $\tilde{t}(\omega)$, equal to $\max(\tilde{t}_1(\omega), \tilde{t}_2(\omega))$, such that

$$q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{j}, \mathbf{x}, w) > q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{j}, \mathbf{x}, v) \text{ for } w \in A,\ v \notin A,\ \mathbf{x} \in S_w(\omega) \tag{EC.27}$$

for all $t \geq \tilde{t}(\omega)$ and any pipeline $\mathbf{j}$. For ease of notation, in the sequel we drop the ω from notations $S_w(\omega)$ and $\tilde{t}(\omega)$.

Let $f_{q,t}(\mathbf{x}) = \arg\max_{w \in \mathscr{W}} q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t, \mathbf{x}, w)$ be the set of treatments with the largest one-step look-ahead value. By assumption, allocation policy π samples each of the treatments in $f_{q,t}(\mathbf{X}_{t+1})$ with probability at least δ. Thus, when $\mathbf{X}_{t+1} \in S_w$ and $w \in f_{q,t}(\mathbf{X}_{t+1})$ for $t \geq \tilde{t}$ (an event with positive probability), treatment w is sampled with probability at least δ. In addition, for $t \geq \tilde{t}$, if $\mathbf{X}_{t+1} \in S_w$ for some $w \in A$, then $f_{q,t}(\mathbf{X}_{t+1}) \subseteq A$ because (EC.27) implies that $q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t, \mathbf{X}_{t+1}, w) > q(\boldsymbol{\theta}_t, \Sigma_t, \mathbf{J}_t, \mathbf{X}_{t+1}, v)$ for $v \notin A$, thus excluding all $v \notin A$ from $f_{q,t}(\mathbf{X}_{t+1})$. In other words, we can separate the so-called q-factors for treatments in A and out of A for all $\mathbf{j}$, for all sufficiently large t.

[12] Recall Example 1. Suppose that you know exactly how treatment 1 works for patient type A but you need lo learn how treatment 2 works on patient type B. If you only observe type A patients (an event in Ω), then you cannot learn enough about type B, and will only sample treatment 2, which is the only treatment providing any learning benefits.

For a given $\omega \in E_A \cap D$, let $S = S(\omega) = \bigcup_{w \in A} S_w(\omega)$. By Lemma EC.3 (item 1), S has measure strictly greater than 0. For $\omega \notin E_A \cap D$, let $S = S(\omega) = \emptyset$. Let $||\cdot||_F$ be the Frobenius matrix norm. By (EC.8),

$$||\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})\tilde{\boldsymbol{\sigma}}^\top(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})||_F = ||\Sigma_{t+\Delta+1} - \Sigma_{t+\Delta}||_F.$$

By Lemma EC.3 (item 3), if $\mathbf{X}_{t+1} \in S$ and $W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})$, there is a scalar $C = C(\omega) > 0$ such that the vector norm $||\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})||_2 \geq C > 0$ for $t \geq \tilde{t}$. In that case, by the definition of the Frobenius norm,

$$||\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})\tilde{\boldsymbol{\sigma}}^\top(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})||_F = ||\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})||_2^2 \geq C^2 > 0 \tag{EC.28}$$

for all $t \geq \tilde{t}$. We can make a similar argument for all $\omega \in E_A \cap D$. Thus, we find that

$$\begin{aligned}
&\mathbb{P}^\pi\left(\limsup_{t\to\infty}\left\{||\Sigma_{t+\Delta+1} - \Sigma_{t+\Delta}||_F \geq C^2\right\} \mid E_A \cap D\right) \\
&\quad = \mathbb{P}^\pi\left(\limsup_{t\to\infty}\left\{||\tilde{\boldsymbol{\sigma}}(\Sigma_{t+\Delta}, W_{t+1}, \mathbf{X}_{t+1})||_2 \geq C\right\} \mid E_A \cap D\right) \\
&\quad \geq \mathbb{P}^\pi\left(\limsup_{t\to\infty}\left\{\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})\right\} \mid E_A \cap D\right),
\end{aligned}$$

where the equality follows from (EC.28), and the inequality follows by the construction of C in terms of the event $\{\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})\}$ in the text before (EC.28). Because $S(\omega) = \emptyset$ for $\omega \notin E_A \cap D$, we have $\mathbb{P}(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1} \mid (E_A \cap D)^c) = 0$ and for all $t > \tilde{t}$

$$\begin{aligned}
&\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1}) \mid E_A \cap D) \\
&\quad = \frac{\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})) - \mathbb{P}^\pi(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1}) \mid (E_A \cap D)^c)\mathbb{P}^\pi((E_A \cap D)^c)}{\mathbb{P}^\pi(E_A \cap D)} \\
&\quad = \frac{\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1}))}{\mathbb{P}^\pi(E_A \cap D)} \\
&\quad = \frac{\mathbb{P}^\pi(W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})) \mid \mathbf{X}_{t+1} \in S)\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S)}{\mathbb{P}^\pi(E_A \cap D)} \\
&\quad \geq \frac{\delta\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S)}{\mathbb{P}^\pi(E_A \cap D)} \\
&\quad = \frac{\delta\left(\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S \mid E_A \cap D)\mathbb{P}^\pi(E_A \cap D) + \mathbb{P}^\pi(\mathbf{X}_{t+1} \in S \mid (E_A \cap D)^c)\mathbb{P}(E_A \cap D)^c\right)}{\mathbb{P}^\pi(E_A \cap D)} \\
&\quad = \frac{\delta\mathbb{P}^\pi(\mathbf{X}_{t+1} \in S \mid E_A \cap D)\mathbb{P}^\pi(E_A \cap D)}{\mathbb{P}^\pi(E_A \cap D)}
\end{aligned}$$

$$= \delta \mathbb{P}^{\pi}(\mathbf{X}_{t+1} \in S \mid E_A \cap D) \tag{EC.29}$$
$$> 0,$$

where the first equality is an application of the total probability theorem, the first inequality follows by the assumption on the policy π, and the last inequality follows because, for $\omega \in E_A \cap D$, $\{\mathbf{X}_{t+1} \in S\}$ is an event with positive probability for all t by Lemma EC.3. Because $\mathbf{X}_{t+1}$ are independent and identically distributed, the expression in (EC.29) is bounded below by a positive constant for any t, and we obtain that $\limsup_{t\to\infty} \mathbb{P}^{\pi}(\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1}) \mid E_A \cap D) > 0$. The probability of the limsup event is at least as large as the limsup of the probability (Billingsley 2008, Theorem 4.1), therefore

$$\mathbb{P}^{\pi}\left(\limsup_{t\to\infty} \{\mathbf{X}_{t+1} \in S, W_{t+1} \in f_{q,t}(\mathbf{X}_{t+1})\} \mid E_A \cap D\right) > 0.$$

However, the convergence of the sequence Σ_t in the event D implies that $\mathbb{P}^{\pi}\left(\limsup_{t\to\infty} \{||\Sigma_{t+\Delta+1} - \Sigma_{t+\Delta}||_F \geq C^2\} \mid E_A \cap D\right) = 0$. This contradicts the result of the preceding displayed equations, that $\mathbb{P}^{\pi}(\limsup_{t\to\infty} \{||\Sigma_{t+\Delta+1} - \Sigma_{t+\Delta}||_F \geq C^2\} \mid E_A \cap D) > 0$. Thus, we conclude that $\mathbb{P}^{\pi}(E_A \cap D) = 0$, and therefore that $\mathbb{P}^{\pi}(E_A) = 0$.

The above argument holds for an arbitrary nonempty subset of treatments A, so we conclude that $\mathbb{P}^{\pi}(E_A) = 0$ for any nonempty A and $\mathbb{P}^{\pi}(E_{\emptyset}) = 1$, i.e., for all $\mathbf{j}$ and for all $w \in \mathscr{W}$, $Q(\boldsymbol{\theta}^{\pi}_{\infty}, \Sigma^{\pi}_{\infty}, \mathbf{j}, w) = G(\boldsymbol{\theta}^{\pi}_{\infty}, \Sigma^{\pi}_{\infty}, \mathbf{j})$ almost surely. By Lemma EC.2, we get

$$\tilde{G}(\boldsymbol{\theta}^{\pi}_{\infty}) = U(\boldsymbol{\theta}^{\pi}_{\infty}, \Sigma^{\pi}_{\infty}) \ a.s. \tag{EC.30}$$

We can now write

$$\begin{aligned}
V^{\pi}(\mathbf{K}_0; \infty) &= \mathbb{E}[\tilde{G}(\boldsymbol{\theta}^{\pi}_{\infty}) \mid \mathbf{K}_0] \\
&= \mathbb{E}[U(\boldsymbol{\theta}^{\pi}_{\infty}, \Sigma^{\pi}_{\infty}) \mid \mathbf{K}_0] \\
&= \mathbb{E}\left[\mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}^{\pi}_{\infty}, \Sigma^{\pi}_{\infty}\right] \,\middle|\, \mathbf{K}_0\right] \\
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \mathbf{K}_0\right] \\
&= \mathbb{E}\left[\max_{\tilde{w}\in\mathscr{W}} (\tilde{w} \otimes \tilde{\mathbf{X}}_1)\boldsymbol{\mu} \mid \boldsymbol{\theta}_0, \Sigma_0\right] \\
&= U(\boldsymbol{\theta}_0, \Sigma_0).
\end{aligned}$$

Here, the first equality is by Corollary EC.3, the second equality is by (EC.30), the third equality is by the definition of U, the fourth equality is by the tower property of expectations, the fifth equality follows because the empty pipeline of $\mathbf{K}_0$ does not affect the expectation, and the sixth equality is by the definition of $U(\boldsymbol{\theta}_0, \Sigma_0)$. $\square$

Finally, we show that each of the allocation policies fEVI, fEVI-rand, and fEVI-MC-rand satisfy the premise of Theorem 2, justifying their asymptotic optimality.

THEOREM 3. *Allocation policies* fEVI*,* fEVI-rand*, and* fEVI-MC-rand *are asymptotically optimal.*

Proof. We show that the allocation policies satisfy the premise of Theorem 2, which will complete the proof. That is, we show that the allocation policies sample the treatment w with the largest $\nu_t(\mathbf{X}_{t+1}, w)$ with positive probability.

Allocation policy fEVI is defined to sample a treatment w with the largest index $\nu_t(\mathbf{X}_{t+1}, w)$. Considering that ties for the largest index are broken uniformly at random, and that there are n treatments, choosing $\delta = 1/n$ allows the hypothesis of Theorem 2 to be satisfied.

Allocation policies fEVI-rand and fEVI-MC-rand potentially sample a treatment with largest index in two ways. They sample a treatment prescribed by the fEVI and fEVI-MC allocation policies, respectively, with probability $1-\epsilon > 0$. Moreover, they sample uniformly at random with probability $\epsilon > 0$. Thus, selecting $\delta = (1-\epsilon)/n + \epsilon/n = 1/n$ allows the hypothesis of Theorem 2 to be satisfied. $\square$

REMARK EC.2. The use of Theorem 2 to prove the asymptotic optimality of the fEVI-MC allocation policy is less obvious. There is a nonstationarity in the fEVI-indices being estimated, and it may be hard to prove that the index with highest value is sampled with a probability bounded below by a specified fixed $\delta > 0$ for all t. However, we observe that the fEVI-MC allocation policy gives an optimal index to maximize the *conditional* expected reward of sample information, for the special case of $\eta^{\text{on}} = \eta^{\text{off}} = 1$, conditional on the sampled posterior at time $\boldsymbol{\theta}_{t+\Delta}$, the patients in pipeline given $\mathbf{K}_t$, and the covariates $\mathbf{X}_{t+1}$ of the next patient to assign to an arm. This follows because the naive MC estimator of the fEVI-index in (12), conditional on these quantities, becomes a correlated knowledge gradient (cKG) calculation – the only remaining uncertainty is the univariate outcome

Y_{t+1}, conditional on the other random quantities. Frazier et al. (2009) has shown the one-step and asymptotic optimality of cKG. Our fEVI-MC indices are MC averages of such (conditionally) optimal cKG indices, which explains their usefulness in numerical examples.

EC.4. Comparator Policies in Numerical Experiments

We describe in detail the TS, TTTS, BATTLE*, and BC* policies used in Section 7.

Thompson sampling (TS). Treatments are allocated to a patient with given covariate vector $\mathbf{X}_{t+1}$ according to the probability each treatment is the best. This is implemented by drawing a sample from the posterior mean reward for each treatment, and selecting the treatment with the largest expected outcome given the sampled mean, as is often done in practice (Russo 2020).

Top-two Thompson sampling (TTTS). Russo (2020) defines this adaptation of Thompson sampling as follows. Draw a sample of the posterior mean $\boldsymbol{\theta}_{t+1} \mid \mathbf{K}_t \sim \mathcal{N}(\boldsymbol{\theta}_t, \Sigma_t)$. Compute the expected outcomes of each treatment $\gamma_w = (w \otimes \mathbf{X}_{t+1})\boldsymbol{\theta}_{t+1}$, where $\boldsymbol{\theta}_{t+1}$ here represents the drawn sample. With probability β allocate the treatment $I = \arg\max_w \gamma_w$ (using the Thompson sampling policy). With probability $1-\beta$, we continue drawing samples of the posterior mean and finding the treatment with the largest expected outcome until we get a draw in which the treatment with the largest expected outcome is not equal to I, and allocate that treatment. We limit the number of draws to 100 to limit the time the policy takes to make an allocation. We use $\beta = 0.5$ following the recommendation of Russo (2020).

BATTLE.* The BATTLE trial policy (Zhou et al. 2008) was designed for 0-1 outcomes, i.e., $Y_t = 1$ if the treatment was successful and $Y_t = 0$ if the treatment failed. It allocates a treatment with probability proportional to the current probability that the treatment will be successful, among treatments whose probability exceeds a threshold.

Because our model considers continuous outcomes ($Y_t \in \mathbb{R}$), we use the following BATTLE* procedure to mimic the spirit of the BATTLE trial policy:

1. Compute the expected outcome for the current patient for each treatment: $\gamma_w = (w \otimes \mathbf{X}_{t+1})\boldsymbol{\theta}_t$ for $w \in \mathscr{W}$.
2. Compute the sample mean μ_γ and sample standard deviation σ_γ of the expected outcomes γ_w for $w \in \mathscr{W}$.
3. Compute a lower threshold $\gamma_{lt} = \mu_\gamma - z_\alpha \sigma_\gamma$. We use $z_\alpha = 2$.
4. Allocate each treatment with probability proportional to $\max\{0, \gamma_w - \gamma_{lt}\}$.

Because treatments with low expected outcomes will have a γ_w below γ_{lt}, their probability of assignment is zero, effectively dropping those treatments from the choice for this specific time t. Therefore, similar to the BATTLE trial policy, BATTLE* allocates with probabilities proportional to the expected outcomes[13] (shifted by a lower threshold) and drops treatments that are unlikely to be best for the next treatment allocation.

Biased coin (BC).* Pocock and Simon (1975) describe a general algorithm to balance prognostic covariates across treatment arms. The biased coin scheme of Pocock and Simon (1975) requires a distance function (for calculating distance between arms with respect to a covariate value), an aggregating function (for aggregating distances across dimensions of covariates), and probabilities of assigning treatments given the resulting scores. In this section, we describe our choices for those.

We split patients into groups with respect to their covariates, where each group corresponds to a unique combination of values for the covariates, and where the covariates are assumed to be discrete. We index the patient groups by $l = 1, 2, \ldots, m_{pred}$. If there are no predictive covariates, then $m_{pred} = 1$ and all patients fall within the same group. Suppose also that we have m_{prog} prognostic covariates and covariate k, with $k = 1, 2, \ldots, m_{prog}$, can take M_k distinct values, which we refer to as levels. We also include a covariate with index $k = 0$ and $M_0 = 1$ to represent the intercept term, a covariate with only one level that all patients share. Let $x_{l,k,j,i}$ represent the number of patients in group $l \in \{1, 2, \ldots, m_{pred}\}$ with prognostic covariate $k \in \{0, 1, \ldots, m_{prog}\}$ at level $j \in \{1, 2, \ldots, M_k\}$ that have been assigned to receive treatment $i \in \mathscr{W}$.

When a new patient is enrolled, let $x^w_{l,k,j,i}$ be the counts that would be obtained if the patient was allocated to treatment $w \in \mathscr{W}$. (These counts depend on the time t, but we suppress t to avoid further indices in subscripts.) For each prognostic covariate $k = 1, 2, \ldots, m_{prog}$, we can compute a *distance* $d_{k,w}$ that captures the imbalance of the allocation of covariate k into the treatment arms if we were to allocate this patient to treatment arm w. Pocock and Simon (1975) propose several distance functions and we use the sample variance (their first suggestion):

$$d_{k,w} = \left(\sum_{i \in \mathscr{W}} \left(x^w_{\hat{l},k,\hat{j},i} - \sum_{i \in \mathscr{W}} x^w_{\hat{l},k,\hat{j},i}/n \right)^2 \right) /(n-1),$$

[13] The expected outcome of a Bernoulli 0-1 variable is the probability of success.

where $\hat{l}$ and $\hat{j}$ are, respectively, the group and the level of the covariate k for the newly arrived patient.

After obtaining the distance for each prognostic covariate, we need to aggregate the distances into a *score* $\mathscr{G}_w$ that represents the total amount of imbalance. We follow the recommendation of Pocock and Simon (1975) to use a weighted sum with weights u_k: $\mathscr{G}_w = \sum_{k=0}^{m_{prog}} u_k d_{k,w}$. We let $u_0 = 0$ and $u_k = 1$ for $k = 1, 2, \ldots, m_{prog}$, unless otherwise specified. Setting $u_0 = 0$ balances prognostic covariates. Setting $u_0 > 0$ not only balances prognostic covariates across treatment arms, but also balances treatment arms for the given group of the new patient to enroll.

Finally, we sort the treatments from lowest to highest score (breaking ties at random) and sample each with probability p_i for $i = 1, 2, \ldots, n$, where p_1 is the probability of assigning the treatment with lowest score and p_n is the probability of assigning the treatment with highest score. To promote a reduction in imbalance, i.e., treatments with lower score, p_i should be non-increasing. We set $p_1 = 0.5$ and $p_i = 0.5/(n-1)$ for $i = 2, 3, \ldots, n$ following one of the recommendations of Pocock and Simon (1975).

Thus, our BC* algorithm is an extension of the biased coin randomization scheme in Pocock and Simon (1975). If $m_{pred} = 1$ and $u_0 = 0$, our BC* allocation policy is equivalent to the biased coin scheme of Pocock and Simon (1975). Moreover, we allow for some levels of covariates to be predictive and some levels to be prognostic. Importantly, the BC* requires a given covariate to be either predictive with respect to all treatments, or prognostic, or neither. It is not possible, for purposes of allocation to treatments, for BC* to allow a covariate to both be prognostic and predictive. This is different from all the other allocation policies we consider. This has implications for comparing BC* with the other comparator allocation policies in Section 7.

Consider our numerical example in Section 7.1.1. There, we have four Mars endotypes. Because BC* does not allow for a covariate to be both prognostic as well as predictive with respect to one or more treatments, we split the Mars endotyping information into two covariates for the purposes of assessing BC* (otherwise, we could not formally apply BC* if Mars3 were both prognostic and predictive with respect to aggressive fluids management). The first is a binary covariate indicating whether the patient is Mars3 or not. The second is a covariate with four levels indicating whether the patient is Mars1, 2, 4, or not. We label the first covariate as predictive and the second one as prognostic. The caveat of this

approach is that patients in the Mars3 endotype group will only have one possible level in the second covariate, namely none. Therefore, this BC* scheme is promoting the balance of the treatment arms for Mars3 patients and it is promoting the balance of all other endotypes otherwise. This covariate splitting is used for the experiments of Figures 2a, 2b, and 3c.

We note that the covariate splitting proposed here for allocation purposes is one possible splitting that is compatible with the setting in Section 7.1.1, while other compatible splittings exist. Although we show results for only one possible covariate splitting in each experiment, the performance of BC* under various compatible covariate splittings used for allocation is similar for large enough sample size in the experiments of Section 7.2 (data not shown). We also note that for implementation purposes, BC* in our experiments uses the assumed regression model defined by the specified labeling, similarly to all the other allocation policies considered.

For the experiment of Figure 2c, which explores a mislabeling where there is no predictive covariate, BC* uses a covariate splitting where all patients belong to the same patient group ($m_{pred} = 1$) and the Mars endotypes, the APACHE score, and the idle real-valued covariate are prognostic.

EC.5. Additional Simulation Results

Here we provide additional experiments to complement the results presented in Section 7. Appendix EC.5.1 presents graphs for the PICS performance metric for the experiments in Section 7 that reported the EOC performance metric. Appendix EC.5.2 presents some additional examples of the performance of our proposed fEVI family of allocations in a setting where the labeling of potentially active covariates is incorrectly specified. Appendix EC.5.3 presents another Monte Carlo algorithm for estimating fEVI indices, and provides empirical support to the assertion that our chosen fEVI-MC algorithm in the main paper can compute indices with orders of magnitude improvement by making use of a variance reduction enabled by results of Frazier et al. (2009). Finally, Appendix EC.5.4 presents a numerical experiment that suggests that performance of fEVI-MC-rand is what one would expect from combining the Monte Carlo estimation of fEVI indices in fEVI-MC together with the randomization of fEVI-rand.

EC.5.1. PICS for experiments in Section 7

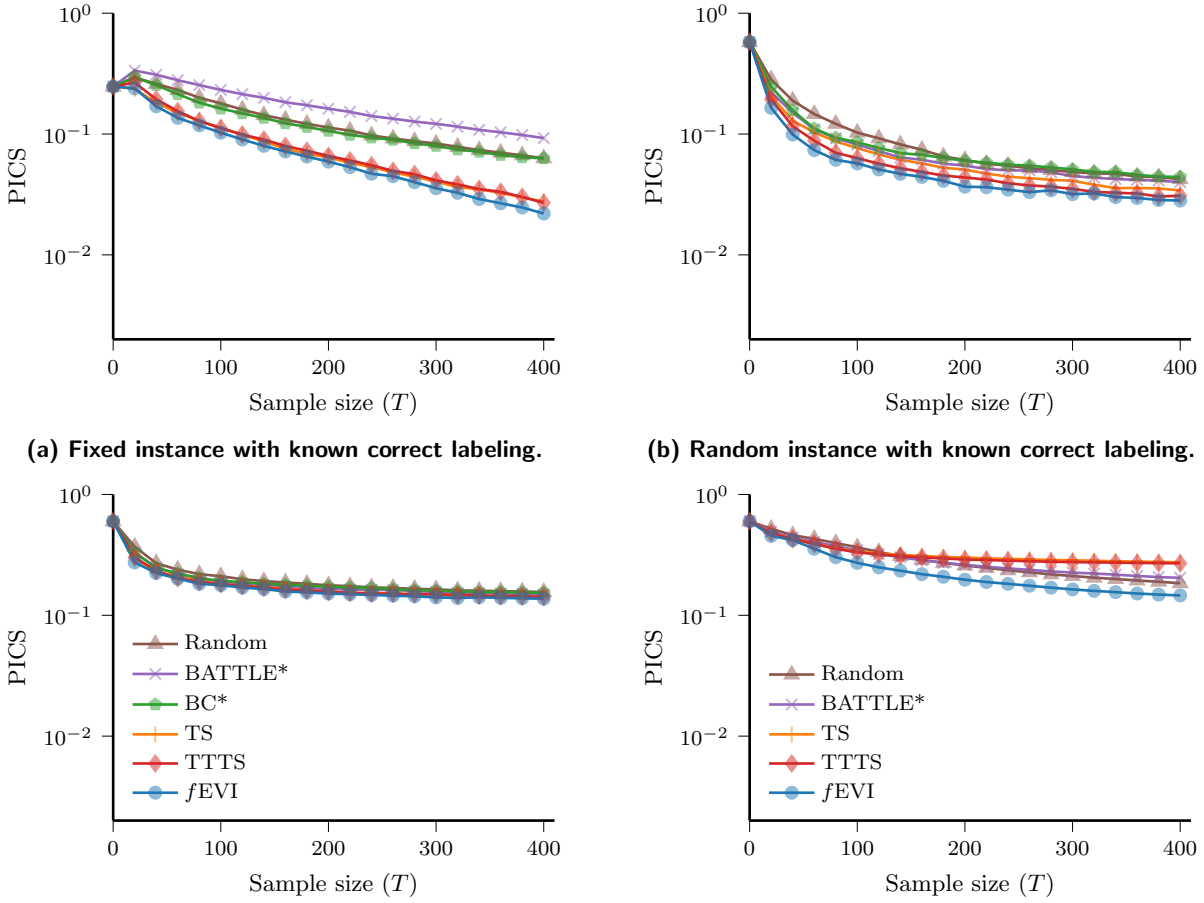
The newly proposed allocation policies and comparator allocation policies presented in the main paper attempt to optimize regret, or EOC. It is also of interest to assess their performance relative to PICS, the probability that a randomly selected patient from the post-trial population does not receive a treatment with the best mean outcome for that patient. EOC weights particularly bad choices more than moderately poorer choices, whereas PICS measures the probability, averaged over all patients, that the true best alternative is not implemented. In this section, we present performance results analogous to those in Section 7, which were measured relative to EOC, for the PICS performance measure.

Figure EC.1 presents results for PICS that correspond to the results in Figure 2 for EOC that are in Section 7.2 and Section 7.3. The graphs have the same qualitative features for the relative ranking of the curves for PICS here as compared to those reported for EOC in the main paper. For example, for our given difficult fixed instance setting problem, that was motivated by the slippage configuration of R&S, we again see in Figure EC.1a that the slope on a log scale of the performance metric is roughly linear, and that fEVI dominates (compare with Figure 2a). For the random instance setting, the slope of the curve is not as steep, as this curve averages over a variety of random configurations, some of which have slower improvement rates for all allocation policies. This explains the apparent slowing of the slopes of the curves for all allocation policies in the random instance setting for PICS, as also observed for EOC (similar qualitative behaviors in Figure 2b and Figure EC.1b).

Figure EC.2 presents results for PICS that correspond to the results in Figure 3 in Section 7.4. We do see some modest changes in the magnitude of the relative performance. For example, Figure EC.2a shows that for performance of fEVI-MC with the smallest number of replications (e.g., $\eta^{\text{on}} = 1$) in the estimator for the fEVI indices, the performance is less severely degraded for PICS as compared to a modest number of replications (e.g., $\eta^{\text{on}} \geq 5$), in comparison with this difference for EOC in the main paper (Figure 3a). We see a similar less pronounced degradation for the PICS curve with randomization in Figure EC.2b as compared to the EOC curve with randomization in Figure 3b. These are questions of magnitude of difference for these graphs, however, rather than differences in the relative performance for the different parameter settings for these curves. As such, they do not shed substantive changes in the insights relative to those presented in the main paper. We see similar relative comparisons for the panels dealing with approximations for

handling continuous-values covariates (Figure EC.2c), and with the effect of delay and ignoring information about pipeline patients in making allocations (Figure EC.2d).

Figure EC.1 Probability of incorrect selection (PICS) for policies with different labelings.



(a) Fixed instance with known correct labeling.

(b) Random instance with known correct labeling.

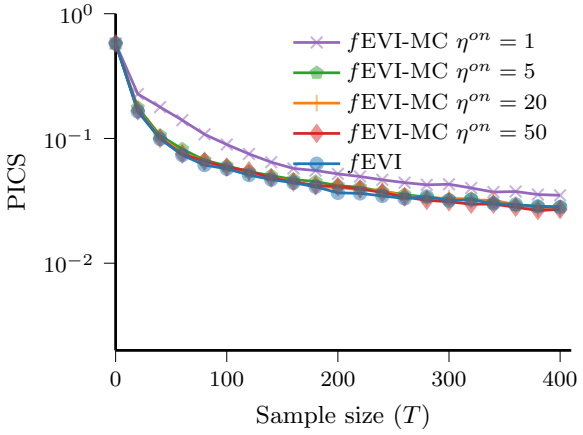
(c) Unknown personalization potential. Random instance.

(d) All covariates potentially predictive. Random instance.

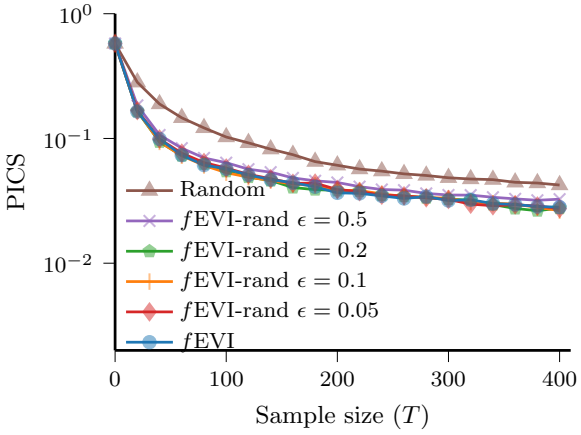
EC.5.2. Additional results on mislabeling predictive, prognostic, and idle covariates

Which treatment interacts with a covariate is unknown. Consider that the trial manager knows that Mars3 is the covariate that enables personalization, but is unaware that it only interacts with treatments with liberal fluid management, and assumes instead that Mars3 is potentially predictive with respect to all treatments (adding three coefficients). Moreover, consider that the trial manager is unaware that treatment 4 is the only treatment with an active treatment effect, and instead labels all treatment effects as potentially active (adding six coefficients, as we add $2^3 - 1$ predictive coefficients but we remove the prognostic coefficient for Mars3 to avoid overspecification). Figure EC.3a shows that this

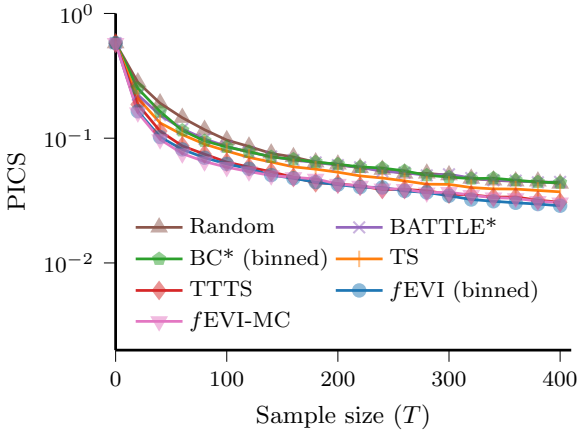
Figure EC.2 Effect of practical considerations on the performance of fEVI. The plots show the PICS of the random instance setting.



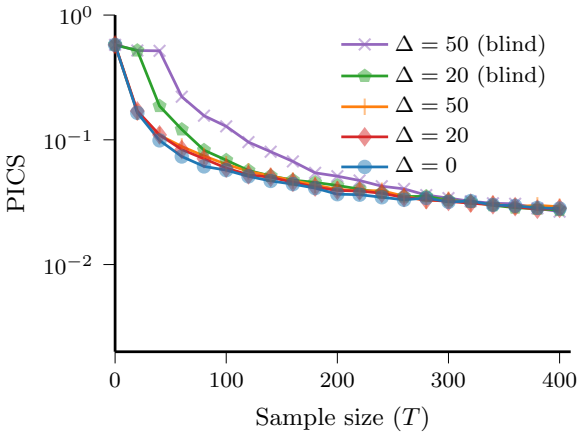
(a) fEVI-MC (with $\eta^{\text{off}} = 1$, several values of η^{on})



(b) fEVI-rand (with several values of ϵ)



(c) PICS when APACHE score is continuous-valued. Allocation policies fEVI and BC* bin the APACHE score into three levels. fEVI-MC uses $\eta^{\text{on}} = 20$, $\eta^{\text{off}} = 1$.



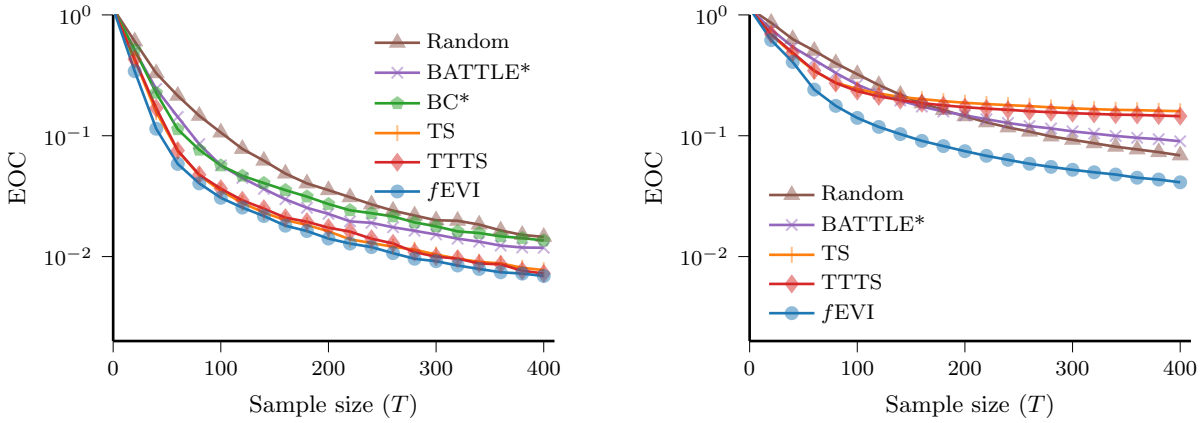
(d) Effect of delay Δ on the PICS when using the fEVI-MC and fEVIblind allocation policies.

mislabeling slows inference down (there are 20 instead of 11 coefficients to learn). This can be seen because the EOC of a given allocation policy is larger in Figure EC.3a compared to the known labeling case in Figure 2b. This labeling still allows for full personalization (the EOC curves do not hit a nonzero asymptote). The performance ranking of the policies remains, with fEVI having the lowest (best) EOC and Random the highest (worst).

Taking all covariates but the idle to be potentially predictive. We consider a different mislabeling, where the trial manager labels all Mars endotypes and APACHE as potentially predictive with respect to all treatments, with no prognostic covariate (40 instead of 11 coefficients), but does not mislabel the idle covariate. Figure EC.3b shows that inference is slowed down compared to knowing the correct labeling (Figure 2b), but less so than in the

setting where all of the Mars endotypes, APACHE, *and* the idle covariate are mislabeled as predictive with respect to all treatments (Figure 2d).

Figure EC.3 Expected opportunity cost (EOC) for policies with different labelings.



(a) Unknown interaction with treatments. Random instance.

(b) Mars endotypes and APACHE potentially predictive. Random instance.

EC.5.3. Simple Monte Carlo estimation of the fEVI policy.

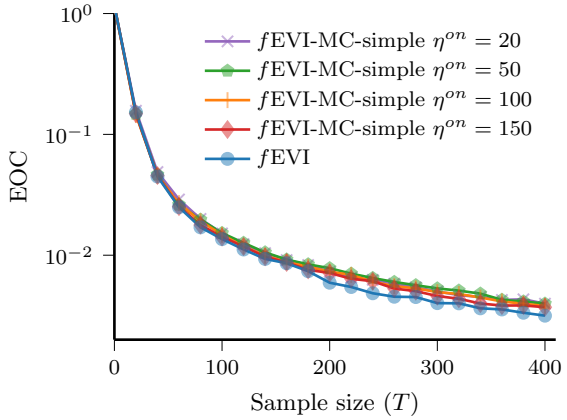
To show that the fEVI-MC policy performs better than a naive Monte Carlo estimate of the fEVI indices, we explore the performance of the fEVI-MC-simple allocation policy. It uses a *simple* Monte Carlo estimation of the fEVI indices as described in Algorithm 2. It is simple in that it draws random samples of the outcome of the current patient instead of using the closed-form computation implemented by the fEVI-MC policy (to compute a conditional EVI in step 13 of Algorithm 1). However, fEVI-MC-simple uses common random numbers for posterior means to be observed when outcomes of pipeline patients are observed as a variance reduction technique.

Figure EC.4 shows the EOC of the fEVI-MC-simple policy for different values of η^{on}. We fixed $\eta^{\text{off}} = 50$ because there are $4 \times 3 = 12$ combinations of active covariates, and we expect to observe each combination at least once with high probability in each simulated set of $\eta^{\text{off}} = 50$ post-trial patients. While the EOC for small sample sizes is not significantly different to that of fEVI, for larger sample sizes ($T > 200$) we observe a small but statistically significant increase in EOC even for the largest η^{on}. Because the EVI decreases exponentially with sample size, it becomes more difficult to estimate the fEVI indices. While fEVI-MC uses a closed-form computation that can compute the logarithm of the indices, thus reducing the difficulty of estimating smaller indices, the performance of the

Algorithm 2 fEVI-MC-simple: Estimate fEVI indices with Monte Carlo estimates of posterior means

function fEVI-MC-simple($\mathbf{K}_t, \mathbf{X}_{t+1}; \eta^{\text{on}}, \eta^{\text{off}}$)
  Let $\Delta'_t = \min\{t, \Delta\}$ ▷ Compute number of patients in pipeline.
  for all $w \in \mathscr{W}$ **do** ▷ For each potential treatment for the next
    $W_{t+1} \leftarrow w$ ▷ patient, compute the preposterior
    $\hat{\boldsymbol{\sigma}}^{(w)} \leftarrow \tilde{\boldsymbol{\sigma}}(\Sigma_t, \mathbf{W}_{(t-\Delta'_t+1):(t+1)}, \mathbf{X}_{(t-\Delta'_t+1):(t+1)})$ ▷ std dev in (EC.5)
  end for
  for j in $\{1, \ldots, \eta^{\text{on}}\}$ **do** ▷ Compute offline rewards for η^{on} replications
    $\hat{\mathbf{X}}_{1:\eta^{\text{off}}} \overset{i.i.d.}{\sim} F_{\tilde{x}}$ ▷ Sample η^{off} post-trial covariates
    $\hat{\mathbf{Z}} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{\Delta'_t+1})$ ▷ Noise vector of length $\Delta'_t + 1$ (pipeline plus extra patient)
    for all $w \in \mathscr{W}$ **do** ▷ Compute offline rewards when sampling each treatment
      $\hat{\boldsymbol{\theta}}_j^{(w)} \leftarrow \boldsymbol{\theta}_t + \hat{\boldsymbol{\sigma}}^{(w)} \hat{\mathbf{Z}}$ ▷ Posterior mean
      $\hat{V}_j^{(w)} \leftarrow (1/\eta^{\text{off}}) \sum_{i=1}^{\eta^{\text{off}}} (\tilde{f}_{\hat{\boldsymbol{\theta}}_j^{(w)}}(\hat{\mathbf{X}}_i) \otimes \hat{\mathbf{X}}_i) \hat{\boldsymbol{\theta}}_j^{(w)}$ ▷ Offline rewards
    end for
  end for
  for all $w \in \mathscr{W}$ **do**
    $\hat{V}_{avg}^{(w)} \leftarrow (1/\eta^{\text{on}}) \sum_{i=1}^{\eta^{\text{on}}} \hat{V}_j^{(w)}$ ▷ Trial value estimated through average of offline rewards
  end for
  return $W_{t+1} = \arg\max_{w \in \mathscr{W}} \hat{V}_{avg}^{(w)}$ ▷ Pick highest estimated value (break ties at random)
end function

Figure EC.4 EOC of the fEVI-MC-simple policy for different values of η^{on}. We fix $\eta^{\text{off}} = 50$.

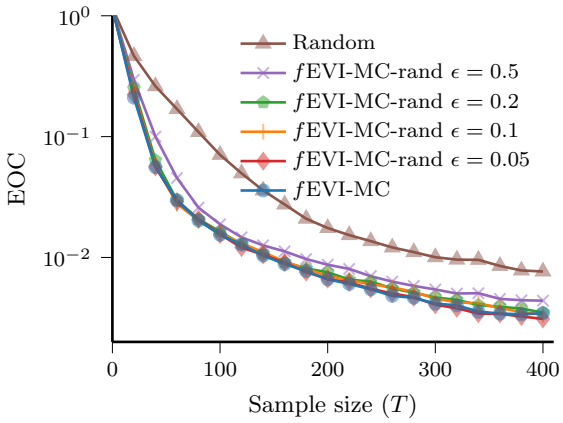


fEVI-MC-simple degrades with sample size despite the use of many more replications. Thus, we prefer the fEVI-MC allocation policy to the fEVI-MC-simple allocation policy, and have therefore used fEVI-MC in the main paper.

EC.5.4. fEVI-MC-rand

To complement the results of Section 7.4, Figure EC.5 shows the EOC when we combine Monte-Carlo simulation and randomization in the fEVI-MC-rand allocation policy. Similar to the results presented in Figure 3b, we do not observe a practically significant difference

Figure EC.5 EOC of the random instance setting for the fEVI-MC-rand for different values of ϵ. The fEVI-MC indices were estimated with $\eta^{\text{on}} = 5$ and $\eta^{\text{off}} = 1$.



between fEVI-MC and fEVI-MC-rand for $\epsilon \leq 0.2$. We only observe a practically significant difference when $\epsilon = 0.5$.

EC.6. Additional Conceptual and Practical Considerations

There are a number of interesting and related theoretical and practical considerations that may arise for clinical trials or for the contexts in which clinical trials are run. They may include a delay in observations, which is already in our base model, and the need for additional randomization, which is treated by our fEVI-rand and fEVI-MC-rand allocation policies. Additional considerations may include the development of statistical power curves, heteroskedastic patient outcomes, unknown statistical parameters, link functions for 0-1 outcomes, longitudinal studies, response-adaptive stopping times, dynamic enrollment decisions, online rewards, unknown labelings of covariates, and more general post-trial populations that correspond to a range of market exclusivity agreements. Some of these are conceptually straightforward to incorporate to our model, while others are not included in our base model because they merit a full, in-depth analysis in separate work.

Statistical power curves. Statistical power curves can reassure trial designers and regulatory bodies that true effects of a given size have a sufficient probability to be identified. This is primarily a frequentist concept, whereas our proposed fEVI-type allocation policies are based on Bayesian methods. Fortunately, power curves can be generated by a sequence of Monte Carlo simulation experiments (Berry 2006), with each Monte Carlo experiment generating a probability of detecting a known true effect using a fixed instance setting (see Section 6), and by varying that known true effect through a range of interest. This is in line with regulatory guidance for reporting complex adaptive trials (FDA 2019).

Heteroskedasticity. The variance of patient outcomes conditional on treatments and patient types, $\mathbb{V}(Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t)$, might not be constant σ^2 in practice. For example, it may be that the outcomes of all patients receiving treatment w have variance $\sigma^2_{\cdot,w}$, or that the outcome of a patient with covariates $\mathbf{x}$ has variance $\sigma^2_{\mathbf{x},\cdot}$, or perhaps the variance depends on the specific patient covariate/treatment combination, $\sigma^2_{\mathbf{x},w}$. For these cases, one could adapt the conjugate normal model in Assumption 3 and the Bayesian inference process in Appendix EC.1 as follows. Firstly, one would modify Assumption 3 so that

$$Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}((W_t \otimes \mathbf{X}_t)\boldsymbol{\mu}, \sigma^2_{\mathbf{X}_t,W_t})$$

with each $\sigma^2_{\mathbf{X}_t,W_t} > 0$. Secondly, the Bayesian updating would need to be adapted to account for different variances. For the one-step updating in (6a)-(6b), we would simply replace σ^2 in the denominators with $\sigma^2_{\mathbf{X}_{t-\Delta},W_{t-\Delta}}$. For the preposterior standard deviation $\tilde{\boldsymbol{\sigma}}$, which we need to characterize in order to implement the fEVI allocation policy, one would replace $\sigma^2 I_t$ in (EC.5) with the matrix that has zeros in all non-diagonal entries and has vector $(\sigma^2_{\mathbf{X}_1,W_1}, \sigma^2_{\mathbf{X}_2,W_2}, \ldots, \sigma^2_{\mathbf{X}_t,W_t})^\top$ as its diagonal.

Unknown patient outcome variance. If the variance of patient outcomes is unknown, then alternative conjugate models can be used, for example a normal-gamma model (Bernardo and Smith 1994). Related work on Bayesian sequential learning suggests that the online updating of estimators of outcome variances and correlations can be pragmatic and useful, in conjunction with the robust prior manipulations of Section 6 (Powell and Ryzhov 2012, Xie et al. 2016, Chick et al. 2021).

Link function for 0-1 outcomes. Our base model assumed real-valued observations that can be well-approximated with a normal distribution with mean $r_{\boldsymbol{\mu}}(\mathbf{X}_t, W_t) = \mathbb{E}[Y_t \mid \boldsymbol{\mu}, \mathbf{X}_t, W_t]$. There are also many trials with 0-1 responses, where logistic or other linkage function estimates for the probability of a positive response are employed: e.g., the probability of a positive response may be given by a link function σ, with $P(Y_t = 1 \mid \mathbf{X}_t, W_t) = \sigma(r_{\boldsymbol{\mu}}(\mathbf{X}_t, W_t))$. For example, the logistic model has $\sigma(a) = 1/(1 + e^{-a})$ and the probit model has $\sigma(a) = \Phi(a)$. Wang et al. (2016) show how to do so for contextual learning in a setting that allows the probability of successful outcomes to depend on characteristics of the treatment and of the patient separately, but do not explicitly consider predictive covariates (in the language of our work). Their work nonetheless seems to provide a direct method to adapt the predictive/prognostic linear model for real-valued outcomes here to logistic, probit, and other models for 0-1 outcomes.

Longitudinal studies and surrogate measures. The current work presumes that outcomes are observed after a specific duration of time. That duration, together with the rate of recruitment of the trial, provide the parameter Δ that represents the number of patients in the pipeline. Here, we consider short to medium term delays, during which a small to moderate number of patients are recruited, so that there is time to adapt the response to incoming data before the trial ends. In some studies, monitoring data might be available before the final primary outcome is observed, such as in longer-term survival studies. It is therefore of interest to incorporate surrogate measures or data arriving through time that correlate with final outcomes before they are fully observed. Techniques of Anderer et al. (2022) would be promising for incorporation in this context.

Adaptive stopping times. The analysis above assumed a known, fixed sample size T, as is commonly done in related simulation optimization literature (Frazier et al. 2008, 2009, Ryzhov et al. 2012). This is convenient for asymptotic analysis. There is practical interest in trials whose sample size is response adaptive (e.g., Berry 2011, Williamson et al. 2017, Pallmann et al. 2018, Villar and Rosenberger 2018, Rojas-Cordova and Bish 2018, Wang and Yee 2019, Chick et al. 2021). The base model in Section 2 can have a response-adaptive duration by allowing T to be a stopping time, by defining $V_t(\mathbf{k})$ in the first line of (11) for arbitrary $t = 0, 1, \ldots$, and by allowing $G(\mathbf{K}_t)$ to be selected at each time t in the first line of (11) to represent the trial's conclusion at time $T = t$ (in addition to the treatment choice options in that equation). We would make similar changes to the definition of the fEVI-index in (12) to be valid for arbitrary $t = 0, 1, \ldots$. Key to implementation in practice would be an efficient tool to compute a stopping time. See Chick et al. (2021), Eckman and Henderson (2022) for related work on computing stopping rules.

Dynamic enrollment decisions. The main model assumes that T trial participants arrive sequentially, and the cost of enrollment is not explicitly modeled. Our model of sequential enrollment assumes that all patients that arrive are enrolled, yet is compatible with the implicit modeling of the selection of T trial participants as part of a larger stream of patients that might be eligible for the trial. Not each patient would bring the same expected value of information if enrolled. For example, if the reward for each treatment is well known for type A patients, but not well known for type B patients, then it may be useful to wait until the next type B patient arrives to make an enrollment in the trial, rather than enrolling more type A patients. To this end, it may therefore be of interest to model the variable

cost for each trial participant, and compare that variable cost with the expected value of information of enrolling that patient in the trial.

Online rewards. The main model in Section 2 assumes offline rewards. We note that the bandit literature (Auer 2002) typically focuses on online rewards, the R&S literature focuses on offline rewards (Kim and Nelson 2006, Chick 2006), and stoppable bandits (Glazebrook 1979) can maximize the sum of online and offline rewards. In our context, online rewards would be the cumulative expected outcomes of patients enrolled in the trial, $\sum_{t=1}^{T} r_{\boldsymbol{\mu}}(\mathbf{X}_t, W_t)$. Online rewards may be particularly relevant for rare diseases (Williamson et al. 2017, Alban et al. 2022), and can be included in our model by adding $r_{\boldsymbol{\mu}}(\mathbf{X}_{t+1}, W_{t+1})$ to each treatment option in the first line of (11).

Related work that does not consider covariates (Ryzhov et al. 2012, Chick et al. 2017) suggests that Bellman's equation in (11) would still give an optimal policy for the cases of online rewards and adaptive stopping times. Moreover, Ryzhov et al. (2012) suggests an effective heuristic for online rewards, and Chick et al. (2021) suggests it may be useful to explore multi-step, multi-arm, lookahead indices for stopping times. A fuller exploration of these topics merits further work.

Unknown labeling. Our base model assumes the labeling is known, corresponding to an assumption that it is known which covariates are potentially predictive (e.g., through biological hypothesis) and which are prognostic (through medical experience). The inference of which factors are potentially predictive is also of interest (e.g., Bastani and Bayati 2020). This is related to the problem of model selection. Krishnamurthy and Athey (2022) discuss model selection from a set of nested models, but the set of model labelings may be more related to partially ordered sets rather than a pure nesting of models (for example, a covariate might be prognostic or not, as well as predictive with respect to different sets of treatments or not). Carranza et al. (2022) also account for predictive and prognostic effects, in the language of the current paper, and suggest that model selection is an area for separate further work, to which we concur. Possible implementations of model selection to infer an unknown labeling include model selection using criteria such as fractional Bayes factors, AIC and BIC (O'Hagan 1997, Kadane and Lazar 2004), variable selection using Lasso (Bastani and Bayati 2020), and the spike and slab approach and its variations (Malsiner-Walli and Wagner 2018).

Market exclusivity. Our base model assumes a fixed post-trial population. This allowed us to assume $P=1$ in (2), and allowed for a simplification of several results. This assumption may be reasonable for some health technology assessments, but might not be reflective of some forms of market exclusivity where the patent protection period is fixed, so that market access occurs for a shorter period of time if the trial is run longer. This does not pose a problem for a trial of fixed duration such as in our base model. If the trial's sample size and duration is a (random) stopping time, then the model could be adapted by allowing the post-trial population to be decreasing in T, for example $P(T)=P\cdot(1-T/H)$ for some fixed maximum patent protection horizon H, and relevant population size P. This would involve replacing the terminal reward $G(\mathbf{k})$ of (9) with an expected reward on stopping at time $T=t$ under this specific model of patent protection,

$$G_{\textit{fixed patent horizon}}(\mathbf{K}_t,t)=P\cdot(1-t/H)\mathbb{E}[\max_{\tilde{w}}(\tilde{w}\otimes\tilde{\mathbf{X}}_1)\boldsymbol{\theta}_{t+\Delta}\mid\mathbf{K}_t],$$

and allowing for the flexibility to stop as in the adaptive stopping time extension above. This modeling may allow for theoretical and computational results. See Alban et al. (2022) for further discussion of these issues.

References

Alban A, Chick SE, Forster M (2022) Value-based clinical trials: selecting trial lengths and recruitment rates in different regulatory contexts. *Management Science* (accepted to appear).

Anderer A, Bastani H, Silberholz J (2022) Adaptive clinical trial designs with surrogates: When should we bother? *Management Science* 68(3):1982–2002.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning* 3:397–422.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates online decision making with high-dimensional covariates. *Operations Research* 68(1):276–294.

Bernardo JM, Smith AFM (1994) *Bayesian Theory* (Chichester, UK: Wiley).

Berry D (2006) Bayesian clinical trials. *Nat Rev Drug Discov* 5:27–36.

Berry DA (2011) Adaptive clinical trials in oncology. *Nat Rev Clinical Oncology* 9(4):199–207.

Bertsekas DP, Shreve SE (1978) *Stochastic optimal control: The discrete time case* (Belmont, Massachusetts: Athena Scientific).

Billingsley P (2008) *Probability and measure* (John Wiley & Sons).

Carranza AG, Krishnamurthy SK, Athey S (2022) Flexible and efficient contextual bandits with heterogeneous treatment effect oracle. Available at https://arxiv.org/abs/2203.16668.

Chick SE (2006) Subjective probability and Bayesian methodology. Henderson S, Nelson B, eds., *Handbooks in Operations Research and Management Science: Simulation*, chapter 9 (Elsevier).

Chick SE, Forster M, Pertile P (2017) A Bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society. Series B* 79(5):1439–1462.

Chick SE, Gans N, Yapar O (2021) Bayesian sequential learning for clinical trials of multiple correlated medical interventions. *Management Science* accepted to appear:doi.org/10.1287/mnsc.2021.4137.

Eckman DJ, Henderson SG (2022) Posterior-based stopping rules for bayesian ranking-and-selection procedures. *INFORMS Journal on Computing* URL https://doi.org/10.1287/ijoc.2021.1132.

FDA (2019) Interacting with the FDA on complex innovative trial designs for drugs and biological products. US Food and Drug Administration, https://www.fda.gov/media/130897/download.

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Frazier PI, Powell WB, Dayanik S (2009) The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* 21(4):599–613.

Gelman A, Carlin JB, Stern HS, et al. (2013) *Bayesian data analysis* (CRC press).

Glazebrook KD (1979) Stoppable families of alternative bandit processes. *Journal of Applied Probability* 843–854.

Kadane J, Lazar N (2004) Methods and criteria for model selection. *JASA* 99:279–290.

Kim SH, Nelson BL (2006) Selecting the best system. Henderson S, Nelson B, eds., *Handbooks in Operations Research and Management Science*, chapter 17 (Elsevier).

Krishnamurthy SK, Athey S (2022) Optimal model selection in contextual bandits with many classes via offline oracles. Available at https://arxiv.org/abs/2106.06483.

Malsiner-Walli G, Wagner H (2018) Comparing spike and slab priors for Bayesian variable selection. *arXiv preprint arXiv:1812.07259* .

O'Hagan A (1997) Properties of intrinsic and fractional bayes factors. *Test* 6:101–118.

Pallmann P, et al. (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine* 16(29), Accessed May 7, 2018, https://doi.org/10.1186/s12916-018-1017-7.

Pocock SJ, Simon R (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31:103–115.

Powell WB, Ryzhov IO (2012) *Optimal learning* (John Wiley & Sons).

Raiffa H, Schlaifer R (1961) *Applied Statistical Decision Theory*.

Rojas-Cordova A, Bish EK (2018) Optimal patient enrollment in sequential adaptive clinical trials with binary response. Available at SSRN, https://ssrn.com/abstract=3234590.

Russo D (2020) Simple Bayesian algorithms for best arm identification. *Operations Research* 68(6):1625–1931.

Ryzhov IO, Powell WB, Frazier PI (2012) The knowledge gradient algorithm for a general class of online learning problems. *Operations Research* 60(1):180–195.

Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge University Press).

Villar SS, Rosenberger WF (2018) Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics* 74(1):49–57.

Wang H, Yee D (2019) I-SPY 2: A neoadjuvant adaptive clinical trial designed to improve outcomes in high-risk breast cancer. *Curr Breast Cancer Rep.* 11(4):303–310.

Wang Y, Wang C, Powell W (2016) The knowledge gradient for sequential decision making with stochastic binary feedbacks. *Proc. 33rd International Conference on Machine Learning*, PMLR 48:1138–1147.

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

Xie J, Frazier PI, Chick SE (2016) Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research* 64(2):542–559.

Zhou X, Liu S, Kim ES, Herbst RS, Lee JJJ (2008) Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials* 5(3):181–193, ISSN 17407745.