UNIVERSITY of York

Discussion Papers in Economics

No. 15/09

A Bayesian Decision-Theoretic Model of Sequential Experimentation with Delayed Response

Version 2

Stephen Chick, Martin Forster and Paolo Pertile

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

A Bayesian Decision-Theoretic Model of Sequential Experimentation with Delayed Response

Stephen Chick[*] Martin Forster,[†] Paolo Pertile[‡]

August 12, 2015

Abstract

We propose a Bayesian decision-theoretic model of a fully sequential experiment in which the real-valued primary end point is observed with delay. The goal is to identify the sequential experiment which maximises the expected benefits of technology adoption decisions, minus sampling costs. The solution yields a unified policy defining the optimal 'do not experiment'/'fixed sample size experiment'/'sequential experiment' regions and optimal stopping boundaries for sequential sampling, as a function of the prior mean benefit and the size of the delay. We apply the model to the field of medical statistics, using data from a published trial investigating the clinical- and cost-effectiveness of drug-eluting stents versus bare metal stents.

Keywords: Bayesian inference; Clinical trials; Delayed observations; Health economics; Sequential experimentation

[*] Corresponding author. Novartis Chair for Healthcare Management, Technology and Operations Management Area, INSEAD, Boulevard de Constance, 77300 Fontainebleau, FRANCE. (33) 1.60.72.41.57. stephen.chick@insead.edu

[†] Department of Economics and Related Studies, University of York, York, U.K. mf8@york.ac.uk

[‡] Department of Economics, University of Verona, Verona, Italy. paolo.pertile@univr.it

1 Introduction

The ethical and economic advantages of sequential and adaptive trial designs are well documented (Armitage, 1975; Berry, 1985; Whitehead, 1997; Jennison and Turnbull, 1999; Emerson et al., 2007a,b). It is also common to observe data on patient outcomes some time after treatment has taken place. For example, Brown et al. (2000) measured outcomes immediately following surgery and again at one and twenty four hours post-surgery; Connor et al. (2015) measured the primary end point at 90 days and Moses et al. (2003) measured outcomes over one year. Less well researched is the question of how sequential experiments should be adjusted when the primary end point arrives with delay.

This question is especially important given the increasing interest of both academics and regulators in the ethical and economic advantages that sequential designs present (European Medicines Agency, 2006; United States Food and Drug Administration, 2010; Stallard and Todd, 2011). Concern about avoiding unnecessary recruitment to the trial, past the point at which evidence is deemed to be conclusive, means there is a growing focus on valuing the cost of carrying out research, together with the benefits that accrue to trial participants and patients who may benefit from a new technology (Lewis et al., 2007; Willan and Kowgier, 2008; Pertile et al., 2014). Indeed, the UK National Institute for Health and Care Excellence (NICE, 2012) examines cost and effectiveness when making tradeoffs in care, and the value based health care movement (Porter, 2010) calls for increased attention to the health benefits obtained for a given level of expenditure.

Hampson and Jennison (2013) provide a useful survey of the literature on sequential trial design with delay. They solve frequentist models which minimise the expected sample size of the trial, with penalty terms for, or constraints on, Type I and II error rates. Broglio et al. (2014) present a Bayes adaptive design which stops recruitment to a trial if the predictive probability of success upon immediate cessation of recruitment and follow-up of pipeline patients exceeds a predefined probability, or if the predictive probability of success at the maximum sample size is lower than a predefined futility probability.

In discussing Hampson and Jennison (2013), Draper (2013) comments that the Neyman-Pearson framework, with its two simple hypotheses, unnecessarily 'dichotomises' the measurement of a treatment effect. As an alternative, he suggests using the Quality Adjusted Life Year, or QALY, citing its use by NICE (2012). Burman (2013) suggests using a Bayesian decision-theoretic framework which permits a prior distribution to account for previous information and which incorporates both the cost of sampling and value of information calculations for the trial's pipeline subjects.

We implement the recommendations of Draper (2013) and Burman (2013) by proposing a model for experimental design which compares two health technologies and which is fully sequential (as opposed to one which allocates all patients at once, or in a small number of groups), where outcomes are observed with a specified delay and converted to economic values using standard cost utility analysis. The model uses a Bayesian, decision-theoretic, framework and seeks to maximise the expected benefits of the technology adoption decision which is made on the basis of experimental data, less the expected cost of the sequential experiment itself. The model allows the treatment effect to accrue to trial participants as the study progresses, and it accounts for a potential fixed cost of switching technologies. Discounting of future costs and benefits is permitted, so that the model may be applied to health technology assessments such as those considered by NICE. To the best of our knowledge, ours is the first to combine all of these features within a unified framework and, in so doing, to show how the size of the delay determines the optimal trial design.

Section 2 presents our model. As with the model of Hampson and Jennison (2013), dynamic programming techniques (Bertsekas and Shreve, 1978) characterise the optimal sampling and technology adoption policy analytically for responses which are from the regular exponential family when parameter uncertainty is modeled by the conjugate family of prior distributions. Additional results are provided for the base case of the paper: normally distributed responses with unknown mean and known sampling variance.

Section 3 highlights the main features of the optimal policy with an illustrative example. We show how prior information is used to establish whether or not it is worth starting the trial in the first place, whether it is optimal to conclude recruitment to the trial before observations on primary end points start arriving, or whether observation of end points and recruitment to the trial should take place concurrently. It also provides optimal stopping boundaries for the sequential analysis which are typical of similar models which ignore delay.

Section 4 presents an application using data from a published clinical trial for drug eluting stents to provide an assessment of the operating characteristics of the model's optimal policy. Comparisons with alternative approaches to trial design show the net expected benefit that results from applying the optimal policy. In order to account for the competing, and sometimes conflicting, perspectives of the various disciplines that are involved in trial design (Emerson et al., 2007b), we also estimate the probability that the correct technology is selected. We show how delay and other parameters influence the optimal policy and stopping boundaries. The expected benefit of sequential sampling over other experimental designs is found to decrease as the number of pipeline data points increases, a result that is consistent with Hampson and Jennison's finding that the advantage of group sequential methods, in terms of lower expected sample sizes, are lower the longer is the delay.

Appendix S of the Online Supplemental Material describes how we numerically compute the optimal stopping policy of the core model of section 2. We convert our discrete time problem into a suitable continuous time model using the approach of Chernoff (1961). The solution of the resulting free boundary problem is convenient for approximating the optimal stopping policy for the examples in sections 3 and 4. As such, the work extends results for stochastic simulation optimisation problems in management science (Chick and Gans, 2009; Chick and Frazier, 2012) and for economic evaluation (Pertile et al., 2014). Appendix S also points to related literature which characterise asymptotics for optimal stopping boundaries for some special cases, and to the multi-armed bandit literature.

2 The model

We consider a two-armed, sequential clinical trial in which study units are allocated at random, and in a pairwise manner, to either a control (the current best available standard) health technology or a new one. There is a sampling cost $c \in \mathbb{R}_{\geq 0} \equiv [0, \infty)$ per pairwise allocation made. The purpose of the trial is to evaluate which technology should be used to treat $P \in \mathbb{R}_{>0} \equiv (0, \infty)$ patients upon stopping the trial. A one-time switching cost $I \in \mathbb{R}_{\geq 0}$ is incurred if the decision is made to adopt the new technology. No such cost is incurred if the decision is made to continue with the standard technology.

Effectiveness is denoted by the random variable $E_n \in \mathbb{R}$ if a patient is assigned to the new technology and $E_s \in \mathbb{R}$ if the patient is assigned to the control. The patient-level costs of using each technology are the random variables $C_n \in \mathbb{R}_{\geq 0}$ and $C_s \in \mathbb{R}_{\geq 0}$. It is assumed that all patients complete their assigned course of treatment, there is no loss to follow up, and E_n, E_s, C_n and C_s are observed without measurement error.

Following standard approaches in Bayesian decision-theoretic models (see, for example, Berry and Ho 1988, Lewis et al. 2007 and Pertile et al. 2014) and in line with the suggestion of Burman (2013), a common unit of measurement is used to value benefits and costs. We assume that effectiveness is valued in monetary terms, using survey data or information provided by a regulatory body such as NICE.[1] Define $\lambda \in \mathbb{R}_{>0}$ as the monetary value of one unit of effectiveness. Then the individual level incremental net monetary benefit (INMB) of the new technology versus the existing one for pairwise allocation i is:

$$X_i = \lambda(E_{n,i} - E_{s,i}) - \delta_{\text{CE}}(C_{n,i} - C_{s,i}), \tag{1}$$

where $\delta_{\text{CE}} = 1$ if the experiment assesses cost-effectiveness and $\delta_{\text{CE}} = 0$ if it assesses effectiveness only. It is assumed that $X_i \sim \mathcal{N}(W, \sigma_X^2)$, $i = 1, 2, \ldots, T_{\max}$, where $T_{\max} \in \mathbb{Z}_{>0}$ is the maximum number of pairwise allocations which can be made in the trial. W is assumed to be unknown and σ_X^2 is assumed known. The prior distribution on W is assumed to be $\mathcal{N}(\mu_0, \sigma_0^2)$. Given knowledge of both σ_X^2 and σ_0^2, the effective sample size of the prior distribution is $n_0 = \sigma_X^2/\sigma_0^2$.

The annual rate of accrual to the trial is assumed to be constant and equal to $R \in \mathbb{R}_{>0}$. In contrast to the model of Pertile et al. (2014), the X_i arrive with a delay of $\tau \in \mathbb{Z}_{\geq 0}$, $\tau < T_{\max}$, pairwise allocations, at which point they are used to update the prior/posterior distribution of W in a sequential manner. The number of pairwise allocations τ of delay therefore depends on the rate of accrual, R, and the time delay in observing the outcome.

The model permits future benefits and costs to be down-weighted using an annual continuous discount rate $\rho_{\text{year}} \geq 0$. With time measured as the number of pairwise allocations made, $\rho = \rho_{\text{year}}/R$ is the continuous time discount rate and $\tilde{\rho} = e^{\rho_{\text{year}}/R} - 1$ is the discrete time discount rate.

2.1 The decision problem in discrete time

Define $\mathbb{T} \equiv \{0, 1, \ldots, T_{\max}\}$, and define $T \in \mathbb{T}$ as the time at which pairwise allocations cease to be made. Define $\bar{\mathbb{T}} \equiv \{0, 1, \ldots, T_{\max} + \tau\}$ as the set of equally spaced times where pairwise allocations and/or a choice to adopt one of the two technologies may be made.

It is assumed that, once pairwise allocations cease to be made, sampling cannot be restarted. At each $t \in \mathbb{T} \backslash \{T_{\max}\}$, an action a_t is chosen from the set of available actions, $\mathcal{A} \equiv \{0, 1\}$, such that $a_t = 1$ denotes choosing to make a pairwise allocation (so that $T > t$) and $a_t = 0$

[1] For example, NICE values one Quality Adjusted Life Year (QALY) at between £20,000 and £30,000.

denotes choosing not to make a pairwise allocation. At the first occurrence of $a_t = 0$, pairwise allocations cease (so that $T = t$ and $a_t = 0$ for all $t > T$).

For $t \leq \tau$, a_t is chosen only on the basis of prior information. For $\tau < t < T_{\max}$, the action can be a function of the $\{X_i\}_{1 \leq i \leq t-\tau}$. For $t = \tau, \ldots, T_{\max} - 1$, the ordering of events is as follows: action a_t is chosen; realisation $X_{t+1-\tau} = x_{t+1-\tau}$ is observed; prior distribution on W is updated. If sampling continues as far as $t = T_{\max}$, $T = T_{\max}$ and sampling stops.

Once sampling is stopped, one must wait to observe all outcomes for the 'pipeline subjects' – those who have been treated but whose outcomes have yet to be observed – before making the technology adoption decision. Define $\mathcal{D} \in \{n, s\}$ as the decision concerning whether to choose the new technology ($\mathcal{D} = n$) or the standard ($\mathcal{D} = s$). This adoption decision is made at time 0 if $a_0 = 0$, because no pairwise allocations will be made. It is made at time $T + \tau$ if $a_0 = 1$, because of the delay, and in this case $T > 0$.

More compactly, the adoption decision is made at time $\mathbf{1}_{T>0}(T+\tau)$, where $\mathbf{1}_F$ is the indicator function, equal to 1 if the event F is realized and 0 otherwise. The expected reward from selecting technology $\mathcal{D}$, ignoring the cost of sampling and discounting, is $\mathbf{1}_{\mathcal{D}=n}(PW - I)$. A policy π is a dynamic method of deciding, at each time t, to take an action from $\mathcal{A}$ using the history of choices and realisations that have so far accrued, and a technology adoption decision from $\mathcal{D}$. The objective is to establish a policy π^* which maximises the expected reward of the sequential sampling process and adoption decision.

More formally, define $\mathcal{F} = (\mathcal{F}_t)_{t \in \bar{\mathbb{T}}}$ as the natural filtration generated by the $\{X_i\}_{1 \leq i \leq t-\tau}$ for $t \in \bar{\mathbb{T}}$. Note that $\mathcal{F}_t = \mathcal{F}_0$ for $t \in \{0, 1, \ldots, \tau\}$ due to the delay. Define variables tracking the 'effective sample size' (in terms of the number of realisations of pairwise allocations) in the posterior distribution for W, and 'effective cumulative sum' of realisations, given information available to time $t \in \bar{\mathbb{T}}$,

$$n_t = n_0 + (t - \tau)^+, \tag{2}$$

$$Y_t = \mu_0 n_0 + \sum_{i=1}^{(t-\tau)^+} X_i, \tag{3}$$

where $(m)^+ = \max(0, m)$ and the sum is equal to 0 if the upper bound for the summation is 0.

The posterior beliefs about W at time t have a normal distribution

$$W \mid \mathcal{F}_t \sim \mathcal{N}(\mu_t, \sigma_X^2 / n_t), \text{ where:} \tag{4a}$$

$$\mu_t = Y_t / n_t. \tag{4b}$$

We may use (y_t, n_t) as a sufficient statistic for W conditional on $\mathcal{F}_t$ and we use (y_t, t) as a state because it also provides information about the number of pipeline subjects.

The policy π defines a mapping $f(y_t, t) : \mathbb{R} \times \mathbb{T} \backslash \{T_{\max}\} \rightarrow \mathcal{A}$ from states to a decision to stop or to continue sampling, which in turn determines T. A policy π also specifies the choice of the new technology or control, $\mathcal{D} \in \{n, s\}$, as discussed above.

By construction, T is a stopping time with respect to the filtration $\mathcal{F}$ taking values in $\mathbb{T}$; $\mathcal{D} \in \{n, s\}$ is $\mathcal{F}_{\mathbf{1}_{T>0}(T+\tau)}$-measurable and π is measurable with respect to $\mathcal{F}$. Let Π be the set of all policies which are so measurable with respect to $\mathcal{F}$. We write $\mathbb{E}_\pi$ to denote the expectation

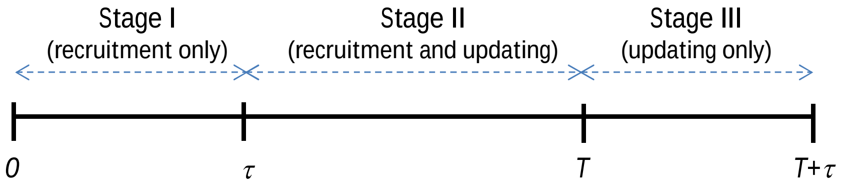


Figure 1: Stages of the problem with stopping time T and delay τ

with respect to the measure π induces on the sequence of observations and decisions, and $\mathbb{E}$ to indicate the expectation when it does not depend on π. A summary of notation may be found in Table 2 at the end of Appendix S in the Online Supplemental Material.

The expected reward from a policy $\pi \in \Pi$ depends on the parameters of the prior distribution (μ_0, n_0), and is determined by the cost of sampling, benefits to patients during the trial (if permitted), and benefits from the technology adoption decision:

$$V^{\pi}(\mu_0, n_0) = \mathbb{E}_{\pi}\left[\left\{\sum_{t=0}^{T-1}\frac{-c+\delta_{\text{on}}X_{t+1}}{(1+\tilde{\rho})^{t}}\right\} + \frac{\mathbf{1}_{\mathcal{D}=n}(PW-I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}}\middle|\mu_0, n_0\right]. \tag{5}$$

Here, $\delta_{\text{on}} = 1$ if the rewards for patients participating in the trial are to be included in the reward function (known as ‘online learning’) and $\delta_{\text{on}} = 0$ if they are not (‘offline learning’).[2] The term $\mathbf{1}_{T>0}(T+\tau)$ indicates that a penalty for discounting is only relevant for the terminal reward if at least one pairwise allocation is made.

The objective is defined to be that of finding a policy $\pi^* \in \Pi$ such that

$$V^{\pi^*}(\mu_0, n_0) = \sup_{\pi\in\Pi} V^{\pi}(\mu_0, n_0). \tag{6}$$

It will be useful to analyze three distinct stages of the trial in order to characterise the optimal policy. These are illustrated in Figure 1. During *stage I* ($t \in \{0, 1, \ldots, \tau-1\}$) pairwise allocations are made sequentially and no outcomes are observed, owing to the delay. During *stage II* ($t \in \{\tau, \tau+1, \ldots, T-1\}$) pairwise allocations are made, realisations $x_{t+1-\tau}$ for pipeline subjects arrive sequentially and are used to carry out Bayesian updating. During *stage III* ($t \in \{T, T+1, \ldots, T+\tau\}$) no pairwise allocations are made, observations on pipeline subjects arrive sequentially and are used to carry out Bayesian updating.

In sections 2.1.1–2.1.3 we formulate a dynamic program (Bertsekas and Shreve, 1978) by developing Bellman’s equation for the expected reward in reverse time from stage III to stage I. Section 2.2 then justifies how an optimal policy $\pi^* \in \Pi$ can be determined from Bellman’s equation and provides further results for two special cases of the problem.

2.1.1 Optimal rewards in stage III

Stage III is entered when recruitment to the trial stops at time T. The optimal expected reward upon entering stage III depends on the $u = \min(T, \tau)$ pairwise allocations in the pipeline: $u = T$

[2] Traditional trials set $\delta_{\text{on}} = 0$ implicitly. Setting $\delta_{\text{on}} = 1$ values outcomes for both trial participants and patients post-trial.

if stopping takes place during stage I, and $u = \tau$ if it takes place during stage II. Let $Z_{t,u}$ be the posterior expected INMB at the patient level, given that t pairwise allocations have been made and u outcomes are yet to be observed. Then:

$$Z_{t,u} = \mathbb{E}[\, \mu_{t+u} \mid \mathcal{F}_t \,]; \tag{7a}$$

$$Z_{t,u} \sim \mathcal{N}\left(\mu_t, \frac{\sigma_X^2 u}{n_t(n_t+u)}\right). \tag{7b}$$

If $T > 0$, the last of the observations on pipeline subjects will be observed τ time units after stopping. An adoption decision can be made immediately if $T = 0$ because no trial takes place. Once all outcomes on pipeline subjects are observed, it is optimal to adopt the new technology ($\mathcal{D} = n$) if $PZ_{T,\min(T,\tau)} > I$ and the standard one ($\mathcal{D} = s$) otherwise. Define $G : \mathbb{R} \times \mathbb{N}_0 \to \mathbb{R}$ as the optimal discounted expected reward following a decision to stop at time $T = t$ and wait for the observations on pipeline subjects before making an adoption decision:

$$G(y_t, t) = (1+\tilde{\rho})^{-\mathbf{1}_{t>0}\tau} \mathbb{E}[\, (PZ_{T,\min(T,\tau)} - I)^+ \mid Y_T = y_t, T = t \,]. \tag{8}$$

2.1.2 Bellman's equation for stage II

For stage II, let $\mathbb{T}_{\text{II}} \equiv \{\tau, \ldots, T_{\max} - 1\}$ be the set of times at which pairwise allocations are being made, outcomes are being observed and Bayes updating is taking place. The decision about whether to make the next pairwise allocation is based on a comparison of G in Eq. (8) with the expected reward of making that allocation, observing the outcome of the next pairwise allocation in the pipeline, and having the option to continue behaving optimally on the basis of that outcome. Define $B(y_t, t) : \mathbb{R} \times (\mathbb{T}_{\text{II}} \cup \{T_{\max}\}) \to \mathbb{R}$ as having the maximum value of the expected reward for the next allocation decision, given that t pairwise allocations have been made and $(t - \tau)$ have been observed, resulting in a posterior mean of y_t/n_t. Then Bellman's equation in stage II is:

$$\begin{aligned} B(y_t, t) = &\max_{a_t \in \mathcal{A}} \Big\{ G(y_t, t),\ -c + \delta_{\text{on}}(y_t/n_t) \\ &\qquad + (1+\tilde{\rho})^{-1}\, \mathbb{E}_\pi[\, B(y_t + X_{t+1-\tau}, t+1) \mid y_t, t \,] \Big\}, \quad t \in \mathbb{T}_{\text{II}}, \end{aligned} \tag{9a}$$

$$B(y_{T_{\max}}, T_{\max}) = G(y_{T_{\max}}, T_{\max}). \tag{9b}$$

If the second term in the maximand of Eq. (9a) exceeds the first, $a_t = 1$ and stage II continues with an additional pairwise allocation so that $T > t$. For first occurrence at which the first term exceeds the second, $a_t = 0$ and the stopping time is $T = t$; if the first term never exceeds the second, the trial runs to the maximum sample size ($T = T_{\max}$).

2.1.3 Bellman's equation for stage I

Bellman's equation for stage I is similar to that in Eq. (9a) for stage II, except that some simplifications can be made due to the structure of delayed sampling information when $\tau > 0$. The existence of delay implies that no observations on $X_{t+1-\tau}$ are available during stage I, so that $y_t = y_0$ and $n_t = n_0$ for $t \in \mathbb{T}_{\text{I}} \equiv \{0, 1, \ldots, \tau - 1\}$. Thus,

$$B(y_t, t) = \max_{a_t \in \mathcal{A}} \left\{ G(y_t, t), -c + \delta_{\text{on}}(y_0/n_0) + (1 + \tilde{\rho})^{-1} B(y_0, t + 1) \right\}, \quad t \in \mathbb{T}_{\text{I}}. \tag{10}$$

The special case of $\tau = 0$ is modeled by letting $\mathbb{T}_{\text{I}}$ be the empty set, letting stage II commence at time $t = 0$, and noting the simplification $G(y_t, t) = (Py_t/n_t - I)^+$ in Eq. (8).

2.2 Characterization of the optimal policy

This section shows that a policy $\pi \in \Pi$ is optimal for the sequential sampling problem in Eq. (6) if it selects (almost surely) the argmax of Bellman's equation in Eqs. 9 and 10. It provides additional structural results which characterise the optimal solution for some special cases.

We first observe that, for the special case of free, undiscounted sampling ($c = 0, \tilde{\rho} = 0$) with offline learning ($\delta_{\text{on}} = 0$), the following policy is optimal: sample as much as possible ($T = T_{\max}$) and select the new technology if the posterior mean net reward is positive ($P\mu_{T+\tau} > I$) once all outcomes are observed and the control otherwise. This result is trivial from the observation that information, in expectation, has a nonnegative value.

The special case of offline learning ($\delta_{\text{on}} = 0$), positive discounting ($\tilde{\rho} > 0$) and no time delay ($\tau = 0$) reduces to the special case of Chick and Gans (2009) for comparing a known alternative (control) with known mean reward 0 with an unknown alternative (new technology) with unknown mean reward $PW - I$. The special case of offline learning ($\delta_{\text{on}} = 0$), positive sampling costs ($c > 0$), no discounting ($\tilde{\rho} = 0$) and no time delay ($\tau = 0$) reduces to the special case of Chick and Frazier (2012) for the same comparison. We now draw upon, and extend, those results to account for general costs (that is, at least one of c and $\tilde{\rho}$ positive), delayed responses ($\tau \geq 0$), as well as both offline and online learning ($\delta_{\text{on}} \in \{0, 1\}$).

It will be useful to define $\bar{V}$ as the expected reward of an oracle who adopts the prior distribution for W and who will become aware of the true value of W immediately before starting the trial. The oracle then has the option to adopt one of the two technologies immediately, based on that information, and still run patients through the trial if there exists online learning and the expected reward for those patients exceeds the cost of sampling them. Let $T_{\max,\tilde{\rho}} = \sum_{t=0}^{T_{\max}-1} (1 + \tilde{\rho})^{-t}$ be the discounted maximum number of pairwise allocations in the trial. Then, given μ_0 and n_0 and prior to knowing W, define:

$$\bar{V}(\mu_0, n_0) = \mathbb{E}[(PW - I)^+ + \delta_{\text{on}}(W - c)^+ T_{\max,\tilde{\rho}} | \mu_0, n_0]. \tag{11}$$

The term $\mathbb{E}[(PW - I)^+ | \mu_0, n_0]$ is the oracle's expected reward from selecting the best alternative immediately before executing the trial (that is, assuming no penalties for discounting). The term $\mathbb{E}[\delta_{\text{on}}(W - c)^+ T_{\max,\tilde{\rho}} | \mu_0, n_0]$ is the oracle's expected reward from sampling all patient pairs if online learning is permitted and such sampling has positive net reward.

The first proposition links the expected reward of the oracle with the expected reward of any given policy, and will be useful for characterizing the optimal policies in Propositions 2.2 and 2.3 below.

Proposition 2.1 *For policies $\pi \in \Pi$:*

$$V^{\pi}(\mu_0, n_0) = \bar{V}(\mu_0, n_0) - \mathbb{E}_{\pi}\left[\mathcal{K}_{\pi} + \mathcal{S}_{\pi} + L_{\pi}\middle|\mu_0, n_0\right], \tag{12}$$

where the following terms are each non-negative for all sample paths:

$$\mathcal{K}_{\pi} \equiv \sum_{t=0}^{T-1} (c - \delta_{\text{on}}(W - (W-c)^{+}))/(1+\tilde{\rho})^{t}, \tag{13a}$$

$$\mathcal{S}_{\pi} \equiv \sum_{t=T}^{T_{\max}-1} \delta_{\text{on}}(W-c)^{+}/(1+\tilde{\rho})^{t}, \tag{13b}$$

$$\textit{and } L_{\pi} \equiv (PW - I)^{+} - \mathbf{1}_{\mathcal{D}=n}(PW-I)/(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}. \tag{13c}$$

Proofs of mathematical claims in the paper can be found in Appendix A below.

Here, $\mathcal{K}_{\pi}$ is the opportunity cost associated with sampling and $\mathcal{S}_{\pi}$ is a residual penalty in the presence of online sampling if the sample size is not optimal for the oracle. L_{π} is the opportunity cost associated with selecting a potentially suboptimal technology $\mathcal{D}$ after outcomes for all T pairwise allocations are observed, accounting for any discounting associated with a delayed selection.

From Eq. (12) it is clear that a policy π maximises V^{π} if and only if it minimises $\mathbb{E}_{\pi}[\mathcal{K}_{\pi}+\mathcal{S}_{\pi}+L_{\pi}|\mu_0, n_0]$. This minimisation is itself a sequential optimal stopping problem whose continuation costs per unit time are the terms in the summand of $\mathcal{K}_{\pi}$, and whose terminal cost is $\mathcal{S}_{\pi} + L_{\pi}$. This observation, together with the nonnegativity of $\mathcal{K}_{\pi}$, $\mathcal{S}_{\pi}$, and L_{π}, allow us state and prove the following results which show that Bellman's equation determines the optimal policy.

Proposition 2.2 *If all decisions of a policy $\pi \in \Pi$ attain the maximum in Bellman's equation in Eq. (10) for stage I decisions and in Eq. (9) for stage II decisions, and make technology adoption decisions as described in section 2.1.1 (π-almost surely), then that policy is optimal, i.e.,*

$$V^{\pi}(\mu_0, n_0) = V^{\pi^*}(\mu_0, n_0) = B(\mu_0 n_0, 0). \tag{14}$$

Proposition 2.3 *If $\tilde{\rho} > 0$ then the conclusions of Prop. 2.2 are also true when $T_{\max} = \infty$.*

The optimal policy might not be unique. The continuity of the values of the terms in Bellman's equation implies that there may be ties for certain parameter combinations. In applications one might choose to break such ties by picking the action which samples more rather than less. Such a choice offers no loss of expected reward, nor quality of inference.

The preceding propositions do not depend on properties of the normal distribution or the assumption of known sampling variances. Their proofs use the *a priori* integrability of W, the Markovian nature of Bayes' rule, and a finite state vector to describe the posterior distribution (e.g., as for sampling in the regular exponential family with a conjugate prior distribution for unknown parameters), assuming that vector replaces (y_t, t) as the state vector.

The next two results use properties of the normal distribution in their proofs. Prop. 2.4 uses the symmetrical nature of the normal distribution to derive a symmetry result for the value function when there is no discounting and no online learning. Prop. 2.5 makes explicit use of properties of the normal distribution and the assumption that σ_X^2 is known to provide an upper bound on the total number of pairwise allocations required by the optimal policy.

Proposition 2.4 *If $\tilde{\rho} = 0$ and $\delta_{\mathrm{on}} = 0$ then (i) $V^{\pi^*}(I/P+\Delta\mu, n_0) - P\Delta\mu = V^{\pi^*}(I/P-\Delta\mu, n_0)$, for all real valued $\Delta\mu$; (ii) $B((I/P+\Delta\mu)n_t, t) - P\Delta\mu = B((I/P-\Delta\mu)n_t, t)$, for all real valued $\Delta\mu$ and $t = 0, 1, \dots, T_{\max}$; and (iii) the set of states (μ_t, t) for which it is optimal to continue sampling is symmetric above and below the line $\mu = I/P$.*

Proposition 2.5 *If $\tilde{\rho} = 0$, $\delta_{\mathrm{on}} = 0$ and $c > 0$ then the optimal stopping time satisfies $T \leq 1 + (P^2\sigma_X^2)/(2\pi c^2) + \tau - n_0$ almost surely, even if $T_{\max}$ is larger than that upper bound.*

2.3 Approximation of the optimal policy

Solving for the optimal discrete time policy in Eq. (6) is challenging even with its characterisation in section 2.2 with Bellman's equation. We approximate the optimal solution using a related continuous time model in the spirit of the work of Chernoff (1961). Appendix S provides mathematical formalism and an overview of computational methods for doing so. In summary, the continuous time analog of Bellman's equation is a free boundary problem for a heat equation, the solution of which determines a continuation set $\mathcal{C}$, such that it is optimal at time t to continue sampling if $(\mu_t, t) \in \mathcal{C}$ and to stop sampling if (μ_t, t) is not in the closure of $\mathcal{C}$. A Matlab implementation is provided, and is used for the applications of sections 3 and 4.

3 Illustration of features of the optimal policy

This section illustrates the main features of the optimal policy and assesses some of its characteristics. The optimal policy π^* of Eq. (6) is called the 'Optimal Bayes Sequential' policy and is computed using techniques described in Appendix S. The stopping boundaries of the optimal policy are then used in Monte Carlo simulations of the discrete time problem. Parameter values are chosen for convenience and are not based on any real-life application. The material in this section is preparatory for the application of section 4, where data from a clinical trial is used to populate the model and to assess statistical and economic performance.

We compare the Optimal Bayes Sequential policy with two alternative policies. One, called the 'Fixed' policy, always makes a fixed number of pairwise allocations (in this section we set $T = T_{\max}$) and selects the new technology in preference to the existing one if the posterior mean, once all observations on pipeline subjects have arrived, strictly exceeds I/P. The 'Optimal Bayes One Stage' policy chooses a sample size $u^*(\mu_0)$ which maximises the net benefit of sampling in expectation,

$$u^*(\mu_0) = \arg\max_{u \in \mathbb{T}} \left\{ \left(\sum_{t=0}^{u-1} \frac{-c + \delta_{\mathrm{on}}\mu_0}{(1+\tilde{\rho})^t} \right) + \frac{\mathbb{E}[(PZ_{0,u} - I)^+ \mid \mu_0, n_0]}{(1+\tilde{\rho})^{\mathbf{1}_{u>0}(u+\tau)}} \right\}. \tag{15}$$

Two scenarios are considered. The ‘baseline’ scenario sets the time delay for observing the primary end point at one year and the rate of recruitment to the trial, R, at 1000 per year, so that τ (= 1000 patient pairs) is relatively high compared with $T_{\max}$ (= 2000 patient pairs); the ‘comparator’ halves the time delay to half a year, while keeping R and $T_{\max}$ unchanged, so that τ is also halved (to τ = 500). Other parameters are common to both baseline and comparator scenarios. The annual discount rate and the fixed cost of switching technologies are set to zero ($\rho = 0, I = 0$). The marginal cost of sampling is $c = 500$. $P = 20000$ patients benefit from the technology adoption decision. The sampling standard deviation is $\sigma_X = 20000$ and the standard deviation of the unknown mean is $\sigma_0 = 2000$. The effective sample size of the prior distribution is $n_0 = (\sigma_X/\sigma_0)^2 = 100$. There is no online learning (δ_{on} = 0).

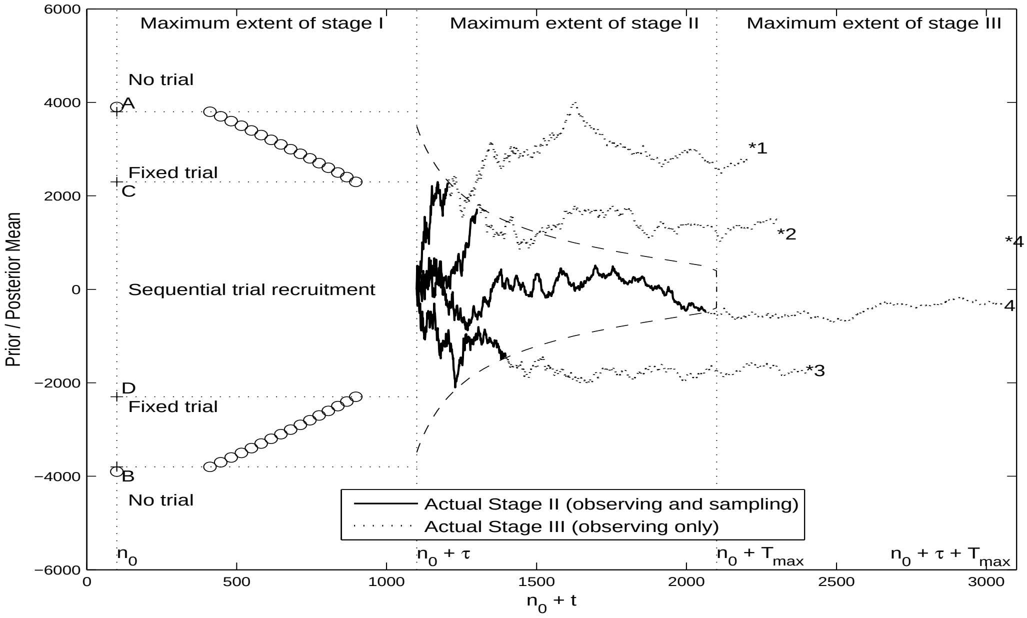
Figures 2(a) (baseline) and 2(b) (comparator) plot the optimal stopping boundaries for the scenarios in $(n_0 + t) \times$ prior/posterior mean space, together with some stage I optimal sample sizes and four stage II/III paths of the posterior mean. In Figure 2(a), the boundaries between the ‘no trial’/‘fixed trial’/‘sequential trial’ ranges for the prior mean are marked with a ‘+’ and labelled A, B, C and D. If the prior mean is above A or below B, it is optimal not to carry out any trial and instead base the technology adoption decision on the value of μ_0 alone. If the prior mean is between A and C or D and B, it is optimal to carry out a fixed sample trial (do stage I sampling and continue to stage III, with no stage II sampling). For the baseline scenario, the optimal fixed sample sizes for such trials for some values of the prior mean are indicated by ‘$\circ$’ in these two regions. If μ_0 lies between C and D, it is optimal to carry out a sequential trial, with stage II sampling. The stage II free boundaries are shown as dashed lines.

For the baseline scenario, Figure 2(a) shows that stage II starts at an effective sample size of $n_0 + \tau = 1100$. Because there is no discounting, there is symmetry above and below $\mu = I/P = 0$ in the stage II stopping rules and the stage I fixed sample sizes (recall Prop. 2.4). For the comparator scenario, Figure 2(b) shows that stage III starts earlier, at an effective sample size of 600, owing to the reduced time delay. In the comparator scenario, the stage I regions between A and C and D and B that are shown in Figure 2(a) are eliminated, implying that it is not optimal to carry out a trial of a fixed sample size for any value of μ_0.

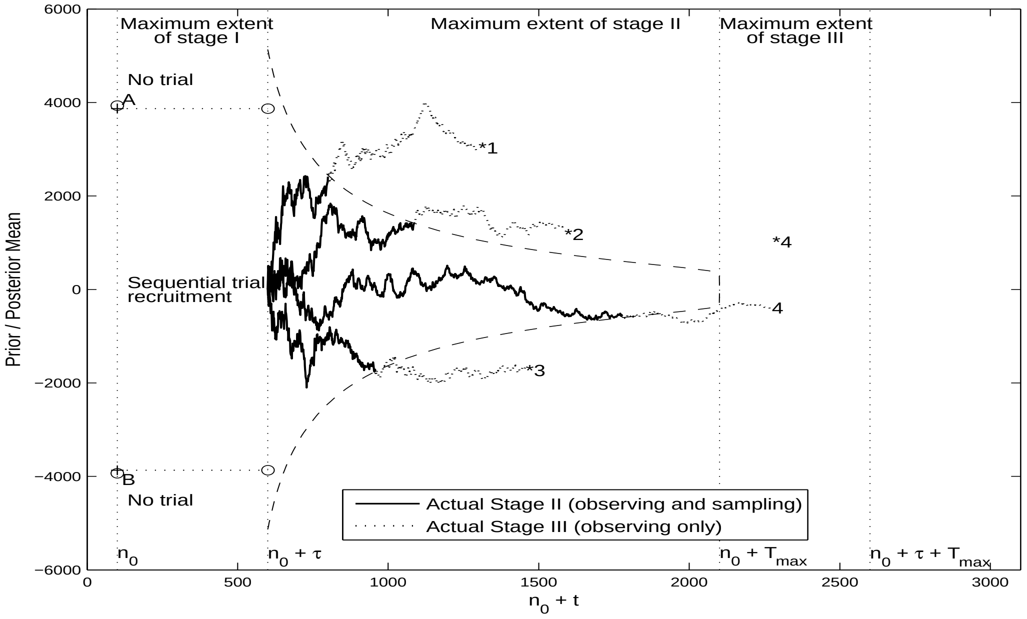
We ran Monte Carlo simulations to study the behavior of sample paths and to explore other operating characteristics of the Optimal Bayes Sequential policy. Each sample path is generated by making an independent draw for the drift W using Eq. (4a), followed by generation of the X_i given that draw, to generate the sample path μ_t using Eqs. (3) and (4b).

Figure 2(a) shows four sample paths for a prior mean of $\mu_0 \approx 17$ lying in the ‘sequential trial’ region, meaning that it is optimal to proceed to stage II. The realised values of W_i, i = 1,. . . ,4 are indicted by ‘*’s. Stage II sections of the paths are marked as continuous lines. When a stage II path first touches the upper or lower Stage II stopping boundary (dashed line), it is optimal to proceed to stage III, at which point the paths are shown as dotted lines.

For the baseline scenario of Figure 2(a), each path is briefly described. For path $i = 1$, $w_1 > 0$ and the path crosses the upper stopping boundary soon after entering stage II. The new technology is selected upon the conclusion of stage III, because the posterior mean is positive. This is the correct decision, given that $w_1 > 0$. The same applies for $i = 2$, with the posterior mean hitting the upper boundary a little later than for path $i = 1$. For path 3, $w_3 < 0$ and the new technology is rejected once all pipeline subjects have been observed (again the correct decision).

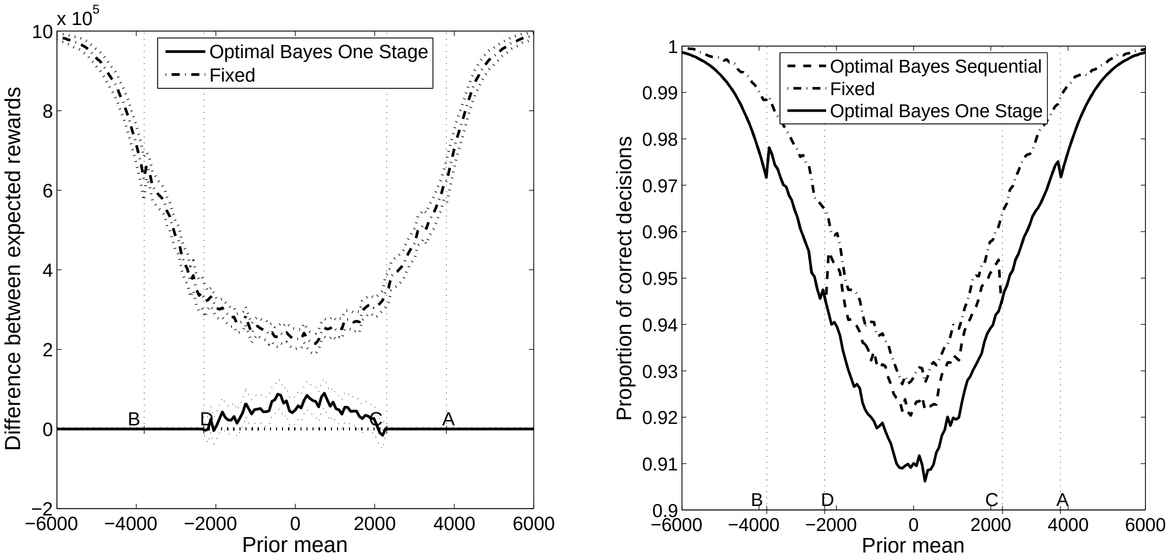


(a) Baseline scenario (large τ relative to $T_{\max}$). KEY: '+' thresholds A, B, C, D delineate the ranges for 'no trial'/'fixed trial'/'sequential trial'; '○' Optimal Bayes One Stage sample sizes.



(b) Comparator scenario (smaller τ relative to $T_{\max}$). KEY: '+' thresholds A and B delineate the ranges for 'no trial'/'sequential trial'.

Figure 2: Optimal Bayes Sequential policy, together with four stage II / III paths of the posterior mean with prior mean $\mu_0 \approx 17$. KEY: '*' value of the sampling mean w_i for each path i; '—-' path of posterior mean when in stage II; '$\cdots$' path of posterior mean when in stage III.



(a) 'Net gain' (difference between expected rewards of Optimal Bayes Sequential and comparator policies).

(b) Proportion of simulations which select the best technology.

Figure 3: Operating characteristics for baseline illustration.

Path 4 results in an incorrect decision: $w_4 > 0$, but the path exits stage II close to $T_{\max}$ (on the lower free boundary) and, upon conclusion of stage III, the new technology is rejected because the posterior mean is negative after all pipeline subjects have been observed.

For the baseline scenario, Figure 3(a) plots the difference between the averages of the realised rewards obtained from the Optimal Bayes Sequential policy and those from the two alternative policies: the 'Fixed' policy, which always makes $T_{\max} = 2000$ pairwise allocations, and the Optimal Bayes One Stage policy of Eq. (15). Thick lines represent the averages, dashed lines 95% confidence intervals. For convenience, we call this difference the 'net gain'. Figure 3(b) shows the proportion of iterations which make the correct adoption decision. To derive each graph, we chose 400 equally-spaced values of μ_0 in the range $[-6000, 6000]$ and, for each value of μ_0, the results from 15,000 sample paths were averaged.[3]

Figure 3(a) shows that, as expected, the Optimal Bayes Sequential policy outperforms the other two policies when judged according to net gain. Compared with the 'Fixed' policy, the greatest gains for the Optimal Bayes Sequential policy may be seen at extreme values of the prior mean, which is unsurprising: there is little point running a trial with a large fixed sample size when the prior mean is far from zero. The net gain is lowest around $\tilde{\mu}_0 = I/P$ (= 0). These findings are reversed for the Optimal Bayes One Stage policy, which yields an optimal sample size equal to that of the Optimal Bayes Sequential policy to the left of D and to the right of C, so that there is no difference between the expected rewards. Between D and C, the Optimal Bayes Sequential policy benefits from the arrival of observations on the pipeline subjects to update the

[3] The jaggedness in these sample averages are on the order of the size of the confidence intervals for the plotted means. This is true for other Monte Carlo results below, so confidence interval lines for other Monte Carlo results are suppressed for clarity. Smooth curves are obtained from the partial differential equation methods of Appendix S.

prior distribution and offers the flexibility to stop stage II according to the value of the posterior mean and variance. No such luxury is available for the Fixed policy, which commits to sampling and observing a predetermined number of observations regardless of the information that arrives.

Figure 3(b) plots the estimate of the probability that each of the three sampling policies correctly selects the best technology for the baseline scenario. The probabilities for the Optimal Bayes Sequential policy and the Optimal Bayes One Stage policy coincide to the left of D and the right of C for the reasons just stated. Between D and C, the Optimal Bayes Sequential policy is superior because it permits a decision rule which sequentially updates the information set and stops the sequential trial if the evidence is sufficiently convincing. The Fixed policy performs best because it guarantees the highest amount of information. This is obtained at a cost, however (refer to Figure 3(a)).

4 Application: drug-eluting stents

Moses et al. (2003) and Cohen et al. (2004) compared the performance of drug-eluting stents (DES, the new technology) with bare metal stents (BMS, the standard) for the treatment of complex coronary stenoses using percutaneous coronary intervention (PCI) in the ‘SIRIUS’ trial. The authors randomised 1058 patients to either DES or BMS and measured clinical outcomes, resource use and costs over a one year follow-up period. The trial’s recruitment phase lasted approximately seven months, so it did not include a period during which observations on the primary end points were being made while recruitment was taking place.

We compare the performance of the Optimal Bayes Sequential policy of section 2 with what is

Parameter	*Value*	*Source*
n_0	20	Assumption
c	\$200	Assumption
λ	\$50000/QALY	Assumption
σ_X	\$17538.00	Derived from Cohen et al. (2004), choice of λ
ρ	0.01	Assumption
P	2000000	Assumption
I	\$0	Assumption
$T_{\max}$	2000	Assumption
End point	QALY	Cohen et al. (2004)
Delay in observing the primary end point	1 year	Cohen et al. (2004)
δ_{on}	0	Assumption
Study size (number of pairs)	529	Cohen et al. (2004)
Duration of recruitment period	7 months	Cohen et al. (2004)
Equivalent annual rate of recruitment R	907	Derived from Cohen et al. (2004)
τ	907	Derived from Cohen et al. (2004)

Table 1: Parameter values used for the drug-eluting stents application

a Fixed policy with the same sample size as the SIRIUS study (529 patient pairs) and the Optimal Bayes One Stage policy. For the purposes of this section, we concentrate on the cost and QALY results at one year of follow-up that are reported in Cohen et al. (2004). We also investigate how the policies change when the cost of sampling changes, and assess how the optimal choice of trial design changes as the value of τ changes by changing the rate of recruitment to the trail. This section is intended to illustrate how our model may be populated with data from an actual application; it is not intended to represent a comment on the particular health technology.

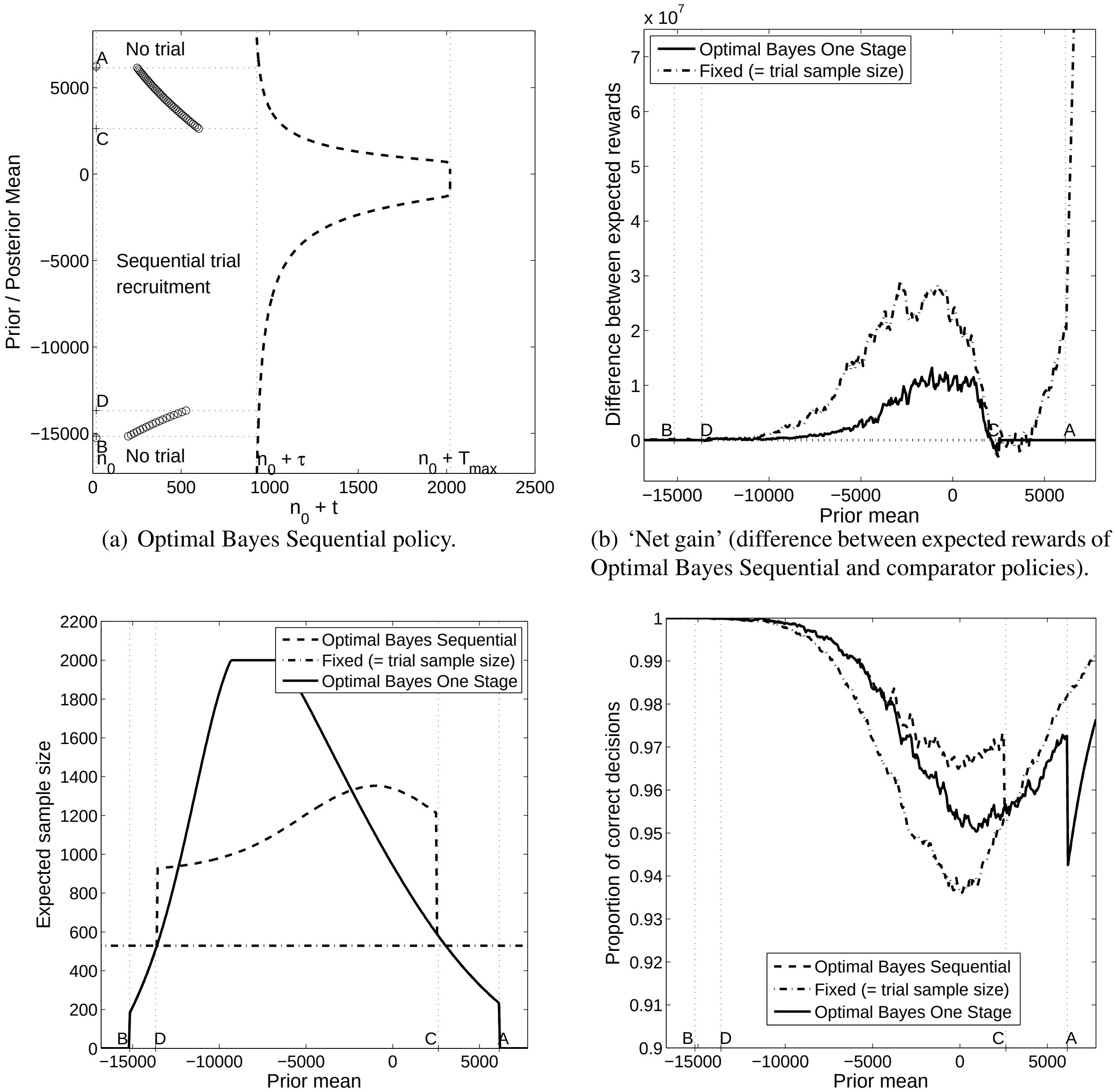
Parameter values are reported in Table 1 and, where possible, are derived from Moses et al. (2003) and Cohen et al. (2004). Where this was not possible, the value is marked as an assumption. The value of σ_X is derived from point estimates in Cohen et al. (2004) and the assumption $\lambda = \$50000/\text{QALY}$. For our model, we choose to set $T_{\max} = 2000$, which is higher than the annual rate of recruitment to the study (calculated to be 529 × 12/7 = 907 patient pairs). For simplicity, a zero switching cost is assumed ($I = 0$) and the effective sample size in the prior distribution is assumed to be $n_0 = 20$. In contrast to the illustrations of section 3, the discount rate is chosen to be greater than zero ($\rho = 0.01$).

Figure 4(a) shows the optimal stopping boundaries. As in Figure 2(a), the '$\circ$'s in Figure 4(a) indicate that, within the ranges AC and DB, the Optimal Bayes Sequential policy fixes a sample size that is neither too close to 0, nor too close to τ. This is because, at points A and B, the expected value of taking a small, fixed, sample size is more than offset by the cost of postponing a decision that is implied by starting to experiment (by experimenting, one must wait for at least a year before making an adoption decision, and rewards are discounted). Between points C and D it is optimal to go to stage II.

In the absence of discounting, Prop. 2.4 implies that there is a greater expected reward when the mean is above $I/P = 0$ than when it is below that value by the same amount. With a positive discount rate, the expected benefit of continued sampling is penalized more for values of the posterior mean above I/P than for values below it by the same absolute amount. Consequently, the upper stage II boundary in Figure 4(a) is shifted down relative to the upper stage II boundary for the case of zero discounting (latter not shown). The change is greater in magnitude than the corresponding change for the lower boundary, resulting in asymmetric stopping boundaries.

This asymmetry about $\mu_0 = I/P$ is reflected in the plots of the differences between the expected rewards of the Optimal Bayes Sequential policy and the two comparators (Figure 4(b)) and the average sample sizes (Figure 4(c)). In Figure 4(b), the negative values (indicating that the Optimal Bayes Sequential policy performs less well than its comparator) close to point C are due to the noise in the Monte Carlo estimates and the fact that the expected sample sizes of the three comparators are quite close to each other in the vicinity of point C (Figure 4(c)). Figure 4(c) also illustrates the 'jumps' in the expected number of pairwise allocations for the different trials at points A, B, C and D. This is as expected, given the discussion of Figure 4(a) above.

For a value of the prior mean close to zero, Figure 4(b) shows that the expected net gain of the Optimal Bayes Sequential policy over the Fixed policy is approximately \$20m and over the Optimal Bayes One Stage it is around \$10m. Not apparent from Figure 4(b), because of the scaling, is that fact that, for extremely low values of the prior mean, the difference in rewards between the Optimal Bayes Sequential policy and the Fixed policy converges to a positive value equal to the net discounted value of sampling patients under the Fixed policy. This is because



(a) Optimal Bayes Sequential policy.

(b) 'Net gain' (difference between expected rewards of Optimal Bayes Sequential and comparator policies).

(c) Expected sample size.

(d) Proportion of simulations which select the best technology

Figure 4: Optimal Bayes Sequential policy, net gain, expected sample size, and proportion of simulations which select the best technology for stents application of section 4.

the pessimistic prior mean implies that it is optimal not to start the trial under the Optimal Bayes Sequential policy, whereas the Fixed policy will always make 529 pairwise allocations and then reject the new technology (providing the value of the prior mean is low enough). As $\mu_0 \to \infty$, the net gain of the Optimal Bayes Sequential policy over the Fixed policy grows without bound.

This is because, with a very optimistic prior mean, it is optimal to adopt immediately rather than commission a trial under the Optimal Bayes Sequential policy, whereas the Fixed policy is committed to incurring trial costs and discounting rewards.

Importantly, Figures 4(b) and 4(c) show that the range of the prior mean over which the Optimal Bayes Sequential policy performs best in terms of the net gain is also the range over which its expected sample size is close to, or greater than, the expected sample sizes of the Optimal Bayes One Stage and Fixed policies. This highlights the Optimal Bayes Sequential policy's maximisation of the expected reward of the trial as defined in Eq. (6), an objective which requires achieving the sample size which appropriately balances the benefits to patients with the costs of learning (as opposed to minimisation of the sample size). With this objective in mind, for some values of μ_0 the Optimal Bayes Sequential policy may sample more than the Optimal Bayes One Stage and the Fixed policies, for other values of μ_0, it may sample the same as, or less.

An estimate of the probability of correct selection after all outcomes have been observed is shown in Figure 4(d). This shows that the Optimal Bayes Sequential policy is superior to both the Fixed and the Optimal Bayes One Stage policies in the region DC, where the probability of selecting correctly is no lower than 0.96, and is similar to the comparators to the left of D. Over the majority of the range CA, Figure 4(d) shows that the Fixed policy performs best because it tends to sample more (Figure 4(c)). Figure 4(d) shows that the proportion of correct decisions for the Optimal Bayes Sequential policy drops at points C and A. These drops mirror the jumps in the sample sizes that occur at those points (Figure 4(c)).

The estimate of the probability of a 'decision reversal' (commonly defined as the probability that the adoption decision that would have been made at the time of stopping to sample sequentially is overturned once all realisations on pipeline subjects have arrived) did not exceed 0.03 in this example.

4.1 Comparative analysis

Changes in parameter values such as the marginal cost of sampling or the discount rate can change the shape of the continuation set. Figure 5 shows the result of increasing c from $c =$ \$200 to $c =$ \$5000: a higher sampling cost is shown to shrink the stage II continuation set. Furthermore, it increases the range of values of the prior mean over which it is optimal not to enter stage II. The opposite effect may be seen by increasing the size of the population to benefit, P (figure not shown). In simulations, the higher is P, the wider is the stage II continuation set and the more attractive is the sequential trial.

4.2 Delay influences optimal design

The illustrations in Figures 2(a) and 2(b) showed how changing the time delay, holding all other parameters constant, changed τ and the Optimal Bayes Sequential boundaries. A change to τ may also be achieved by holding the time delay constant and changing the recruitment rate R. When applying the model to a particular clinical trial, changing the recruitment rate is the more

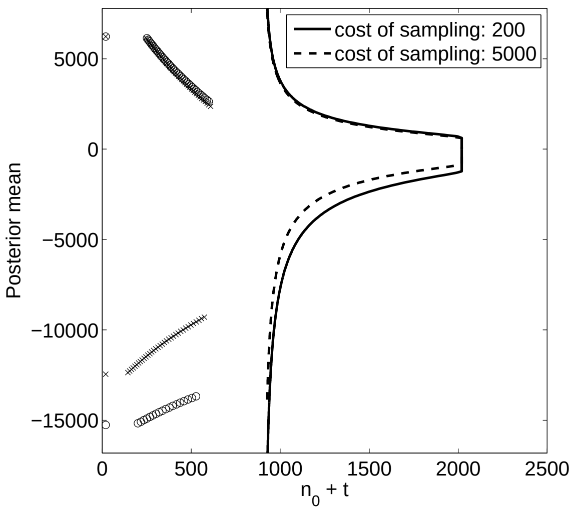


Figure 5: The effect of changing c on the optimal policies. KEY: '$\circ$' c =\$200; '$\times$' c =\$5000

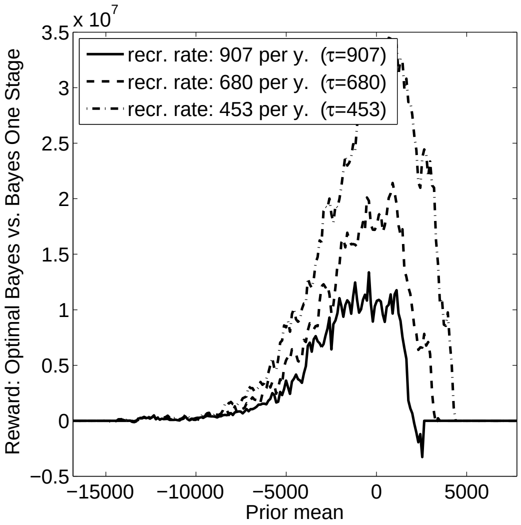


Figure 6: The difference between expected rewards of Optimal Bayes Sequential and Optimal Bayes One Stage policies for several recruitment rates.

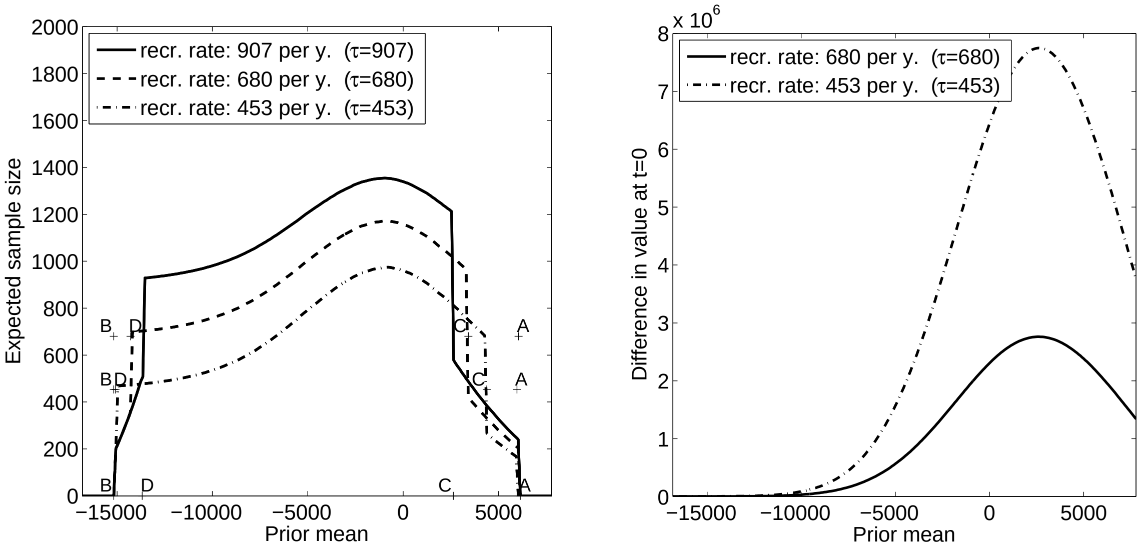
natural way to think about changing τ, because changing τ by changing the time delay could also change the sampling mean W and variance σ_X^2 if they are a function of the time to follow-up.

Figure 6 shows the net gain of the Optimal Bayes Sequential policy over the Optimal Bayes One Stage policy, assuming a time delay of one year and a recruitment rates of 907 (the baseline case), 680 and 453 patient pairs per year (reductions of 25% and 50%, respectively). The net gain increases as the recruitment rate (and hence τ) decreases. We also found a higher net gain at lower recruitment rates when comparing the expected reward of the Optimal Bayes Sequential policy with that of the Fixed policy with 529 patient pairs. These findings are consistent with results in Hampson and Jennison (2013), albeit for a different objective function, in that fewer observations in the pipeline were associated with a greater benefit of sequential sampling.

Figure 7 shows the expected sample sizes and rewards for these three different recruitment rates. The effect on the expected sample size is shown to vary according to the value of the prior mean: when the prior mean is such that, for each of the three simulations, it is optimal to enter Stage II, the higher is τ, the higher is the expected sample size (Figure 7(a)). The comparison is less straightforward for other values of the prior mean. Figure 7(b) shows the higher is the recruitment rate, the higher is the expected reward in this example.

5 Discussion

We have presented a Bayesian decision-theoretic model of sequential experimentation with delay and applied it to the field of medical statistics. Given prior information, we have established the optimal policy and have thereby addressed, in part, the following important questions for trial design: is there value in conducting a sequential trial, in which recruitment to the trial takes place while end points are being observed, or should recruitment to the trial conclude before the first



(a) Expected sample sizes for Optimal Bayes Sequential policy for several recruitment rates.

(b) Expected rewards of Optimal Bayes Sequential policy when recruitment rate is 907 per year minus expected reward with 680 and 453 per year

Figure 7: The effect of changing τ by changing the recruitment rate.

end point is observed? If a sequential trial is conducted, how should stopping rules be adjusted so as to account for delay? How does the length of the delay or rate of patient recruitment influence the optimal trial design? Is it worth carrying out any trial at all?

Results show that the value of the prior mean, combined with the amount of delay that exists between allocating a patient to treatment and observing the primary end point, plays a key role in determining whether it is optimal to carry out no trial at all, a trial of a fixed sample size, or a sequential trial. In particular, the higher is the delay, the less attractive is the sequential design over the fixed design and optimal Bayes One Stage policy. Thus our results are consistent with the results of Hampson and Jennison (2013), even though the objective functions of the two models differ. At the same time, the application highlights that the Optimal Bayes Sequential policy's focus on maximising the total expected reward of the trial and the adoption decision, taking into account both the sampling costs and benefits to patients, results in the highest net gain over the competing policies when the expected sample size is close to, or greater than, the expected sample sizes of those policies.

Monte Carlo simulations which compare the performance of the discrete time model with alternative designs show that it is superior in terms of the net gain and that it performs well with regards to the probability of correctly selecting the best alternative even though the optimal stopping boundaries were derived from a continuous time approximation. Clearly, the precise performance of such a model will depend on the particular trial of interest.

Directions for future research are numerous. A major assumption is that the population variance, σ_X^2, is known. The model also assumes that only one alternative treatment is being considered, it does not incorporate intermediate outcomes that are correlated with the primary end point. We assumed that all data is observed before an adoption decision is made. Future work includes relaxing these assumptions.

Acknowledgements

We thank participants in the 2014 annual conference of the Royal Statistical Society, Sheffield, the 2014 IDEAL (Integrated Design and Analysis of Clinical Trials) consortium Annual Scientific Meeting, Paris, and 2015 seminars at the Chicago Booth School of Business and at INSEAD and Jacco Thijssen, for comments on earlier versions of this work. Priming monies came from fund RIS6.1(2013–2014) of the Department of Economics and Related Studies, University of York. The usual disclaimer applies.

A Mathematical proofs for the discrete time model

Proof of Prop. 2.1. Referring to Eq. (5), conditioning on T and W, and using the tower property of conditional expectation gives

$$\begin{aligned} V^{\pi}(\mu_0, n_0) &= \mathbb{E}_{\pi}\left[\mathbb{E}\left[\left\{\sum_{t=0}^{T-1}\frac{-c+\delta_{\text{on}}X_{t+1}}{(1+\tilde{\rho})^t}\right\} + \frac{\mathbf{1}_{\mathcal{D}=n}(PW-I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}}\middle| T, W\right]\middle|\mu_0, n_0\right] \\ &= \mathbb{E}_{\pi}\left[\left\{\sum_{t=0}^{T-1}\frac{-c+\delta_{\text{on}}W}{(1+\tilde{\rho})^t}\right\} + \frac{\mathbf{1}_{\mathcal{D}=n}(PW-I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}}\middle|\mu_0, n_0\right]. \end{aligned} \tag{16}$$

Let $\mathcal{K}_\pi$, $\mathcal{S}_\pi$ and L_π be as defined in Eq. (13). Adding and subtracting $\bar{V}(\mu_0, n_0)$ from the right hand side of Eq. (16) gives Eq. (12). $\mathcal{K}_\pi$ is not negative for offline learning ($\delta_{\text{on}} = 0$) because $c \geq 0$. For online learning, it is strictly positive for $W < c$ and 0 otherwise. $\mathcal{S}_\pi$ is not negative because $(W-c)^+ \geq 0$. Note that $\tilde{\rho} \geq 0$, $\tau \geq 0$, $T \geq 0$ together imply that $(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)} \geq 1$. Also observe that $(PW-I)^+ \geq 0$ and $(PW-I)^+ \geq \mathbf{1}_{\mathcal{D}=n}(PW-I)$. Hence, $(PW-I)^+ \geq \mathbf{1}_{\mathcal{D}=n}(PW-I)/(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}$ independent of the sign of $\mathbf{1}_{\mathcal{D}=n}(PW-I)$. Thus, $L_\pi \geq 0$. □

Proof of Prop. 2.2. From Eq. (12), it is clear that a policy π maximises $V^\pi(\mu_0, n_0)$ if and only if it minimises the sum of the expectations of $\mathcal{K}_\pi$, $\mathcal{S}_\pi$ and L_π. Minimisation of the expectation of $\mathcal{K}_\pi + \mathcal{S}_\pi + L_\pi$ is equivalent to a sequential optimisation problem to minimise the expected discounted costs when the cost is $\mathcal{K}_\pi$ when the action to continue is chosen, and is $\mathcal{S}_\pi + L_\pi$ if the action to stop sampling is chosen.

Because $T_{\max}$ is finite, this equivalent problem is a finite horizon stochastic optimal control problem in discrete time (Bertsekas and Shreve, 1978, Definition 8.1). The (F$^+$) condition of Bertsekas and Shreve (1978, Chapter 8, page 192) holds because $\mathcal{K}_\pi$ and $\mathcal{S}_\pi + L_\pi$ are not negative on every sample path. Proposition 8.1 of Bertsekas and Shreve (1978) states that there exists for the problem a nonrandomised Markovian policy within the set of all policies. Because of the Markovian nature of Bayes' rule, Bertsekas and Shreve (1978, Corollary 8.1.1) then show that an additional dependence of the state evolution on the past can not bring additional expected reward. Bertsekas and Shreve (1978, Prop. 8.2) then formally justify the claim that the value function in

Eq. (6) satisfies Bellman's equation. This proves that $V^{\pi^*}(\mu_0, n_0) = B(\mu_0 n_0, 0)$. The claim that $V^{\pi}(\mu_0, n_0) = V^{\pi^*}(\mu_0, n_0)$ then follows directly from Bertsekas and Shreve (1978, Prop. 8.5). □

Proof of Prop. 2.3. Given $\tilde{\rho} > 0$, $T_{\max,\tilde{\rho}}$ is finite even though the total potential sampling size $T_{\max}$ is unbounded. Therefore the expectation in Eq. (11) is finite. Given $T_{\max} = \infty$, we have an infinite horizon stochastic optimal control problem in discrete time.

The proof is like that of the proof of Prop. 2.2, except that the infinite horizon results of Bertsekas and Shreve (1978, Chapter 9) are employed. Because $\mathcal{K}_\pi$, $\mathcal{S}_\pi$, and L_π are all non-negative, the (P) assumption of Bertsekas and Shreve (1978, Chapter 9, page 214) is satisfied for the minimisation of the expectation of $\mathcal{K}_\pi + \mathcal{S}_\pi + L_\pi$. The (P) assumption is the infinite horizon analog of the (F$^+$) condition for the finite horizon. Because of the Markovian nature of Bayes' rule, Bertsekas and Shreve (1978, Prop. 9.1) show that an additional dependence of the state evolution on the past can not bring additional expected reward. Prop. 9.8 of Bertsekas and Shreve (1978) then justifies the claim that the value function in Eq. (6) satisfies Bellman's equation. That is, $V^{\pi^*}(\mu_0, n_0) = B(\mu_0 n_0, 0)$. The claim that $V^{\pi}(\mu_0, n_0) = V^{\pi^*}(\mu_0, n_0)$ follows directly from Bertsekas and Shreve (1978, Prop. 9.12). □

Proof of Prop. 2.4. We first prove claim (ii), that $B((I/P + \Delta\mu)n_t, t) = B((I/P - \Delta\mu)n_t, t) - P\Delta\mu$ for $t = 0, 1, \ldots, T_{\max}$, in two steps: we show that the first term in the maximand of Eq. (9a), $G(\cdot)$, satisfies a similar relation involving $\Delta\mu$ for all t, so that $B(\cdot)$ satisfies the claimed relationship when $t = T_{\max}$. Then an induction argument in $-t$ will prove the result for $t = 0, 1, \ldots, T_{\max} - 1$. Claims (i) and (iii) will follow from the proof of claim (ii).

The expectation in Eq. (8), which defines $G(\cdot)$, simplifies due to the Gaussian inference process: if $Z \sim \mathcal{N}(\xi, \sigma^2)$ then $\mathbb{E}[Z^+] = \sigma[\phi(\xi') + \xi'\Phi(\xi')]$, where $\xi' = \xi/\sigma$ and ϕ and Φ are, respectively, the probability density function and the cumulative distribution function of a standard normal random variable (DeGroot, 1970). Moreover,

$$\mathbb{E}[(-Z)^+] = \sigma[\phi(-\xi') - \xi'\Phi(-\xi')] = \sigma[\phi(\xi') + \xi'\Phi(\xi')] - \xi. \tag{17}$$

Consider an arbitrary state, (μ_t, t), and pick $\Delta\mu$ so that $\mu_t = I/P + \Delta\mu$. Then $P\mu_t = I + P\Delta\mu$ and $y_t = (I/P + \Delta\mu)n_t$. We define some additional notation to help us proceed. Define $\tilde{\mu}_t = I/P - \Delta\mu$, so that $P\tilde{\mu}_t = I - P\Delta\mu$ and $\tilde{y}_t = (I/P - \Delta\mu)n_t$. Recall that, given information to time t, $Z_{T,\min(T,\tau)}$ has mean $\mu_t = y_t/n_t$. Let $\tilde{Z}_{T,\min(T,\tau)}$ be the predictive distribution for the posterior mean given stopping at time t with $Y_t = \tilde{y}_t$. Then:

$$\mathbb{E}[PZ_{T,\min(T,\tau)} - I \mid Y_T = y_t, T = t] = P\Delta\mu, \tag{18a}$$

$$\text{and } \mathbb{E}[P\tilde{Z}_{T,\min(T,\tau)} - I \mid Y_T = y_t, T = t] \;=\; -P\Delta\mu. \tag{18b}$$

Define $\sigma^2 = \mathrm{Var}[(PZ_{T,\min(T,\tau)} - I) \mid Y_T, T = t]$, which depends on t but not on Y_T.

Given the assumption $\tilde{\rho} = 0$, we may simplify Eq. (8) using Eqs. (18a) and (18b):

$$\begin{aligned} G(y_t, t) &= \mathbb{E}[(PZ_{T,\min(T,\tau)} - I)^+ \mid Y_T = y_t, T = t] \\ &= \sigma[\phi(P\Delta\mu/\sigma) + (P\Delta\mu/\sigma)\Phi(P\Delta\mu/\sigma)] \\ &= \sigma[\phi(-P\Delta\mu/\sigma) + (-P\Delta\mu/\sigma)\Phi(-P\Delta\mu/\sigma)] - (-P\Delta\mu) \\ &= \mathbb{E}[P(\tilde{Z}_{T,\min(T,\tau)} - I/P)^+ \mid Y_T = \tilde{y}_t, T = t] + P\Delta\mu \\ &= G(\tilde{y}_t, t) + P\Delta\mu. \end{aligned} \tag{19}$$

Thus, given Eq. (9b), $B(y_t, t) - P\Delta\mu = B(\tilde{y}_t, t)$ for $t = T_{\max}$.

Suppose now that $B(y_{t+1}, t{+}1) - P\Delta\mu = B(\tilde{y}_{t+1}, t{+}1)$ for some $t \in \{\tau, \tau{+}1, \ldots, T_{\max}{-}1\}$, so that the claimed relation holds at time $t+1$. We now show that this relation holds at time t by proving a similar relation for each maximand which determines $B(\cdot)$.

By Eq. (19), the first maximand on the right hand side of Eq. (9a) differs by $P\Delta\mu$ when evaluated at $y_t = (I/P + \Delta\mu)n_t$ and $\tilde{y}_t = (I/P - \Delta\mu)n_t$, as desired. Let B_2 be the second maximand in the right hand side of Eq. (9a). If $t \geq \tau$, let $\hat{X}$ be a normal random variable with mean 0 and variance σ_X^2. If $\delta_{\text{on}} = 0$ and $\tilde{\rho} = 0$ then

$$\begin{aligned}
B_2(y_t, t) - B_2(\tilde{y}_t, t) &= \mathbb{E}_\pi[B(y_t + y_t/n_t + (X_{t+1-\tau} - y_t/n_t), t+1) \mid Y_T = y_t, T = t] \\
&\quad -\mathbb{E}_\pi[B(\tilde{y}_t + \tilde{y}_t/n_t + (X_{t+1-\tau} - \tilde{y}_t/n_t), t+1) \mid Y_T = \tilde{y}_t, T = t] \\
&= \mathbb{E}[B(y_t + y_t/n_t + \hat{X}, t+1) \mid Y_T = y_t, T = t] \\
&\quad -\mathbb{E}[B(\tilde{y}_t + \tilde{y}_t/n_t - \hat{X}), t+1) \mid Y_T = \tilde{y}_t, T = t] \qquad (20) \\
&= \mathbb{E}[P(\Delta\mu + \hat{X}/n_{t+1}) \mid Y_T = y_t, T = t] = P\Delta\mu. \qquad (21)
\end{aligned}$$

The first line follows by the definition of $B(\cdot)$. The second line follows by the symmetry of the distribution of $\hat{X}$ about 0. The third line follows because both $y_t + y_t/n_t + \hat{X}$ and $\tilde{y}_t + \tilde{y}_t/n_t - \hat{X}$ differ from $n_{t+1}I/P$ by the same amount, $n_{t+1}\Delta\mu + \hat{X}$. This 'coupling' of expectations implies the posterior means in the two expectations of Eq. (20) change by $\Delta\mu + \hat{X}/n_{t+1}$ in going from time t to time $t+1$, given $\hat{X}$. The fact that $B(\cdot)$ satisfies the claimed relation at time $t+1$, by the induction assumption, then implies Eq. (21).

If $t < \tau$, then it is straightforward to show that $B_2(y_t, t) - B_2(\tilde{y}_t, t) = P\Delta\mu$ from Eq. (10).

By mathematical induction, $B(y_t, t) - P\Delta\mu = B(\tilde{y}_t, t)$ for $t = T_{\max}, T_{\max} - 1, \ldots, 1, 0$. This justifies claim (ii). By setting $t = 0$ and by recalling Eq. (14), we obtain $V^{\pi^*}(I/P + \Delta\mu, n_0) - P\Delta\mu = V^{\pi^*}(I/P - \Delta\mu, n_0)$. This justifies claim (i).

We have shown (a) that the first maximand differs by the same amount (by $-P\Delta\mu$) when evaluated at (y_t, t) and $(\tilde{y}_t, t)$, and (b) that the second maximand in Eq. (9a) differs by the same amount (by $-P\Delta\mu$) when evaluated at (y_t, t) and $(\tilde{y}_t, t)$. Thus, either the first maximand is larger for both (y_t, t) and $(\tilde{y}_t, t)$ or the second maximand is not smaller for both (y_t, t) and $(\tilde{y}_t, t)$. Recall that (y_t, t) and $(\tilde{y}_t, t)$ correspond to the points (μ_t, t) and $(\tilde{\mu}_t, t)$, respectively. This relation among the maximands implies that (μ_t, t) is in the interior of the continuation set when $(\tilde{\mu}_t, t)$ is in the continuation set, and vice versa. This proves claim (iii). □

Proof of Prop. 2.5. The special case of $\tilde{\rho} = 0$, $c > 0$, $\delta_{\text{on}} = 0$ and $\tau = 0$ corresponds exactly to a special case of the undiscounted sampling selection problem of Eq. (4) in Chick and Frazier (2012) for comparing $k = 1$ alternatives with unknown mean with an alternative whose mean reward is known to be 0. Prop. 3 of Chick and Frazier (2012) thus justifies $T \leq \Upsilon \equiv 1 + (P^2\sigma_X^2)/(2\pi c^2) - n_0$ almost surely, when $\tau = 0$ under the stated conditions.

The proof of that result is based on properties of the effective sample size in the posterior distribution for the unknown mean at a given time t, and shows that sampling beyond the stated bound does not give sufficient additional expected reward. Because the number of outcomes observed when there is delay is not more than τ fewer than when there is no delay (formally, $t - (n_t - n_0) \leq \tau$), then $(n_T - n_0) \leq \Upsilon$ implies that $T \leq \Upsilon + \tau$, as desired. □

Appendix S is found in the **Online Supplementary Material**.

References

Armitage, P. (1975). *Sequential Medical Trials*. Blackwell Oxford.

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

Berry, D. A. and Ho, C. (1988). One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics*, 44:219–227.

Bertsekas, D. and Shreve, S. (1978). *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, Belmont, MA.

Broglio, K. R., Connor, J. T., and Berry, S. M. (2014). Not too big, not too small: a Goldilocks approach to sample size selection. *Journal of Biopharmaceutical Statistics*, 24(3):685–705.

Brown, J., McElvenny, D., Nixon, J., Bainbridge, J., and Mason, S. (2000). Some practical issues in the design, monitoring and analysis of a sequential randomized trial in pressure sore prevention. *Statistics in Medicine*, 19:3389–3400.

Burman, C.-F. (2013). Discussion of the paper by Hampson and Jennison. *Journal of the Royal Statistical Society, Series B*, 75(1):47.

Chernoff, H. (1961). Sequential tests for the mean of a normal distribution. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 79–91.

Chick, S. E. and Frazier, P. I. (2012). Sequential sampling for selection with economics of selection procedures. *Management Science*, 58(3):550–569.

Chick, S. E. and Gans, N. (2009). Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437.

Cohen, D. J., Bakhai, A., Shi, C., Githiora, L., Lavelle, T., Berezin, R., and others (2004). Cost-effectiveness of sirolimus-eluting stents for treatment of complex coronary stenoses. *Circulation*, 110:508–514.

Connor, J. T., Broglio, K. R., Durkalski, V., Meurer, W. J., and Johnston, K. C. (2015). The stroke hyperglycemia insulin network effort (SHINE) trial: an adaptive trial design case study. *Trials*, 16(72).

DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York, First edition.

Draper, D. (2013). Discussion of the paper by Hampson and Jennison. *Journal of the Royal Statistical Society, Series B*, 75(1):48.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007a). Bayesian evaluation of group sequential clinical trial designs. *Statistics in Medicine*, 26:1431–1449.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007b). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, 26:5047–5080.

European Medicines Agency (2006). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. Scientific Guidelines.

Hampson, L. and Jennison, C. (2013). Group sequential tests for delayed responses. *Journal of the Royal Statistical Society, Series B*, 75:3–54.

Jennison, C. and Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. Chapman and Hall, Boca Raton, Florida, first edition.

Lewis, R. J., Lipsky, A. M., and Berry, D. A. (2007). Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: a frequentist evaluation. *Clinical Trials*, 4:5–14.

Moses, J. W., Leon, M. B., Popma, J. J., et al. (2003). Sirolimus-eluting stents versus standard stents in patients with stenosis in a native coronary artery. *The New England Journal of Medicine*, 349(14):1315–1323.

NICE (2012). Measuring effectiveness and cost-effectiveness: the QALY. National Institute for Health and Clinical Excellence. London.

Pertile, P., Forster, M., and La Torre, D. (2014). Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies. *Journal of the Royal Statistical Society, Series A*, 177(2):419–438.

Porter, M. E. (2010). What is value in health care? *New England Journal of Medicine*, 363:2477–2481.

Stallard, N. and Todd, S. (2011). Seamless phase II/III designs. *Statistical Methods in Medical Research*, 20(6):623–634.

United States Food and Drug Administration (2010). Guidance for the use of Bayesian statistics in medical device clinical trials. Guidance for Industry and FDA Staff.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. J. Wiley and Sons, Chichester, second edition.

Willan, A. and Kowgier, M. (2008). Determining optimal sample sizes for multi-stage randomized clinical trials using value of information methods. *Clinical Trials*, 5:289–300.

S Online Supplemental Material

This document provides supplementary material for the paper "A Bayesian Decision-Theoretic Model of Sequential Experimentation with Delayed Response". References to sections and equations not found in this supplement may be found in that paper.

Solving for the discrete time optimal stopping policy π^* in Eq. (6) is challenging. An approximate solution may be obtained by exploiting continuous time methods which are in the spirit of the work of Chernoff (1961) and other papers cited below. The numerical solution of the associated optimal stopping problem proves to be useful for the numerical results of sections 3 and 4.

Informally, we construct a diffusion whose joint statistics, when sampled at a set of integer times, match those of the original discrete process. We then allow stopping times to be continuous on this diffusion, thereby constructing a CT optimal stopping problem in Eq. (24) below.

Appendix S.1 defines the diffusion and writes the continuous time analog of the discrete time optimal stopping problem in Eq. (6). It also derives the continuous time analog of Bellman's equation using a Taylor expansion of that equation and Ito's lemma. That analog turns out to be a free boundary problem for a heat equation. Appendix S.2 justifies why the solution to the free boundary problem determines the optimal stopping boundaries and continuation set of the continuous time analog of our stopping problem. Appendix S.3 describes computational techniques for approximating the stopping boundaries of the optimal policy π^*_{CT} and value function for the CT problem with general τ. Section 3 and 4 use π^*_{CT} to approximate the optimal policy π^* for the discrete time problem. Appendix S.3 also references some theoretical results for asymptotic approximations which are useful for the special case of $\tau = 0$.

This supplement also summarises the notation of the main paper for convenience in Table 2. Finally, connections of the modeling approach in the main paper to the multi-armed bandit (MAB) literature are provided in Appendix S.4.

S.1 Continuous time analog of discrete time problem

In order to approximate the optimal delayed sequential sampling problem specified by Eq. (6) in continuous time, the definitions of the time t, the sum Y_t defined in Eq. (3), the induced filtration $\mathcal{F}_t$, and a policy π must be suitably modified. Given such a modification, the definitions of n_t, μ_t and $Z_{t,u}$ are naturally extended to be real valued for real valued t and u, as is the definition of the terminal reward function G of the discrete time problem. We shall refer to the CT discount rate per patient pair, ρ, and recall that it is linked to the discrete time discount rate by $(1+\tilde{\rho})^t = e^{t\rho}$.

Assume that $t \in [0, T_{\max}+\tau]$, that $t = 0$ is the time when the decision maker posits a prior distribution for W, and that sequential sampling commences in the instant immediately following $t = 0$. Let the cumulative sum $Y_t = \sum_{i=1}^{(t-\tau)^+} X_i$ accumulate as a diffusion, that is, a shifted and scaled Brownian motion which has the appropriate joint marginal distribution when sampled at integer times:

$$\mathrm{d}Y_t = W\mathrm{d}t + \sigma_X \mathrm{d}V_{t-\tau}, \quad \tau \le t \le T_{\max} + \tau, \tag{22}$$

where V_u for $u \ge 0$ is a standard Brownian motion and the drift W is inferred with Bayes' rule as the process Y is observed. The delay implies that $Y_t = Y_0 = \mu_0 n_0$ for $t \in [0, \tau]$ and that $V_{\lfloor u \rfloor}$ is a diffusion approximation for the first $\lfloor u \rfloor$ observations.

Define $\mathcal{F}_{\text{CT}} = (\mathcal{F}_{\text{CT},t})_{t\in[0,T_{\max}+\tau]}$ as the natural filtration of the process $\{Y_t\}_{t\in[0,T_{\max}+\tau]}$. By construction, it has the same joint distribution as the discrete time process above at sets of integer valued times in $[0, T_{\max}+\tau]$, as desired.

Define the CT policy π_{CT} as a continuous-valued sample size, T_{CT} (a stopping time with respect to the filtration $\mathcal{F}_{\text{CT}}$ taking values in $[0, T_{\max}]$), and a decision $\mathcal{D}_{\text{CT}} \in \{n, s\}$ for an alternative to select after all outcomes on pipeline subjects are observed. Define Π_{CT} as the set of all policies $\pi_{\text{CT}} = (T_{\text{CT}}, \mathcal{D}_{\text{CT}})$ such that T_{CT} is measurable with respect to $\mathcal{F}_{\text{CT}}$ and $\mathcal{D}_{\text{CT}}$ is measurable with respect to $\mathcal{F}_{\text{CT},\mathbf{1}_{T_{\text{CT}}>0}(T_{\text{CT}}+\tau)}$. The expected reward of a policy $\pi_{\text{CT}} \in \Pi_{\text{CT}}$ is

$$V_{\text{CT}}^{\pi}(\mu_0, n_0) = \mathbb{E}_{\pi_{\text{CT}}}\left[\int_0^{T_{\text{CT}}} \frac{-c}{e^{t\rho}}\,\mathrm{d}t + \int_0^{T_{\text{CT}}} \frac{\delta_{\text{on}}}{e^{t\rho}}\,\mathrm{d}Y_{t+\tau} + \frac{\mathbf{1}_{\mathcal{D}_{\text{CT}}=n}(PW - I)}{e^{\mathbf{1}_{T_{\text{CT}}>0}(T_{\text{CT}}+\tau)\rho}}\,\middle|\, \mu_0,\, n_0\right]. \tag{23}$$

The apparent asymmetry between $\mathrm{d}Y_{t+\tau}$ in Eq. (23) and the summand X_t in Eq. (5) is explained because increments in Y at time $t+\tau$ are due to decisions made at time t.

The optimal delayed sequential sampling problem in continuous time is defined formally as that of finding a policy $\pi_{\text{CT}}^* \in \Pi_{\text{CT}}$ such that

$$V_{\text{CT}}^{\pi_{\text{CT}}^*}(\mu_0, n_0) = \sup_{\pi_{\text{CT}} \in \Pi_{\text{CT}}} V_{\text{CT}}^{\pi_{\text{CT}}}(\mu_0, n_0). \tag{24}$$

In what follows, we show that the optimal solution to this problem is characterised by a continuation set, $\mathcal{C} \subseteq \mathbb{R} \times [0, T_{\max})$ such that, when $(y_t, t) \in \mathcal{C}$ on a realisation, sampling should continue, and otherwise sampling should stop and stage III entered. On the boundary of $\mathcal{C}$, one is indifferent between continuing and stopping. We propose using $\mathcal{C}$ for the continuous time problem to approximate the optimal continuation set for the discrete time problem when evaluating whether or not to continue sampling at integer t.

S.1.1 Continuous time approximation during stage III

The expected reward upon stopping is extended to continuous time by rewriting G in Eq. (8) as:

$$G(y_t, t) = e^{-\mathbf{1}_{t>0}\tau\rho}\,\mathbb{E}[(PZ_{T_{\text{CT}},\min(T_{\text{CT}},\tau)} - I)^+ \mid \mathcal{F}_{\text{CT},t}]. \tag{25}$$

It is optimal to choose $\mathcal{D}_{\text{CT}} = n$ if $PZ_{T_{\text{CT}},\min(T_{\text{CT}},\tau)} > I$ and $\mathcal{D}_{\text{CT}} = s$ otherwise.

S.1.2 Continuous time approximation during stage II

We turn to the problem of solving for the CT approximation in stage II. The problem will reduce to one of establishing a 'free boundary' in (y_t, t) space, which determines $\mathcal{C}$ for $t \in [\tau, T_{\max}]$. The key to doing so is to rewrite Eq. (9) in continuous time. Following Chernoff (1961), a CT diffusion model approximation of Bellman's equation in Eq. (9) is:

$$\begin{aligned} B_{\text{CT}}(y_t, t) = \max\Big\{G(y_t, t), \lim_{h\downarrow 0}\Big[-c + \delta_{\text{on}}(y_t/n_t)\Big]h& \qquad\qquad\qquad\qquad\qquad\qquad \text{(26a)}\\ + e^{-h\rho}\mathbb{E}_{\pi_{\text{CT}}}[B_{\text{CT}}(Y_{t+h}, t+h) \mid \mathcal{F}_{\text{CT},t}]\Big\},& \quad t \in [\tau, T_{\max}),\\ B_{\text{CT}}(y_{\max}, T_{\max}) = G(y_{\max}, T_{\max}),& \qquad\qquad\qquad\qquad\qquad\qquad \text{(26b)} \end{aligned}$$

where $h > 0$ is a small time step and B_{CT} is the continuous time equivalent of B.

States $(y_t, t) \in \mathbb{R} \times [\tau, T_{\max})$ such that the second term in the maximand of Eq. (26a) exceeds the first are in $\mathcal{C}$. States where the first term in the maximand strictly exceeds the second are in the complement of $\mathcal{C}$. Given the assumptions of the model (refer to Eq. (3)), the increment $U_t = Y_{t+h} - y_t$

has a $\mathcal{N}\left(hy_t/n_t, \sigma_X^2[h+h^2/n_t]\right)$ distribution. Equating the left hand side of Eq. (26a) and the second maximand, expanding the second maximand in a Taylor series expansion, and applying Ito's Lemma gives

$$\begin{aligned} B_{\text{CT}}(y_t,t) &= -c+\delta_{\text{on}}\left(y_t/n_t\right)h+(1-h\rho) \\ &\quad \times\mathbb{E}[B_{\text{CT}}(y_t,t)+U_tB_{\text{CT},y}(y_t,t)+hB_{\text{CT},t}(y_t,t)+U_t^2B_{\text{CT},yy}(y_t,t)/2]+o(h) \end{aligned} \tag{27}$$

for (y_t,t) in $\mathcal{C}$, and where the second index in the subscript for B_{CT} refers to derivatives. Collecting terms and simplifying gives the following partial differential equation describing the change in B_{CT} for stage II of the problem:

$$0=-c-\rho B_{\text{CT}}+B_{\text{CT},t}+(B_{\text{CT},y}+\delta_{\text{on}})(y_t/n_t)+\sigma_X^2B_{\text{CT},yy}/2. \tag{28}$$

The boundary of the optimal continuation set, $\partial\mathcal{C}$, is characterised by a free boundary condition and a so-called smooth pasting condition where the two terms in the maximisation in Eq. (26a) are equal and are smoothly matched (Chernoff, 1961). Here, these conditions are:

$$B_{\text{CT}}(y,t)=G(y,t) \text{ on } \partial\mathcal{C} \text{ (free boundary);} \tag{29a}$$

$$B_{\text{CT},y}(y,t)=G_y(y,t) \text{ on } \partial\mathcal{C} \text{ (smooth pasting).} \tag{29b}$$

Equations (28) and (29) are similar to the partial differential equation (PDE) in Pertile et al. (2014) and Chick and Gans (2009), with three notable exceptions:

1. the posterior mean is now multiplied by $B_{\text{CT},y}+\delta_{\text{on}}$ instead of $B_{\text{CT},y}$, to reflect the potential inclusion of online learning;

2. the independent variable in the PDE, $t\in[\tau,T_{\max}]$, is the cumulative number of pairwise allocations made, which no longer coincides with the number of outcomes observed because of the delay.

3. the reward for stopping, G, is defined to include the expected reward from observing the outcomes for the pipeline subjects and acting optimally.

S.1.3 Continuous time approximation during stage I

The first term in the maximand of Bellman's equation in Eq. (10) for discrete time is extended to continuous time by using Eq. (25). The other term can be handled by observing that if $T=u\in(0,\tau]$, then the expected reward of sampling is naturally modeled in continuous time by

$$H(u)\equiv\int_0^u e^{-t\rho}(-c+\delta_{\text{on}}\mu_0)\mathrm{d}t=\begin{cases}(-c+\delta_{\text{on}}\mu_0)u & \text{if } \rho=0\\ (-c+\delta_{\text{on}}\mu_0)(1-e^{-u\rho})/\rho & \text{if } \rho>0.\end{cases}$$

The reward function at time $t=0$ is found by checking all $T_{\text{CT}}=u\in[0,\tau]$ and $T_{\text{CT}}>\tau$:

$$B_{\text{CT}}(y_0,0)=\max\Big\{\sup_{u\in[0,\tau]}\left\{H(u)+e^{-u\rho}G(y_0,u)\right\},\, H(\tau)+e^{-\tau\rho}B_{\text{CT}}(y_0,\tau)\Big\}. \tag{30}$$

This determines the continuation set on $\mathbb{R}\times[0,\tau)$: let u_y be the smallest u which maximises the supremum in Eq. (30) when $y_0=y$. Such a u exists: $[0,\tau]$ is compact and the maximands are continuous in u. Then $(y,t)\in\mathcal{C}$ for all $t\in[0,u_y)$.

S.1.4 Analysis and computation of the PDE for stage II

The analysis for the discrete time stopping problem in Eq. (6) consisted of proving that the solution to the discrete time Bellman's equation determines an optimal policy π^*. In a similar way, the crux of the optimal solution to the CT problem in Eq. (24) can be reduced to the solution of the continuous time version of Bellman's equation, the free boundary problem in Eq. (28) subject to the implicit boundary conditions in Eq. (29), for $t \in [\tau, T_{\max}]$. The optimal solution of the CT problem for $t \in [0, \tau]$ in stage I and for stage III are more straightforward to analyze.

S.2 Analysis for optimal stopping and the free boundary problem

The link between optimal stopping times of a continuous time Markovian process and the free boundary problem has been formalized in two different ways. First, Bather (1970) characterized the solution of a broad class optimal stopping problems for Brownian motion. He showed that, under certain conditions, it is optimal to continue sampling for continuous time stopping problems for Brownian motion when the state is in the interior of the continuation set of a suitably defined free boundary problem for a heat equation, and to stop sampling otherwise. For the stage II and stage III analysis, our stopping problem and free boundary problem fall into the class of problems considered by Bather (1970). For example, the conditions for which Bather's results hold can be verified for the special case $\rho = 0$, $c > 0$, $\delta_{\text{on}} = 0$ by noting that (a) this special case corresponds to a finite horizon version of the example in Chernoff (1961) which provided motivation for Bather (1970), and (b) our terminal reward $G(y, t) \geq 0$ and its derivatives are continuous (except near $t = 0$ when $\tau > 0$, which may cause a discontinuity for stage I analysis).

The above arguments justify that Bellman's equation for the continuous time problem gives the optimal expected reward, given $T \geq \tau$. To handle $T \in [0, \tau)$, observe that sampling costs and online learning benefits in Eq. (30) and Eq. (23) are equal for all $T = u \in [0, \tau]$, and that terminal rewards are identical because $Y_t = y_0$ on that interval. Moreover, the relevant costs through time τ and the expected reward to go given $T > \tau$ are also equal, if $\mathcal{D}_{\text{CT}}$ is as defined in section S.1.1. Thus, $V_{\text{CT}}^{\pi_{\text{CT}}^*}(\mu_0, n_0)$ in Eq. (24) equals $B_{\text{CT}}(\mu_0 n_0, 0)$ in Eq. (30) for the special case $\rho = 0$, $c > 0$, $\delta_{\text{on}} = 0$. Moreover, the optimal stopping time $T_{\text{CT}} \leq \tau$ if the first term in the maximand in Eq. (30) exceeds the second, and $T > \tau$ if the opposite is true. In that second case, the solution to the free boundary problem determines the boundaries of the optimal continuation set for stage II.

A second approach can be used to show continuity of the solution to the free boundary problem and a form of uniqueness to handle the remaining case of $\rho > 0$: general dynamic programming principles for continuous time stochastic control, such as the analysis of Pham (2009, Section 5.2.1). For this case too, then, $V_{\text{CT}}^{\pi_{\text{CT}}^*}(\mu_0, n_0) = B_{\text{CT}}(\mu_0 n_0, 0)$.

The above arguments justify situations when the free boundary problem defines the optimal stopping boundary but do not describe its shape. The next section describes how the free boundary PDE problem in Eqs. (28) and (29) may be solved using numerical methods.

S.3 Numerical solution of PDE free boundary problem

The solution to the free boundary PDE problem which describes the continuation set and its boundary, $\partial\mathcal{C}$, have been studied for some interesting special cases which do not have sampling delays. We use those principles here for computing the solution to the free boundary problem which solves Eq. (24).

For stage II, we solve the PDE with a trinomial tree in (μ_t, t) coordinates by recursing backward from time $n_0 + T_{\max}$, the point at which stage III must be entered, to time $n_0 + \tau$ in steps of size Δt that are specified by the analyst. For a more detailed description of the principles for doing so, see Arlotto et al. (2010), who did so for a project on employment decisions which had Bayesian learning, sampling costs, and online learning, but not the other variations of the model in section 2. See also Chernoff and Petkau (1986), Brezzi and Lai (2002) and Chick and Gans (2009) for discussions of computing solutions to related problems in a reverse time scaling (based on the transformation $u = 1/\gamma t$) which take advantage of some standardizations which are more difficult in our context due to the generality (and hence number of parameters) in our model.

It may seem odd to approximate a discrete time optimal stopping problem with a PDE in continuous time, and to solve that PDE with the time discretization of a trinomial tree. The reason is that time discretization of the trinomial tree is typically different from the integer time step of the original stopping problem. Increasing the number of steps in the trinomial tree per patient pair sampled improves the normal approximation to the observations of the patient pairs. Numerical error can be controlled by refining the grid of the trinomial tree.

Easily computed numerical approximations are available for some special cases with $\tau = 0$. We validated our code to verify that the relevant bounds from those papers correspond well to the solutions found for the code in this paper when τ is small (data not shown).

For the special case $\tau = 0$, $T_{\max}$ arbitrarily large, $c = 0$, $\rho > 0$ and online learning ($\delta_{\text{on}} = 1$), Brezzi and Lai (2002) show the relationship of this problem to the multi-armed bandit problem with normally distributed rewards and mean reward which is inferred through time. They present theory to characterize $\partial\mathcal{C}$ asymptotically as $t \to \infty$ and as $t \to 0$, and give an easy-to-compute approximation for the upper boundary of $\partial\mathcal{C}$.

In a study of a context similar to Brezzi and Lai (2002), except that the bandits are stoppable and there is offline learning with a fixed benefit of selecting a technology ($\delta_{\text{on}} = 1$), with $\tau = 0$, $T_{\max}$ arbitrarily large, $c = 0$, $\rho > 0$, Chick and Gans (2009) show a structural relationship between the boundary of the continuation sets for the cases of online and offline learning in the setting studied by Brezzi and Lai (2002). Chick and Gans (2009, Online Companion) give a numerically useful approximation to the upper boundary of $\partial\mathcal{C}$ for this special case.

For the case $\tau = 0$, $T_{\max}$ arbitrarily large, $c > 0$, $\rho = 0$, and offline learning, Chick and Frazier (2012) provide a numerical approximation for the upper and lower boundaries of $\mathcal{C}$.

The Matlab code used to compute the optimal stopping boundaries for stage I and stage II sampling is available at `https://github.com/sechick/htadelay`.

S.4 Related multi-armed bandit (MAB) literature.

We also note several connections to the MAB literature, which also addresses questions which are related to the current work. A central theme is the exploration-exploitation tradeoff between learning about alternatives with unknown mean performance and exploiting the performance of alternatives with better-known performance, when the goal is to maximise expected discounted rewards. Bellman (1956) studied this with backward induction techniques. Gittins and Jones (1974) proposed an index policy for bandit problems of a particular structure and showed optimality. Glazebrook (1979) extended this framework to allow for “stoppable” bandits. The discrete time optimal stopping problem in Eq. (6) can be considered to be a one-armed stoppable bandit.

The MAB framing has also been useful in adaptive trial design Berry and Eick (1995) use adaptive assignment rules to balance the goal of treating patients within a trial effectively with the goal of correctly identifying the relative efficacy of the treatments. Ahuja and Birge (2015) explore this further by assessing the role of group size in adaptive group sequential designs for each of these objectives for Bernoulli end points. Those works model how to assign patients to different treatments (which we do not) but do not study for how long the trial should run or explore the economics of the trial plus adoption decision. Related non-clinical applications include assortment planning in retail (Caro and Gallien, 2007), employee performance assessment for hiring and retention decisions (Arlotto et al., 2014), and interactive marketing (Bertsimas and Mersereau, 2007). Much of that work does not account explicitly for delays. Hardwick et al. (2006) account for Poisson arrivals and exponential delays, and develop heuristics to minimise patient loss. Caro and Yoo (2010) show that certain bandit problems with stationary random delays satisfy an indexability criterion as long as the delayed responses are observed in the same order they are allocated (as is the case here) and compute indices for a beta-binomial model.

References

Ahuja, V. and Birge, J. (2015). Response-adaptive designs for clinical trials: simultaneous learning from multiple patients. *European Journal of Operations Research*. in press.

Arlotto, A., Chick, S. E., and Gans, N. (2014). Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, 60(1):110–129.

Arlotto, A., Gans, N., and Chick, S. E. (2010). Optimal employee retention when inferring unknown learning curves. In *Proceedings of the 2010 Winter Simulation Conference*, pages 1178–1188.

Bather, J. A. (1970). Optimal stopping problems for Brownian motion. *Adv. Appl. Probab.*, 2:259–286.

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

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

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

Brezzi, M. and Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27:87–108.

Caro, F. and Gallien, J. (2007). Dyamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292.

Caro, F. and Yoo, O. S. (2010). Indexability of bandit problems with response delays. *Probability in the Engineering and Informational Sciences*, 24:349–374.

Chernoff, H. (1961). Sequential tests for the mean of a normal distribution. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 79–91.

Chernoff, H. and Petkau, A. J. (1986). Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Stat. Comput.*, 7(1):46–59.

Chick, S. E. and Frazier, P. I. (2012). Sequential sampling for selection with economics of selection procedures. *Management Science*, 58(3):550–569.

Chick, S. E. and Gans, N. (2009). Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., editor, *Progress in Statistics*, pages 241–266, Amsterdam. North-Holland.

Glazebrook, K. D. (1979). Stoppable families of alternative bandit processes. *J. Appl. Prob.*, 16:843–854.

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

Pertile, P., Forster, M., and La Torre, D. (2014). Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies. *Journal of the Royal Statistical Society, Series A*, 177(2):419–438.

Pham, H. (2009). *Continuous-time Stochastic Control and Optimization with Financial Applications*. Springer.

Parameter	Definition
$P \in \mathbb{R}_{>0}$	Number of patients to receive the new or existing technology once the adoption decision is made
$I \in \mathbb{R}_{\geq 0}$	Fixed cost of switching to the new technology from standard technology
$X \in \mathbb{R}$	Random variable: incremental effectiveness or incremental net monetary benefit (INMB) of the new technology over standard
$\sigma_X^2 \in \mathbb{R}_{>0}$	Variance of X (assumed known)
$W \in \mathbb{R}$	Expected value of X (assumed unknown, hence a random variable)
$\mu_0 \in \mathbb{R}$, $\sigma_0^2 \in \mathbb{R}_{>0}$	Mean and variance of prior distribution for W
$n_0 = \sigma_X^2/\sigma_0^2$	Effective sample size of prior distribution
$\tau \in \mathbb{Z}_{\geq 0}, \tau < T_{\max}$	Number of pairwise allocations that are made between making a pairwise allocation and observing its realisation (assumed known)
$\mathbf{1}_F$	Indicator function for the event F
$T_{\max} \in \mathbb{Z}_{>0}$	Maximum number of pairwise allocations which can be made
$\mathbb{T} \equiv \{0, 1, \dots, T_{\max}\}$	Set of potential patient pairs to be allocated
$\mathbb{T}_{\mathrm{I}} \equiv \{0, 1, \dots, \tau - 1\}$	Recruitment of trial participants only
$\mathbb{T}_{\mathrm{II}} \equiv \{\tau, \dots, T_{\max} - 1\}$	Parallel recruitment, observation of outcomes and Bayes updating is possible
$\bar{\mathbb{T}} \equiv \{0, 1, \dots, T_{\max} + \tau\}$	Set of times where pairwise allocations and/or a treatment choice may be made
$t \in \bar{\mathbb{T}}$	For times $t \in \mathbb{T}$, t is the number of pairwise allocations so far. For $t > \tau$, observations of outcomes may occur
$a_t \in \mathcal{A} \equiv \{1, 0\}$	Actions denoting whether to make a pairwise allocation ($a_t = 1$) or not ($a_t = 0$) for $t \in \mathbb{T}_{\mathrm{I}} \cup \mathbb{T}_{\mathrm{II}}$
$T \in \mathbb{T}$	Time at which pairwise allocations cease to be made (hence, number of pairwise allocations made before stopping)
$\mathcal{D} \in \{n, s\}$	The decision to adopt the new technology or control after all pairwise allocations are observed, at time $\mathbf{1}_{T>0}(T + \tau)$
π	A sequence of sampling decisions and an adoption decision
$\mathcal{F} = (\mathcal{F}_t)_{t \in \bar{\mathbb{T}}}$	Natural filtration defined by the observations seen through time t
Π	Set of policies where $T \leq T_{\max}$ is a stopping time with respect to $\mathcal{F}$ and $\mathcal{D}$ is $\mathcal{F}_{\mathbf{1}_{T>0}(T+\tau)}$-measurable
$\mathbb{E}_\pi, \mathbb{E}$	Expectation with respect to filtration induced by π, expectation independent of π
$n_t = n_0 + (t - \tau)^+$	Effective sample size of posterior distribution as tth pairwise allocation is made
$Y_t = \mu_0 n_0 + \sum_{i=1}^{(t-\tau)^+} X_i$	Sum of prior mean weighted by n_0 and X_i, where second term on RHS equals zero if upper limit on $\sum$ equals zero
$\mu_t = Y_t/n_t$	Posterior mean of W when t pairwise allocations have been made
$Z_{t,u} \equiv \mathbb{E}[\mu_{t+u} \mid \mathcal{F}_t]$	Random variable for posterior mean to be obtained, given t pairwise allocations have been made, assuming u pairwise allocations remain in the pipeline
$c \in \mathbb{R}_{>0}$	Recruitment and monitoring cost of making one more pairwise allocation
ρ_{year}	Annual discount rate (assuming continuous compounding)
$R \in \mathbb{R}_{>0}$	Annual rate of recruitment to the trial
$\tilde{\rho} = e^{\rho_{\text{year}}/R} - 1$	Discrete time discount rate at level of patient pair
$\rho = \rho_{\text{year}}/R$	Continuous time discount rate at level of patient pair
δ_{CE}	1 = 'cost effectiveness' (rewards include benefits and costs of treatment); 0 = 'effectiveness only' (rewards include benefits but not costs of treatment)
λ	Monetary value of one unit of effectiveness (e.g., £30,000 / QALY)
δ_{on}	1 = 'online learning' (rewards of treatment accrue to trial participants and those influenced by adoption decision); 0 = 'offline learning' (rewards of treatment do not accrue to trial participants)

Table 2: Notation.