White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/110768/

Version: Accepted Version

Article:
Chick, Stephen, Forster, Martin and Pertile, Paolo (2017) A Bayesian decision-theoretic model of sequential experimentation with delayed response. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY. pp. 1439-1462. ISSN: 1369-7412

https://doi.org/10.1111/rssb.12225

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

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

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

November 30, 2016

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 published clinical trials.

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 clinical trial designs are well documented (Armitage, 1975; Berry, 1985; Whitehead, 1997; Jennison and Turnbull, 1999). 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 policy interest in sequential and adaptive trial designs (European Medicines Agency, 2006; US FDA, 2010). 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 movement (Porter, 2010) calls for increased attention to the health benefits obtained for a given level of expenditure.

Hampson and Jennison (2013) provide an overview of the emerging literature on group sequential trial design with delay. They derive new, frequentist, delayed response group sequential tests for two-treatment comparisons of mean efficacy which minimise the trial's expected sample size, subject to meeting prespecified type I and type II error probabilities. The authors derive their optimal stopping rules by solving Bayes decision problems using dynamic programming. 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) suggests that solving a Bayesian decision-theoretic model, whose utility function measures outcomes on a clinically relevant scale (such as the Quality Adjusted Life Year, or QALY, e.g. see NICE 2012) could provide real gains over the type I/type II error probability scale. Burman (2013) also advocates use of a Bayesian decision-theoretic framework which measures explicitly the cost of sampling and the value of trial results and which incorporates a prior distribution for the expected outcome.

We implement the recommendations of Draper (2013) and Burman (2013) by proposing a Bayesian decision-theoretic model for experimental design which compares two health technologies and which is fully sequential (as opposed to one which uses a fixed sample size, or which allocates patients in a group sequential manner). Outcomes are observed with a specified delay and converted to economic values using standard cost utility analysis; costs of sampling and switching technologies are explicitly accounted for. The model selects the policy which maximises the expected benefits of the technology adoption decision that is made on the basis of experimental data, minus the expected cost of the sequential experiment itself. Expected benefits can include the treatment effect accruing to trial participants as the study progresses, as

well as expected benefits which accrue to the patients who benefit from the adoption decision. 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.

Section 2 presents and solves the model for the case of a known sampling variance. Section 3 highlights the main features of the optimal policy using an illustrative example. Section 4 considers the case of an unknown sampling variance. Section 5 presents an application using data from a published clinical trial for drug-eluting stents and assesses the operating characteristics of the model's optimal policy. Directions for future research are presented in section 6. Appendix A and the Online Supplementary Material (OSM, Appendix S) provide mathematical proofs and further details on our methods, as well as an additional application. Matlab code which implements these computations is provided at `https://github.com/sechick/htadelay`.

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 $E_{\text{N}} \in \mathbb{R}$ if a patient is assigned to the new technology and $E_{\text{S}} \in \mathbb{R}$ if the patient is assigned to the standard one. The patient-level costs of using each technology are the random variables $C_{\text{N}} \in \mathbb{R}_{\geq 0}$ and $C_{\text{S}} \in \mathbb{R}_{\geq 0}$. 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 (for example, NICE values one Quality Adjusted Life Year (QALY) at between £20,000 and £30,000). 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:

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

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 (we discuss unknown σ_X^2 in section 4). The prior distribution for W is assumed to be $\mathcal{N}(\mu_0, \sigma_0^2)$. The choice of σ_0^2 might be guided by expert judgment and available

related data. For example, choice of the so-called 'effective sample size' of the prior distribution, $n_0 = \sigma_X^2/\sigma_0^2$, might be guided by the sample size of a related Phase II clinical trial or pilot study. O'Hagan et al. (2006) provide additional guidance on specifying prior probability distributions.

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 for 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. Future benefits and costs may be down-weighted using a discount rate, defined at the level of one pairwise allocation as $\tilde{\rho} \geq 0$.

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.

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 $a_t = 0$ denotes choosing not to make a pairwise allocation. It is assumed that, once pairwise allocations cease to be made, sampling cannot be restarted: 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 for 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 $\mathcal{D} \in \{\mathrm{N}, \mathrm{S}\}$ as the decision concerning whether to choose the new technology (N) or the standard (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 $T + \tau$, $T > 0$, if $a_0 = 1$, because of the delay.

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 $\mathbf{1}_{\mathcal{D}=\mathrm{N}}(PW - I)$. 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.

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 $t \in \bar{\mathbb{T}}$. Due to the delay, $\mathcal{F}_t = \mathcal{F}_0$ for $t \in \{0, 1, \ldots, \tau\}$. Define variables tracking the 'effective sample size' (as a function of the number of realisations of pairwise allocations) in the posterior distribution for W, and the 'effective cumulative sum' of realisations, given information available to time $t \in \bar{\mathbb{T}}$,

$$n_t = n_0 + (t-\tau)^+, \text{ and } Y_t = \mu_0 n_0 + \sum_{i=1}^{(t-\tau)^+} X_i, \tag{2}$$

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

The posterior distribution for W at time t has a normal distribution

$$W \mid \mathcal{F}_t \sim \mathcal{N}(\mu_t, \sigma_X^2/n_t), \text{ where:} \tag{3a}$$

$$\mu_t = Y_t/n_t. \tag{3b}$$

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.

A policy π defines a mapping $f(y_t, t) : \mathbb{R} \times \mathbb{T} \backslash \{T_{\max}\} \to \mathcal{A}$ from states to deciding whether to make a pairwise allocation, which in turn determines T. A policy π also specifies the choice of the new technology or standard, $\mathcal{D} \in \{\text{N}, \text{S}\}$, as discussed above.

By construction, T is a stopping time with respect to the filtration $\mathcal{F}$ taking values in $\mathbb{T}$; $\mathcal{D}$ 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 with respect to the measure induced by π on the sequence of observations and decisions, and $\mathbb{E}$ to indicate the expectation when it does not depend on π. Table 1 summarizes the principal notation.

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:

$$V^\pi(\mu_0, n_0) = \mathbb{E}_\pi \left[\left\{ \sum_{t=0}^{T-1} \frac{-c + \delta_{\text{on}} X_{t+1}}{(1+\tilde{\rho})^t} \right\} + \frac{\mathbf{1}_{\mathcal{D}=\text{N}}(PW - I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}} \Big| \mu_0, n_0 \right]. \tag{4}$$

Here, $\delta_{\text{on}} = 1$ if the benefits to patients participating in the trial (in addition to the P post-trial patients) are to be included in the reward function (known as 'online learning'). $\delta_{\text{on}} = 0$ if rewards for participants are not to be included in the reward function ('offline learning'). Traditional trials set $\delta_{\text{on}} = 0$ implicitly. 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

$$V^{\pi^*}(\mu_0, n_0) = \sup_{\pi \in \Pi} V^\pi(\mu_0, n_0). \tag{5}$$

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.

$P \in \mathbb{R}_{>0}$	Number of patients to receive technology once adoption decision 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/net monetary benefit of new over standard
$\sigma_X^2 \in \mathbb{R}_{>0}$	Variance of X
$W \in \mathbb{R}$	Expected value of X
$\mu_0 \in \mathbb{R}$, $\sigma_0^2 \in \mathbb{R}_{>0}$	Mean and variance of prior distribution for W
$n_0 = \sigma_X^2/\sigma_0^2$	Effective sample size of prior distribution
$\tau \in \mathbb{Z}_{\geq 0}$, $\tau < T_{\max}$	Delay in observing realisation of pairwise allocation (in pairwise allocations)
$T_{\max} \in \mathbb{Z}_{>0}$	Maximum number of pairwise allocations which can be made
$\mathbb{T} \equiv \{0, 1, \ldots, T_{\max}\}$	Set of potential patient pairs to be allocated
$\mathbb{T}_{\mathrm{I}} \equiv \{0, 1, \ldots, \tau - 1\}$	Recruitment of trial participants only
$\mathbb{T}_{\mathrm{II}} \equiv \{\tau, \ldots, T_{\max} - 1\}$	Parallel recruitment and Bayes updating possible
$\bar{\mathbb{T}} \equiv \{0, 1, \ldots, T_{\max} + \tau\}$	Set of times when pairwise allocations and/or treatment choice may be made
$a_t \in \mathcal{A} \equiv \{1, 0\}$	Action to make a pairwise allocation ($a_t = 1$) or not ($a_t = 0$), $t \in \mathbb{T}_{\mathrm{I}} \cup \mathbb{T}_{\mathrm{II}}$
$T \in \mathbb{T}$	Time at which pairwise allocations cease to be made (stopping time)
$\mathcal{D} \in \{\mathrm{N}, \mathrm{S}\}$	Decision to adopt new or standard, having observed all realisations
π	Sequence of sampling decisions and an adoption decision
Π	Set of policies where $T \leq T_{\max}$
$\mathcal{F} = (\mathcal{F}_t)_{t \in \bar{\mathbb{T}}}$	Natural filtration defined by the observations seen through time t
$\mathbb{E}_\pi; \mathbb{E}$	Expectations: with respect to filtration induced by π; independent of π
$n_t = n_0 + (t - \tau)^+$	Effective sample size of posterior distribution as tth pairwise allocation is made
$Y_t = \mu_0 n_0 + \sum_{i=1}^{(t-\tau)^+} X_i$	Cumulative sum for posterior mean
$\mu_t = Y_t/n_t$	Posterior mean of W when t pairwise allocations have been made
$Z_{t,u}$	Posterior mean to be obtained, given $\mathcal{F}_t$ and u 'pipeline' observations to arrive
$c \in \mathbb{R}_{>0}$	Recruitment cost of making one more pairwise allocation
$R \in \mathbb{R}_{>0}$	Annual rate of recruitment to the trial
$\tilde{\rho} \geq 0$	Discrete time discount rate at level of one pairwise allocation
λ	Monetary value of one unit of effectiveness (e.g., £30,000 / QALY)
δ_{on}	1 = 'online learning'; 0 = 'offline learning'

Table 1: Table of principal notation.

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 justifies how an optimal policy $\pi^* \in \Pi$ can be determined from Bellman's equation and provides further results for two special cases. Section 2.3 introduces the method that we use to solve 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: $u = T$ 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 the information to time t and that u outcomes are still to be observed. Then in our setting:

$$Z_{t,u} \equiv \mathbb{E}[\, W \mid \mathcal{F}_t, X_{t-u+1}, X_{t-u+2}, \ldots, X_t] \sim \mathcal{N}\left(\mu_t, \frac{\sigma_X^2}{n_t}\frac{u}{(n_t + u)}\right). \tag{6}$$

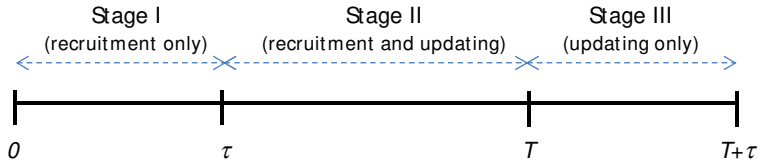


Figure 1: Stages of the problem with stopping time T and delay τ.

An adoption decision can be made immediately if $T = 0$ because no trial takes place. If $T > 0$, the last of the observations on pipeline subjects will be observed τ time units after stopping. Once all outcomes on pipeline subjects are observed, it will be optimal to adopt the new technology ($\mathcal{D} = \text{N}$) if $PZ_{T,\min(T,\tau)} - I > 0$ and the standard one ($\mathcal{D} = \text{S}$) 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:

$$G(y_t, t) = (1 + \tilde{\rho})^{-\mathbf{1}_{t>0}\tau} \mathbb{E}[\,(PZ_{T,\min(T,\tau)} - I)^+ \mid Y_T = y_t, T = t\,]. \tag{7}$$

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 can be 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. (7) with the expected reward of making that allocation, observing the outcome of the next pairwise allocation in the pipeline, and continuing to behave 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:

$$\begin{aligned} B(y_t, t) = \max \Big\{ & G(y_t, t),\ -c + \delta_{\text{on}}(y_t/n_t) \\ & + (1 + \tilde{\rho})^{-1}\, \mathbb{E}_\pi[\, B(y_t + X_{t+1-\tau}, t+1) \mid y_t, t\,] \Big\}, \quad t \in \mathbb{T}_{\text{II}}, \end{aligned} \tag{8a}$$

$$B(y_{T_{\max}}, T_{\max}) = G(y_{T_{\max}}, T_{\max}). \tag{8b}$$

If the second term in the maximand of Eq. (8a) exceeds the first, $a_t = 1$ and stage II continues with an additional pairwise allocation so that $T > t$. For the 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. (8a) 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 are available during stage I, so that $y_t = y_0$ and

$n_t = n_0$ for $t \in \mathbb{T}_{\mathrm{I}} \equiv \{0, 1, \ldots, \tau - 1\}$. Thus,

$$B(y_t, t) = \max\left\{G(y_t, t), -c + \delta_{\mathrm{on}}(y_0/n_0) + (1+\tilde{\rho})^{-1} B(y_0, t+1)\right\}, \quad t \in \mathbb{T}_{\mathrm{I}}. \tag{9}$$

The special case of $\tau = 0$ is modeled by letting $\mathbb{T}_{\mathrm{I}}$ 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. (7).

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. (5) if it selects (almost surely) the argmax of Bellman's equation in Eqs. (8) and (9). 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 ($\delta_{\mathrm{on}} = 0$), 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 ($P\mu_{T+\tau} - I > 0$) once all outcomes have been observed, and the standard otherwise. This result is trivial from the observation that information, in expectation, has a nonnegative value.

The special case of offline learning ($\delta_{\mathrm{on}} = 0$), positive discounting ($\tilde{\rho} > 0$), no sampling costs ($c = 0$) and no time delay ($\tau = 0$) reduces to the special case of Chick and Gans (2009) for comparing a known alternative (standard) with known mean reward 0 with an unknown alternative (new technology) with unknown mean reward $PW - I$. The special case of offline learning ($\delta_{\mathrm{on}} = 0$), 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 ($\delta_{\mathrm{on}} \in \{0, 1\}$).

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:

$$\bar{V}(\mu_0, n_0) = \mathbb{E}[(PW - I)^+ + \delta_{\mathrm{on}}(W - c)^+ T_{\max,\tilde{\rho}} | \mu_0, n_0]. \tag{10}$$

The term $\mathbb{E}[(PW - I)^+ | \mu_0, n_0]$ is the oracle's expected reward from selecting the best technology immediately before executing the trial (that is, assuming no penalties for discounting). The term $\mathbb{E}[\delta_{\mathrm{on}}(W - c)^+ T_{\max,\tilde{\rho}} | \mu_0, n_0]$ 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 $\bar{V}(\mu_0, n_0)$ with the expected reward of any given policy. It will be useful for characterizing the optimal policies in Propositions 2.2 and 2.3 below. Proofs of mathematical claims in the paper can be found in Appendix A.

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

$$V^{\pi}(\mu_0, n_0) = \bar{V}(\mu_0, n_0) - \tilde{V}^{\pi}(\mu_0, n_0), \tag{11}$$

where $\tilde{V}^{\pi}(\mu_0, n_0) \equiv \mathbb{E}_{\pi}[\mathcal{K}_{\pi} + \mathcal{S}_{\pi} + L_{\pi} | \mu_0, n_0]$ and the following terms are each non-negative:

$$\mathcal{K}_{\pi} \equiv \sum_{t=0}^{T-1} \mathcal{K}_{\pi,t}, \text{ where } \mathcal{K}_{\pi,t} = (c - \delta_{\text{on}}(W - (W-c)^{+}))/(1+\tilde{\rho})^{t}, \tag{12a}$$

$$\mathcal{S}_{\pi} \equiv \sum_{t=T}^{T_{\max}-1} \delta_{\text{on}}(W-c)^{+}/(1+\tilde{\rho})^{t}, \tag{12b}$$

$$\text{and } L_{\pi} \equiv (PW - I)^{+} - \mathbf{1}_{\mathcal{D}=N}(PW - I)/(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}. \tag{12c}$$

By Eq. (11), a policy π maximises V^{π} if and only if it minimises $\tilde{V}^{\pi}$. Minimisation of $\tilde{V}^{\pi}$ is itself a sequential optimal stopping problem, in which $\mathcal{K}_{\pi}$ is an opportunity cost of sampling, $\mathcal{S}_{\pi}$ is a residual penalty in the presence of online learning if the stopping time is not equal to the oracle's stopping time and L_{π} is the opportunity cost of selecting a potentially suboptimal technology $\mathcal{D}$ after all outcomes are observed, accounting for any discounting owing to the delay. This observation, together with the nonnegativity of $\mathcal{K}_{\pi}$, $\mathcal{S}_{\pi}$, and L_{π}, allows us to use Bertsekas (2005) and Bertsekas and Shreve (1978) to characterise the optimal policy with Bellman's equation.

Proposition 2.2 *If all decisions of a policy $\pi \in \Pi$ attain the maximum in Bellman's equation in Eq. (9) for stage I decisions and in Eq. (8) 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.,*

$$V^{\pi}(\mu_0, n_0) = V^{\pi^*}(\mu_0, n_0) = B(\mu_0 n_0, 0). \tag{13}$$

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 variance. 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 $\delta_{\text{on}} = 0$ 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 $t = 0, 1, \ldots, T_{\max}$; 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 $\tilde{\rho} = 0$, $\delta_{\text{on}} = 0$ 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. (5) 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 $\mathcal{C}$.

3 Illustration of features of the optimal policy

This section illustrates the main features of the optimal policy and assesses some of its characteristics. We call the optimal policy π^* of Eq. (5) the 'Optimal Bayes Sequential' policy. It 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 5, where data from a clinical trial are 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 $P\mu_{T+\tau} - I > 0$. The 'Optimal Bayes One Stage' policy chooses a sample size $u^*(\mu_0)$ in the set $\mathbb{T}$ which maximises the net benefit of sampling in expectation,

$$u^*(\mu_0) = \underset{u\in\mathbb{T}}{\arg\max} \left\{ \left(\sum_{t=0}^{u-1} \frac{-c + \delta_{\text{on}}\mu_0}{(1+\tilde{\rho})^t} \right) + \frac{\mathbb{E}[(PZ'_{0,u} - I)^+ \mid \mu_0, n_0]}{(1+\tilde{\rho})^{\mathbf{1}_{u>0}(u+\tau)}} \right\}, \tag{14}$$

where $Z'_{0,u} \equiv \mathbb{E}[W \mid \mathcal{F}_0, X_1, X_2, \ldots, X_u] \sim \mathcal{N}(\mu_0, (\sigma_X^2/n_0)(u/(n_0+u)))$.

The time delay for observing the primary end point is set equal to one year and the rate of recruitment to the trial is set to $R = 1000$ pairwise allocations per year. The discount rate and the fixed cost of switching technologies are set to zero ($\tilde{\rho} = 0, I = 0$). 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 $\sigma_0 = 2000$. The effective sample size of the prior distribution is therefore $n_0 = (\sigma_X/\sigma_0)^2 = 100$. There is no online learning ($\delta_{\text{on}} = 0$).

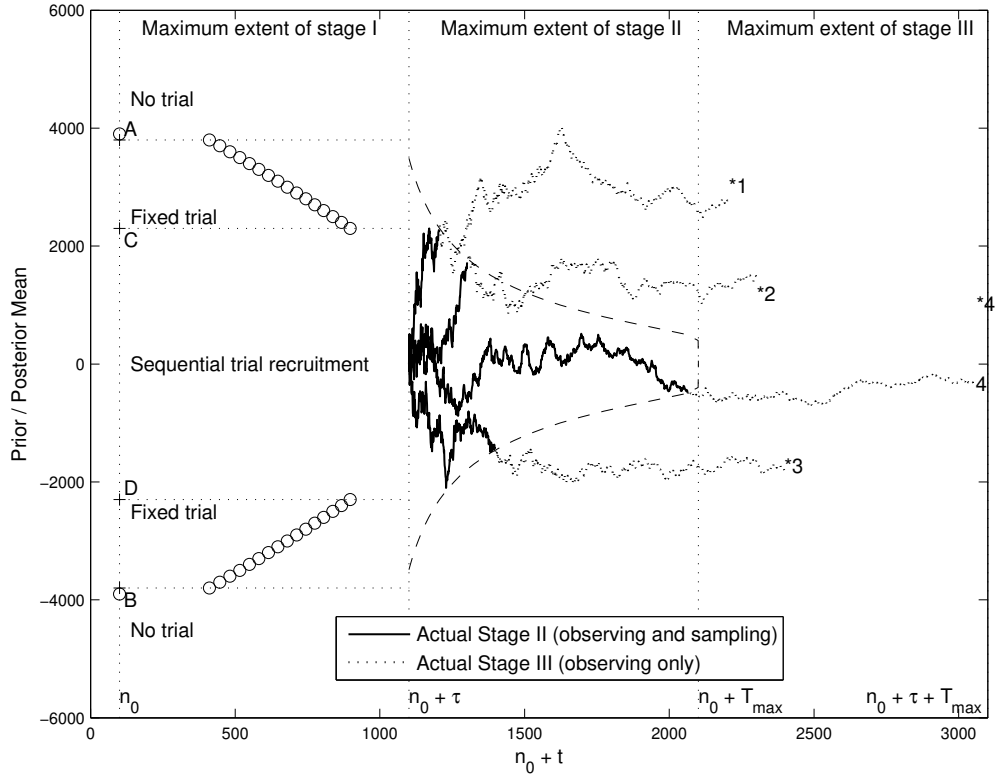


Figure 2: Optimal Bayes Sequential policy, together with four stage II/III paths of the posterior mean with prior mean $\mu_0 \approx 17$. KEY: '*' value of the sampling mean w_i for each path i; '—-' path of posterior mean when in stage II; '$\cdots$' path of posterior mean when in stage III. '+' thresholds A, B, C, D delineate the ranges for 'no trial'/'fixed trial'/'sequential trial recruitment'; '$\circ$' optimal stage I sample sizes.

Figure 2 plots the optimal stopping boundaries 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. The boundaries between the 'no trial'/'fixed trial'/'sequential trial recruitment' 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). 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.

Figure 2 shows that stage II starts at an effective sample size of $n_0 + \tau = 1100$ pairwise allocations. Because there is no discounting, there is symmetry above and below $\mu = I/P = 0$ in the stage II stopping boundary and the stage I fixed sample sizes (recall Prop. 2.4).

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. (3a) with $t = 0$, followed by generation

of the X_i for $i = 1, 2, \ldots$ given that draw, to generate the sample path μ_t using Eqs. (2) and (3b).

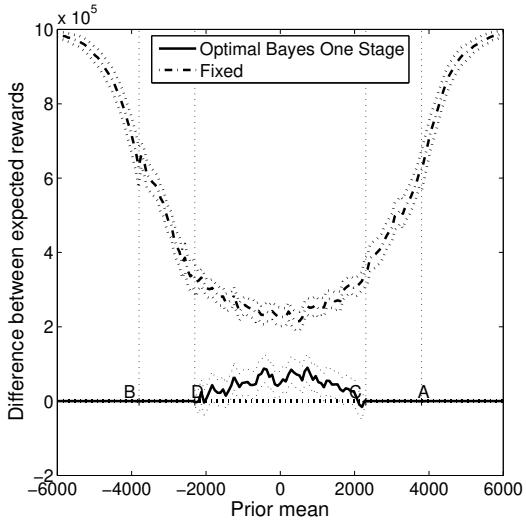
Figure 2 shows four sample paths for a prior mean of $\mu_0 \approx 17$ lying in the 'sequential trial recruitment' region, meaning that it is optimal to proceed to stage II. The realised values of W_i, $i = 1,\ldots,4$ are indicted by '*'s. Stage II sections of the paths are marked as continuous lines. 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 path 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 path 2, with the posterior mean hitting the upper boundary a little later than for path 1. For path 3, $w_3 < 0$ and the new technology is rejected once all pipeline subjects have been observed (again the correct decision). 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.

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. (14). Thick lines represent the averages, dotted 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.[1]

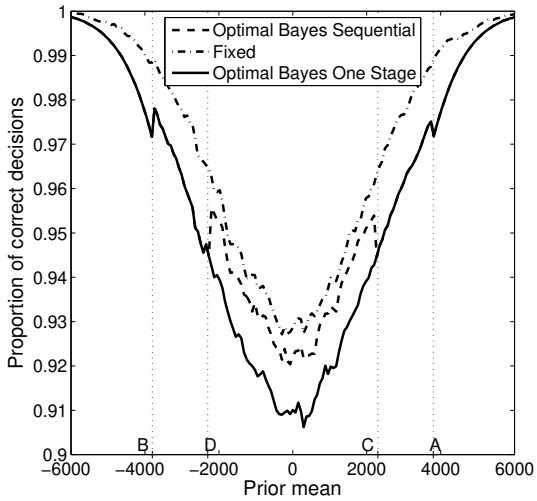
Figure 3(a) shows that, as expected, the Optimal Bayes Sequential policy outperforms the other two policies when judged according to net gain. Compared with the 'Fixed' policy, the greatest gains for the Optimal Bayes Sequential policy may be seen at extreme values of the prior mean, which is unsurprising: 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 $\mu_0 = I/P$ (= 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 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 other policies, which commit 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. 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 its decision rule sequentially updates the information after each observation. The Fixed policy performs best for the probability of correct selection because it guarantees the highest amount of

[1] Smooth curves in the paper are obtained from the partial differential equation methods of Appendix S. The jaggedness in the sample averages is approximately the size of the confidence intervals for the plotted averages.



(a) 'Net gain' in expected reward of Optimal Bayes Sequential over comparator policies.



(b) Proportion of simulations which make the correct adoption decision.

Figure 3: Operating characteristics for the illustration of section 3.

information. This is obtained at an economic cost, however (refer to Figure 3(a)).

In section S.6.1 we illustrate the effect of reducing the delay from τ = 1000 to 500 pairwise allocations. Stage II boundaries change shape slightly but are shifted left (τ is smaller). There are no stage I optimal sample sizes (point D moves to point B, and point C moves to point A).

4 Unknown sampling variance

The analysis to date has assumed that the sampling variance, σ_X^2, is known, but in practice this will not be the case. This section extends the analysis to the case when σ_X^2 is unknown, adopting and developing the framework proposed by Chick et al. (2015).

Define T_ν as a standard Student t random variable with ν degrees of freedom (dof) and define ϕ_ν and Φ_ν as, respectively, its pdf and cdf. Denote the distribution of the three parameter Student t random variable, $\mu + T_\nu/\sqrt{\kappa}$, as $\texttt{St}(\mu, \kappa, \nu)$, with precision κ. If $\nu > 2$, $\mathrm{Var}[T_\nu] = \nu/(\nu-2)$.

As before, assume that X_i are normally distributed and conditionally independent, given the unknown expected value, W, and *unknown* σ_X^2. Let ς be the random variable whose realization is σ_X^2. We choose a prior distribution in the conjugate family for normally distributed samples with unknown mean, W, and variance, ς (DeGroot, 1970, § 9.6). Then:

$$\begin{aligned} X_i \mid W, \varsigma &\overset{iid}{\sim} \mathcal{N}(W, \varsigma), \\ \varsigma &\sim \texttt{InvGamma}(\xi_0, \chi_0), \\ W \mid \varsigma &\sim \mathcal{N}(\mu_0, \varsigma/\eta_0), \end{aligned} \tag{15}$$

where $\xi_0 > 1$ and χ_0 are shape and scale parameters of an inverse-gamma distribution with mode $\chi_0/(\xi_0+1)$, expected value $\mathbb{E}[\varsigma] = \chi_0/(\xi_0-1)$, $\mathbb{E}[1/\varsigma] = \xi_0/\chi_0$ and $\mathrm{Var}[1/\varsigma] = \xi_0/\chi_0^2$ and μ_0

and η_0 determine the a priori mean and variance of the unknown sampling mean. It follows that W is a $\texttt{St}\,(\mu_0, \xi_0\eta_0/\chi_0, 2\xi_0)$ random variable and $\mathrm{Var}[W] = \chi_0/[(\xi_0 - 1)\eta_0]$ when $\xi_0 > 1$.

For $t = 0, 1, \ldots, \tau - 1$, no observations arrive owing to the delay, so $\xi_{t+1} = \xi_t$, $\chi_{t+1} = \chi_t$, $\eta_{t+1} = \eta_t$, and $\mu_{t+1} = \mu_t$. For $t = \tau, \tau + 1, \ldots$, the posterior distribution can be updated by adapting DeGroot (1970) to account for observations on the pipeline subjects as follows:

$$\begin{aligned} \varsigma \,|\, X_{t+1-\tau}, \mathcal{F}_t &\sim \texttt{InvGamma}\,(\xi_{t+1}, \chi_{t+1}), \\ W \,|\, \varsigma, X_{t+1-\tau}, \mathcal{F}_t &\sim \mathcal{N}(\mu_{t+1}, \varsigma/\eta_{t+1}), \\ W \,|\, \mathcal{F}_t &\sim \texttt{St}\,(\mu_t, \eta_t\xi_t/\chi_t, 2\xi_t), \end{aligned}$$

where $\xi_{t+1} = \xi_t + 1/2$, $\chi_{t+1} = \chi_t + \frac{\eta_t}{2(\eta_t+1)}(\mu_t - X_{t+1-\tau})^2$, $\eta_{t+1} = \eta_t + 1$, and $\mu_{t+1} = (\eta_t\mu_t + X_{t+1-\tau})/\eta_{t+1}$. We note that ξ_t and η_t are deterministic functions of t, given ξ_0 and η_0. The posterior precision, $\xi_t\eta_t/\chi_t$, is the Bayesian analog of the frequentist observed information $\mathcal{I}_{(t-\tau)^+}$ given $(t-\tau)^+$ observations (Hampson and Jennison, 2013, § 6 on unknown variance).

The predictive distribution for the posterior mean given that sampling stops at time $T = t$, with state (μ_t, χ_t, t), and with $u = \min(T, \tau)$ pipeline subjects to arrive, is (DeGroot, 1970):

$$Z_{t,u} \sim \texttt{St}\left(\mu_t, \frac{\xi_t\eta_t}{\chi_t}\frac{(\eta_t + u)}{u}, 2\xi_t\right) \tag{16}$$

(compare with Eq. (6) for the case of known variance).

Just as for the case of known σ_X^2, the optimal solution involves solving stages I, II and III as illustrated in Figure 1. In contrast to the case of known σ_X^2, the state vector (μ_t, χ_t, t), and not (μ_t, t), is sufficient to summarize $\mathcal{F}_t$ for the purposes of inference about W. Stages I and III are straightforward to modify: the stage III terminal reward function in Eq. (7) is modified by taking the expectation in its RHS with respect to the Student t distribution for $Z_{t,u}$ in Eq. (16), rather than the normal distribution of Eq. (6). A similar change is sufficient to modify the expectation in the RHS of Eq. (14) for stage I, to determine the Optimal Bayes One Stage policy.

Chick et al. (2015) proposed three approaches to solving stage II for the case of unknown σ_X^2. Here we extend the so-called KG_* variant of the knowledge gradient (Chick and Frazier, 2009; Frazier and Powell, 2010) which, given information to hand, continues sampling if and only if there exists a feasible one-stage sampling policy giving a greater expected reward than would be gained from stopping.

Define $\hat{B}_\beta(\mu_t, \chi_t, t)$ to be the expected value of making $\beta \geq 0$ more pairwise allocations at time t, observing the remaining $\beta + \min(t, \tau)$ outcomes, accruing online rewards (if applicable), and selecting the better technology:

$$\hat{B}_\beta(\mu_t, \chi_t, t) = \mathbb{E}\left[\left\{\sum_{i=0}^{\beta-1} \frac{-c + \delta_{\mathrm{on}} X_{t+i+1}}{(1+\tilde{\rho})^i}\right\} + \frac{(PZ'_{t,\beta+\min(t,\tau)} - I)^+}{(1+\tilde{\rho})^{\mathbf{1}_{\beta+t>0}(\beta+\tau)}}\middle|\mathcal{F}_t\right]. \tag{17}$$

where $Z'_{t,u} \sim \texttt{St}\,(\mu_t, ((\xi_t\eta_t)/\chi_t)((\eta_t + u)/u), 2\xi_t)$ adapts $Z'_{0,u}$ from Eq. (14) to the case of unknown sampling variance given $\mathcal{F}_t$. We note that $\hat{B}_0(\mu_t, \chi_t, t)$, with $\beta = 0$ additional samples, is precisely the expected value of stopping, $G(y_t, t)$ in Eq. (7), extended to handle unknown variances, for $t \in \mathbb{T}_{\mathrm{I}} \cup \mathbb{T}_{\mathrm{II}}$.

To adapt stage II to the case of unknown sampling variance, we replace the value of continuing over all nonanticipative policies (the second term in the maximand of Eq. (8a)) with a set of one-step lookahead policies, $\mathcal{B}_t$. Thus, Eq. (8a) is approximated by

$$\hat{B}^*(\mu_t, \chi_t, t) = \max\left\{\hat{B}_0(\mu_t, \chi_t, t), \max_{\beta\in\mathcal{B}_t} \hat{B}_\beta(\mu_t, \chi_t, t)\right\}, \quad t \in \mathbb{T}_{\mathrm{II}}. \tag{18}$$

The set $\mathcal{B}_t = \{1, 2, \ldots, T_{\max} - t\}$ contains the nonzero pairwise allocations which remain. The choice $\mathcal{B}_t = \{2^{-1/2}, 1, 2^{1/2}, \ldots, \min(128, T_{\max} - t)\}$ proved useful as an approximation and is used in numerical results here. Let $\beta^* = \arg\max_{\beta\in\mathcal{B}_t} \hat{B}_\beta(\mu_t, \chi_t, t)$.

We define the KG_* continuation set $\mathcal{C}_{\mathrm{KG}_*}$ here to be the set of (μ_t, χ_t, t) such that one continues to allocate if and only if $\hat{B}_{\beta^*}(\mu_t, \chi_t, t) > \hat{B}_0(\mu_t, \chi_t, t)$ (i.e., there is a non-zero, feasible one stage sampling plan whose expected reward exceeds that of stopping immediately and acting optimally once the observations on the pipeline subjects are observed).

A second approach to solving stage II that was proposed by Chick et al. (2015) is based on the numerical solution of the PDE free boundary problem. This adjusts the variance of the diffusion process to account for the uncertainty on σ_X^2 and is briefly described in Appendix S.6.3. The following application considers the operating characteristics of both approaches to dealing with an unknown sampling variance.

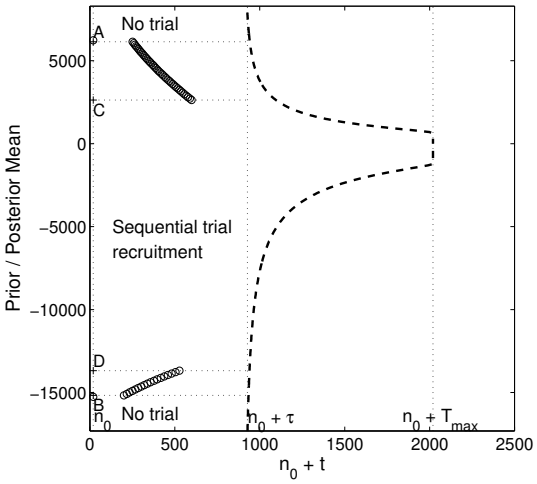
5 Application: drug-eluting stents

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.

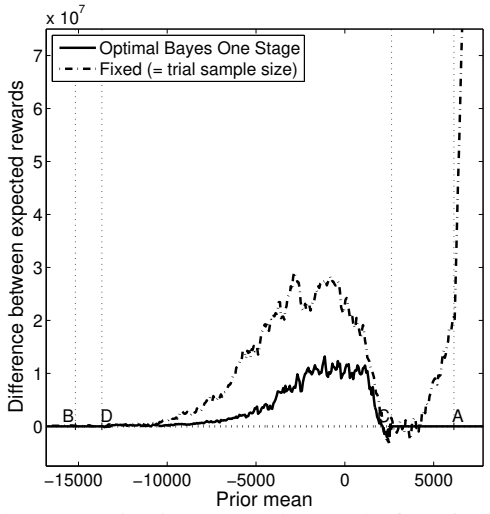
We consider the performance of the Optimal Bayes Sequential policy of section 2 (known sampling variance) and the policy of section 4 (unknown sampling variance) with what is a Fixed policy with the same sample size as the SIRIUS study (529 patient pairs in 7 months) and the Optimal Bayes One Stage policy. For the purposes of this section, we set $\delta_{\text{CE}} = 1$ in Eq. (1) to concentrate on the cost and QALY results at one year of follow-up that are reported in Cohen et al. (2004). This section is intended to illustrate how our model may be populated with data from a health technology assessment; it is not intended to represent a comment on the health technology itself.

5.1 Known sampling variance

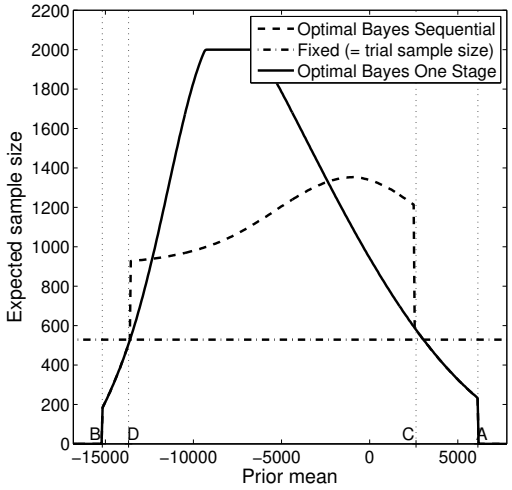
Where possible, parameter values are derived from Moses et al. (2003) and Cohen et al. (2004). Otherwise they are based on assumptions. The value of $\sigma_X = \$17358$ is derived from point estimates in Cohen et al. (2004) and the assumption $\lambda = \$50000/\text{QALY}$. We set $T_{\max} = 2000$,



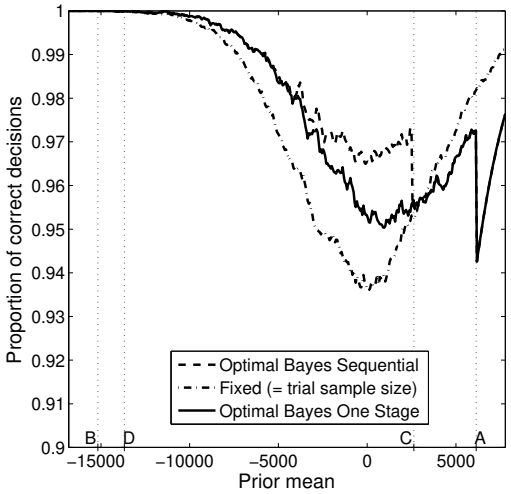
(a) Optimal Bayes Sequential policy.



(b) 'Net gain' in expected reward of Optimal Bayes Sequential over comparator policies.



(c) Expected sample size.



(d) Proportion of simulations which make the correct adoption decision.

Figure 4: Optimal Bayes Sequential policy and operating characteristics for the stents application of section 5 (known variance).

which is higher than the annual rate of recruitment to the study (calculated to be $R = 529 \times 12/7 = 907$ patient pairs per year). The delay in response is one year, so $\tau = 907$. A zero switching cost is assumed ($I = 0$) and the effective sample size in the prior distribution is assumed to be $n_0 = 20$. We assume $c = \$200$ and $P = 2 \times 10^6$. In contrast to the illustration of section 3, the discount rate is chosen to be 1% per annum ($\tilde{\rho} = (1 + 0.01)^{-R} - 1$). Benefits accruing to trial participants are not valued ($\delta_{\text{on}} = 0$).

Figure 4(a) shows the optimal stopping boundaries. The '$\circ$'s in Figure 4(a) indicate that, for a prior mean lying 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 the adoption decision that is implied by starting to experiment (by experimenting, one must wait for at least a year before making the adoption decision, and rewards are discounted). For a prior mean lying between points C and D it is optimal to proceed to stage II.

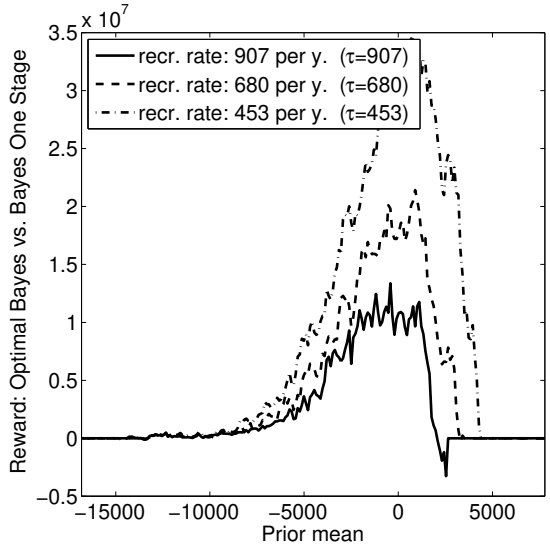
In the absence of discounting, Prop. 2.4 implies that there is a greater expected reward when the posterior mean is above $I/P = 0$ than when it is below that value by the same absolute amount. With a positive discount rate, the expected benefit of continued sampling is penalized more for values of the 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 for a positive discount rate.

This asymmetry is reflected in the plots of the 'net gains' (the differences between the expected reward of the Optimal Bayes Sequential policy and the two comparators) in Figure 4(b) and the expected sample sizes in 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 sample sizes for the different trial designs at points A–D (see the discussion of Figure 4(a) above).

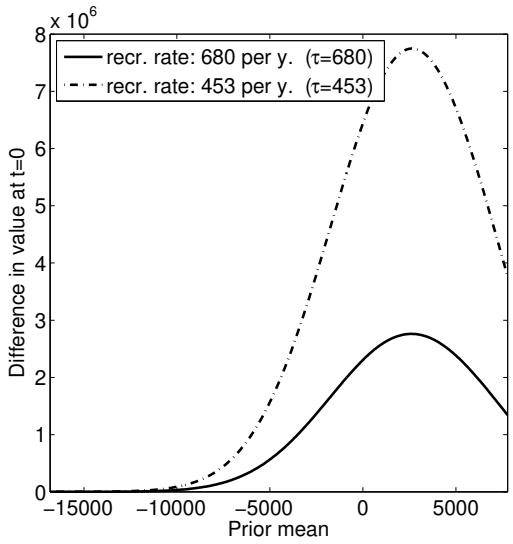
For a value of the prior mean close to zero, Figure 4(b) shows that the expected net gain of the Optimal Bayes Sequential policy over the Fixed policy is approximately \$20m and over the Optimal Bayes One Stage policy it is \$10m. Not apparent from Figure 4(b), due to the scaling, is that fact that, 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 discounted cost of sampling patients under the Fixed policy. This is because, if the value of the prior mean is low enough, it will be 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 with very high probability. As $\mu_0 \to \infty$, the net gain of the Optimal Bayes Sequential policy over the Fixed policy grows without bound: with a very optimistic prior mean, it is optimal to adopt immediately 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. (5), an objective which requires achieving the sample size which appropriately balances the benefits to patients with the costs of learning. With this objective in mind, given μ_0, the Optimal Bayes Sequential policy may sample more, the same as, or less than the Optimal Bayes One Stage and the Fixed policies according to the filtration defined by the observations seen through time t, which depends on w_i.

An estimate of the probability of correct selection after all outcomes have been observed is



(a) Difference between expected reward of Optimal Bayes Sequential and Optimal Bayes One Stage policies for several recruitment rates.



(b) Expected reward of Optimal Bayes Sequential policy: expected reward when recruitment rate is 907/year minus expected reward when it is 680/year and 453/year.

Figure 5: Effect of changing τ by changing the recruitment rate.

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. It is similar to the comparators to the left of D. Over the majority of the range CA, 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 expected sample sizes that occur at those points (Figure 4(c)).

The estimate of the probability of a 'decision reversal' (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 application.

Assessing the sensitivity of these results to changes in parameter values is straightforward. Here we consider changing τ by changing the recruitment rate, R. Such a change can be caused by a change in the recruitment rate at a single facility or a change in the number of facilities which participate in the trial. Figure 5(a) 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 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. 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 5(b) focuses on the Optimal Bayes Sequential policy showing that, for this application, the higher is the recruitment rate, the higher is the expected reward.

Figure 8 of Appendix S.6.2 shows the impact of changing the recruitment rate on the expected sample size. Figure 9 shows the stopping boundaries for different values of the sampling cost c.

5.2 Unknown sampling variance

When the sampling variance is unknown, the following additional parameters are required to implement the model of section 4. For the prior distributions, we set $\eta_0 = n_0$ (η_0 and n_0 both represent the sample size in the prior distribution for the unknown mean), $\xi_0 = 2n_0 - 1$ and $\chi_0 = 17358^2(\xi_0 - 1)$, so that $\mathbb{E}[\varsigma]$ equals the point estimate of the sampling variance ($\sigma_X^2 = 17358^2$) from the study. The KG_* continuation set isestablished for the value of χ_t such that χ_t/ξ_t equals 17358^2. The continuation set for other χ_t can be found by rescaling states (namely, $(\mu_t, \chi_t, t) \in \mathcal{C}_{\mathrm{KG}_*}$ if and only if $(a\mu_t, a^2\chi_t, t) \in \mathcal{C}_{\mathrm{KG}_*}$).

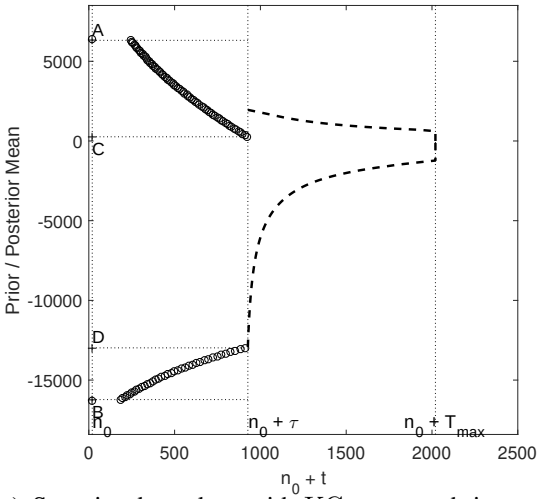
Figure 6 replicates Figure 4 for the case of unknown variance, solved using KG_*. A comparison of Figure 6(a) with Figure 4(a) shows that the continuation set of Stage I is slightly wider when the variance is unknown, owing to the additional dimension of uncertainty. The opposite effect is seen in Stage II, because the KG_* approach is one stage and so the value of continuing is lower than the fully sequential approach adopted for the case of known variance. As a result, the expected sample size is smaller for KG_* (Figures 6(c) and 4(c)) owing to earlier stopping, on average, and there is a smaller advantage in terms of the proportion of correct decisions (Figures 6(d) and 4(d)). This, in turn, implies that the net gain in comparison with alternative policies is slightly reduced (Figures 6(b) and 4(b)).

Section S.6 of the OSM discusses the results obtained by replacing the KG_* approach with one based on the numerical solution of the PDE free boundary problem and plug-in estimates of the sampling variance (Chick et al., 2015). The net gain (Figure 10) is very similar to the case of known variance.

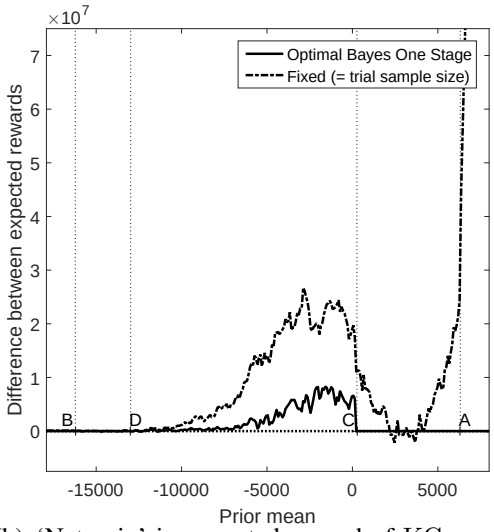
The results of additional simulations (data not shown) suggest that the advantage of the Optimal Bayes Sequential Policy over a Fixed Policy remains when $\tilde{\rho}$ is smaller. Those results also show that the benefit of the Optimal Bayes Sequential Policy over the Optimal Bayes One Stage policy decreases as $\tilde{\rho}$ decreases.

6 Discussion

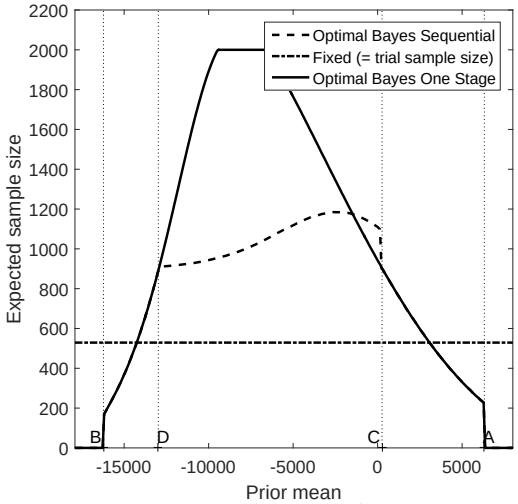
We have solved a Bayesian decision-theoretic model of sequential experimentation with delay and applied it to the field of medical statistics. The model maximises the expected benefits of the technology adoption decision, minus the cost of the sequential experiment itself, and it can value benefits accruing to study participants as well as to those who benefit from the adoption decision. Explicit measurement of these costs and benefits meets a growing demand for a 'value-based' approach to health care decision making at policy level (NICE, 2012; Porter, 2010). At the level of Phase III trial design and health technology assessment, the model helps answer the following questions: is it worth carrying out any trial at all? If it is, should the trial be sequential or of a fixed sample size? If a sequential trial is chosen, how should stopping boundaries be defined in



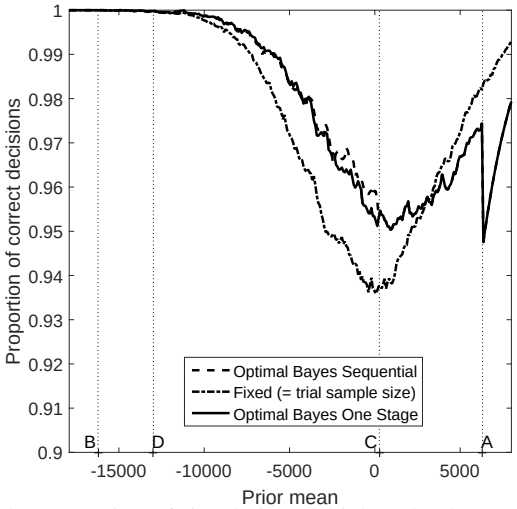
(a) Stopping boundary with KG_* approach in section 5.2.



(b) 'Net gain' in expected reward of KG_* approach over comparator policies.



(c) Expected sample size.



(d) Proportion of simulations which make the correct adoption decision.

Figure 6: Operating characteristics for the stents application of section 5 with KG_* approach in section 5.2 to approximate Optimal Bayes Sequential policy with unknown sampling variance.

the presence of delay in observing the primary end point? How do parameters such as the rate of patient recruitment and the cost of sampling influence optimal design?

Monte Carlo simulations which compare the performance of the 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 are derived from a continuous time approximation. In addition, the applications show that the Optimal Bayes Sequential policy results in the highest net gain over competing policies when

the expected sample size is close to, or greater than, the expected sample sizes of those policies. Further, the higher is the delay, the less attractive is the sequential design over the Fixed and the Optimal Bayes One Stage policies. Clearly, the precise performance of the model will depend on the particular application of interest.

Directions for future research are numerous. The model assumes that only two health technologies are being considered, it does not incorporate intermediate outcomes that are correlated with the primary end point and it is assumed that all pipeline data must be observed before an adoption decision is made. Future work includes relaxing these assumptions and exploring further the issues of unknown sampling variance and sensitivity of the policy to the choice of sampling distribution.

A Mathematical proofs for the discrete time model

Proof of Prop. 2.1. Condition on W and T in Eq. (4) and use the tower property of conditional expectation:

$$\begin{aligned} V^{\pi}(\mu_0, n_0) &= \mathbb{E}_{\pi}\left[\mathbb{E}\Big[\Big\{\sum_{t=0}^{T-1}\frac{-c+\delta_{\text{on}}X_{t+1}}{(1+\tilde{\rho})^t}\Big\} + \frac{\mathbf{1}_{\mathcal{D}=\text{N}}(PW-I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}}\Big|W,T\Big]\Big|\mu_0,n_0\right] \\ &= \mathbb{E}_{\pi}\left[\Big\{\sum_{t=0}^{T-1}\frac{-c+\delta_{\text{on}}W}{(1+\tilde{\rho})^t}\Big\} + \frac{\mathbf{1}_{\mathcal{D}=\text{N}}(PW-I)}{(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}}\Big|\mu_0,n_0\right]. \end{aligned} \tag{19}$$

Substituting Eq. (12) into the RHS of Eq. (11) and simplifying gives Eq. (19). $\mathcal{K}_{\pi} \geq 0$ when $\delta_{\text{on}} = 0$ because $c \geq 0$. When $\delta_{\text{on}} = 1$ it is strictly positive for $W < c$ and 0 otherwise. $\mathcal{S}_{\pi} \geq 0$ because $(W-c)^+ \geq 0$. $\tilde{\rho} \geq 0$, $\tau \geq 0$, $T \geq 0$ imply that $(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)} \geq 1$. Further, $(PW-I)^+ \geq 0$ and $(PW-I)^+ \geq \mathbf{1}_{\mathcal{D}=\text{N}}(PW-I)$. Hence, $(PW-I)^+ \geq \mathbf{1}_{\mathcal{D}=\text{N}}(PW-I)/(1+\tilde{\rho})^{\mathbf{1}_{T>0}(T+\tau)}$ independent of the sign of $\mathbf{1}_{\mathcal{D}=\text{N}}(PW-I)$. Thus, $L_{\pi} \geq 0$. □

Proof of Prop. 2.2. From Eq. (11), a policy π in a given set of policies maximises V^{π} if and only if it minimises $\tilde{V}^{\pi}$. This reformulation is useful, because the non-negativity of $\mathcal{K}_{\pi,t}$, $\mathcal{S}_{\pi}$ and L_{π} for all (y_t, t, W) satisfies the (F^+) property of Bertsekas and Shreve (1978, Chap. 8, p. 192). However, Bertsekas and Shreve (1978, Chap. 8) require rewards which depend on a known state vector, whereas $\mathcal{K}_{\pi,t}$, $\mathcal{S}_{\pi}$, and L_{π} depend on the unknown W. Following Bertsekas (2005, p. 218-222), we augment the state to be (y_t, t, W), define the information vector I_t with $I_0 = (0, y_0)$ and $I_t = (t, y_0, y_1, \ldots, y_t, a_0, \ldots, a_{t-1})$ for $t = \mathbb{T}\backslash\{T_{\max}\}$, and must now allow a broader set of policies $\tilde{\pi}$ whose actions a_t may depend on I_t, not just (y_t, t).

Call the problem of finding a policy $\tilde{\pi}$ to minimize $\tilde{V}^{\tilde{\pi}}$, the 'regret problem.' Its Bellman equation, $\tilde{B}(y_t, t, W)$, consists of minimizing the expected cost of stopping (the first time at which $a_t = 0$),

$$\mathbb{E}\left[-G(y_t,t) + (PW-I)^+ + \sum_{t=0}^{T_{\max}-t-1}\delta_{\text{on}}(W-c)^+/(1+\tilde{\rho})^t \mid I_t\right], \tag{20}$$

and of making an additional pairwise allocation and proceeding optimally thereafter ($a_t = 1$),

$$\mathbb{E}[c - \delta_{\text{on}}(W - (W-c)^+) + (1-(1+\tilde{\rho})^{-1})(PW - I)^+ \\ + (1+\tilde{\rho})^{-1}\tilde{B}(y_t + \mathbf{1}_{t\geq\tau}X_{t+1-\tau}, t+1, W) \mid I_t], \qquad (21)$$

for $t = \mathbb{T}\backslash\{T_{\max}\}$. Its terminal cost is $\tilde{B}(y_{T_{\max}}, T_{\max}, W) = G(y_{T_{\max}}, T_{\max})$.

Because the (F$^+$) property holds for this problem, Prop 8.1 and Cor. 8.1.1 of Bertsekas and Shreve (1978) justify that it suffices to consider nonrandomised Markovian policies within the set of all policies when solving $\inf_{\tilde{\pi}} \tilde{V}^{\tilde{\pi}}$. Let $\tilde{\pi}_{\tilde{B}}$ be determined by $\tilde{B}$ for the regret problem. Although $\tilde{\pi}_{\tilde{B}}$ may depend on I_t for decisions a_t, note that (y_t, t) is sufficient for W: $\tilde{\pi}_{\tilde{B}}$ is Markovian in (y_t, t). Props. 8.2 and 8.5 of Bertsekas and Shreve (1978) show that $\tilde{\pi}_{\tilde{B}}$ is optimal, that is, $\tilde{V}^{\tilde{\pi}_{\tilde{B}}}(\mu_0, n_0) = \tilde{B}(\mu_0 n_0, 0, W) = \inf_{\tilde{\pi}} \tilde{V}^{\tilde{\pi}}(\mu_0, n_0)$.

The expectations in $\tilde{B}$ depend on I_t only through (y_t, t), and $\tilde{\pi}_{\tilde{B}}$ is therefore feasible for the original problem. By Prop. 2.1, $\tilde{\pi}_{\tilde{B}}$ is optimal for the original problem, and $V^{\tilde{\pi}_{\tilde{B}}}(\mu_0, n_0) = V^{\pi^*}(\mu_0, n_0) = \bar{V}(\mu_0, n_0) - \tilde{B}(\mu_0 n_0, 0, W)$.

To complete the proof, we show that $\tilde{\pi}_{\tilde{B}}$ also satisfies Bellman's equation of the original problem. By definition, $\tilde{\pi}_{\tilde{B}}$ chooses the smaller of Eq. (20) and Eq. (21). It therefore makes the same choices if one subtracts the same quantity from both equations. In particular, $\tilde{\pi}_{\tilde{B}}$ is still optimal if one subtracts $\mathbb{E}[(PW - I)^+ + \sum_{t=0}^{T_{\max}-t-1} \delta_{\text{on}}(W-c)^+/(1+\tilde{\rho})^t \mid I_t]$ from both Eq. (20) and Eq. (21). With a bit of algebra, one confirms that these subtractions result in terms which, in expectation, are -1 times the maximands in Bellman's equation, Eq. (8) and Eq. (9), for the original problem. Because $-\min(-a, -b) = \max(a, b)$, $\tilde{\pi}_{\tilde{B}}$ also satisfies Bellman's equation of the original problem. Setting $t = 0$ in the subtracted terms allows us to show $B(\mu_0 n_0, 0) = \bar{V}(\mu_0, n_0) - \tilde{B}(\mu_0 n_0, 0, W)$. Thus $B(\mu_0 n_0, 0) = V^{\pi^*}(\mu_0, n_0)$. □

Proof of Prop. 2.3. The proof is like that 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 nonnegative, the (P) assumption of Bertsekas and Shreve (1978, 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$^+$) property 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. Bertsekas and Shreve (1978, Prop. 9.8) justify the claim that the value function in Eq. (5) satisfies Bellman's equation for the regret problem, $\tilde{V}^{\tilde{\pi}_{\bar{B}}}(\mu_0, n_0) = \bar{B}(\mu_0 n_0, 0)$, and $\tilde{V}^{\tilde{\pi}_{\bar{B}}}(\mu_0, n_0) = \tilde{V}^{\pi^*}(\mu_0, n_0)$ follows from Bertsekas and Shreve (1978, Prop. 9.12). The link from $\tilde{V}$ back to V is as for Prop. 2.2. □

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. (8a), $G(\cdot)$, 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. (7), 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 (pdf) and the cumulative distribution function (cdf)

of a standard normal random variable (DeGroot, 1970). Moreover,

$$\mathbb{E}[(-Z)^+] = \sigma[\phi(-\xi') - \xi'\Phi(-\xi')] = \sigma[\phi(\xi') + \xi'\Phi(\xi')] - \xi. \tag{22}$$

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:

$$\mathbb{E}[PZ_{T,\min(T,\tau)} - I \mid Y_T = y_t, T = t] = P\Delta\mu, \tag{23a}$$

$$\mathbb{E}[P\tilde{Z}_{T,\min(T,\tau)} - I \mid Y_T = y_t, T = t] = -P\Delta\mu. \tag{23b}$$

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. (7) using Eqs. (23a) and (23b):

$$\begin{aligned} G(y_t, t) &= \mathbb{E}[(PZ_{T,\min(T,\tau)} - I)^+ \mid Y_T = y_t, T = t] \\ &= \sigma[\phi(P\Delta\mu/\sigma) + (P\Delta\mu/\sigma)\Phi(P\Delta\mu/\sigma)] \\ &= \sigma[\phi(-P\Delta\mu/\sigma) + (-P\Delta\mu/\sigma)\Phi(-P\Delta\mu/\sigma)] - (-P\Delta\mu) \\ &= \mathbb{E}[P(\tilde{Z}_{T,\min(T,\tau)} - I/P)^+ \mid Y_T = \tilde{y}_t, T = t] + P\Delta\mu \\ &= G(\tilde{y}_t, t) + P\Delta\mu. \end{aligned} \tag{24}$$

Thus, given Eq. (8b), $B(y_t, t) - P\Delta\mu = B(\tilde{y}_t, t)$ 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. (24), the first maximand on the right hand side of Eq. (8a) 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. (8a). If $t \geq \tau$, let $\hat{X}$ be a normal random variable with mean 0 and variance σ^2_X. If $\delta_{\mathrm{on}} = 0$ and $\tilde{\rho} = 0$ then

$$\begin{aligned} B_2(y_t, t) - B_2(\tilde{y}_t, t) &= \mathbb{E}_\pi[B(y_t + y_t/n_t + (X_{t+1-\tau} - y_t/n_t), t+1) \mid Y_T = y_t, T = t] \\ &\quad -\mathbb{E}_\pi[B(\tilde{y}_t + \tilde{y}_t/n_t + (X_{t+1-\tau} - \tilde{y}_t/n_t), t+1) \mid Y_T = \tilde{y}_t, T = t] \\ &= \mathbb{E}[B(y_t + y_t/n_t + \hat{X}, t+1) \mid Y_T = y_t, T = t] \\ &\quad -\mathbb{E}[B(\tilde{y}_t + \tilde{y}_t/n_t - \hat{X}), t+1) \mid Y_T = \tilde{y}_t, T = t] \qquad (25) \\ &= \mathbb{E}[P(\Delta\mu + \hat{X}/n_{t+1}) \mid Y_T = y_t, T = t] = P\Delta\mu. \qquad (26) \end{aligned}$$

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. (25) 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. (26).

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. (9).

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. (13), 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. (8a) 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. Follows directly from Chick and Frazier (2012, Prop. 3). See Appendix S.1 in the OSM for further detail.

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, INSEAD, the Yale School of Management, and Bocconi and Bologna Universities, as well as Noah Gans 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.

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.

Bertsekas, D. P. (2005). *Dynamic Programming and Stochastic Control: Volume I*. Athena Scientific, Belmont, MA, 3 edition.

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. *JRSS, 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., Forster, M., and Pertile, P. (2015). Optimal sequential sampling with delayed observations and unknown variance. In Yilmaz, L., Chan, W., Moon, I., Roeder, T., Macal, C., and Rossetti, M., editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 3789–3800, Piscataway, NJ. IEEE, Inc.

Chick, S. E. and Frazier, P. (2009). The conjunction of the knowledge gradient and the economic approach to simulation selection. In *Proceedings of the 2009 Winter Simulation Conference*, pages 528–539, Piscataway, New Jersey. IEEE, Inc.

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. *JRSS, Series B*, 75(1):48.

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

Frazier, P. I. and Powell, W. B. (2010). Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403.

Hampson, L. and Jennison, C. (2013). Group sequential tests for delayed responses. *JRSS, 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.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting experts' probabilities*. John Wiley & Sons, Chichester.

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

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

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

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley & Sons, Chichester, 2nd 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 Supplementary Material (OSM)

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

Appendix S.1 provides the proof of Prop. 2.5.

Solving for π^* numerically for the model of section 2 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 is useful for the numerical results of sections 3 and 5. 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 continuous time (CT) optimal stopping problem. Appendix S.2 defines the diffusion and writes the continuous time analog of the discrete time optimal stopping problem in Eq. (5). 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.3 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.4 describes computational techniques for approximating the stopping boundaries of the optimal policy π^*_{CT} and value function for the CT problem with general τ. Numerical results in the main paper use π^*_{CT} to approximate the optimal policy π^* for the discrete time problem. 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`.

Appendix S.5 presents connections of the modelling approach in the main paper to the multi-armed bandit (MAB) literature.

Appendix S.6 provides additional analysis for section 3, further results for the application of section 5, as well as an additional application.

S.1 Additional analysis for the discrete time problem

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) shows that $T \leq \Upsilon \equiv 1 + (P^2\sigma_X^2)/(2\pi c^2) - n_0$ 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. □

S.2 Continuous time analog of discrete time problem

In order to approximate the optimal delayed sequential sampling problem specified by Eq. (5) in continuous time, the definitions of the time t, the sum Y_t defined in Eq. (2), the induced filtration $\mathcal{F}_t$, a policy π and the discount rate 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. The continuous time discount rate is $\rho = \ln(1+\tilde{\rho})$.

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:

$$\mathrm{d}Y_t = W\mathrm{d}t + \sigma_X \mathrm{d}\mathcal{V}_{t-\tau}, \quad \tau \le t \le T_{\max}+\tau, \tag{27}$$

where $\mathcal{V}_u$ 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 $\mathcal{V}_{\lfloor u \rfloor}$ 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 $\mathcal{D}_{\text{CT}} \in \{\text{N}, \text{S}\}$ for a technology 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

$$V^{\pi}_{\text{CT}}(\mu_0, n_0) = \mathbb{E}_{\pi_{\text{CT}}}\left[\int_0^{T_{\text{CT}}} \frac{-c}{e^{t\rho}}\,\mathrm{d}t + \int_0^{T_{\text{CT}}} \frac{\delta_{\text{on}}}{e^{t\rho}}\,\mathrm{d}Y_{t+\tau} + \frac{\mathbf{1}_{\mathcal{D}_{\text{CT}}=\text{N}}(PW-I)}{e^{\mathbf{1}_{T_{\text{CT}}>0}(T_{\text{CT}}+\tau)\rho}}\middle|\, \mu_0,\, n_0\right]. \tag{28}$$

The apparent asymmetry between $\mathrm{d}Y_{t+\tau}$ in Eq. (28) and the summand X_t in Eq. (4) 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

$$V^{\pi^*_{\text{CT}}}_{\text{CT}}(\mu_0, n_0) = \sup_{\pi_{\text{CT}}\in\Pi_{\text{CT}}} V^{\pi_{\text{CT}}}_{\text{CT}}(\mu_0, n_0). \tag{29}$$

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.2.1 Continuous time approximation during stage III

The expected reward upon stopping is extended to continuous time by rewriting G in Eq. (7) as:

$$G(y_t,t) = e^{-\mathbf{1}_{t>0}\tau\rho}\mathbb{E}[(PZ_{T_{\text{CT}},\min(T_{\text{CT}},\tau)} - I)^+ \mid \mathcal{F}_{\text{CT},t}]. \tag{30}$$

It is optimal to choose $\mathcal{D}_{\text{CT}} = \text{N}$ if $PZ_{T_{\text{CT}},\min(T_{\text{CT}},\tau)} - I > 0$ and $\mathcal{D}_{\text{CT}} = \text{S}$ otherwise.

S.2.2 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. (8) in continuous time. Following Chernoff (1961), a CT diffusion model approximation of Bellman's equation in Eq. (8) is:

$$B_{\text{CT}}(y_t, t) = \max\Big\{G(y_t, t), \lim_{h\downarrow 0}\Big[-c + \delta_{\text{on}}(y_t/n_t)\Big]h \tag{31a}$$
$$+ e^{-h\rho}\mathbb{E}_{\pi_{\text{CT}}}[B_{\text{CT}}(Y_{t+h}, t+h) \mid \mathcal{F}_{\text{CT},t}]\Big\}, \quad t \in [\tau, T_{\max}),$$
$$B_{\text{CT}}(y_{\max}, T_{\max}) = G(y_{\max}, T_{\max}), \tag{31b}$$

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. (31a) 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. (2)), 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. (31a) and the second maximand, expanding the second maximand in a Taylor series expansion, and applying Ito's Lemma gives

$$\begin{aligned} B_{\text{CT}}(y_t, t) \;=\; & -c + \delta_{\text{on}}\,(y_t/n_t)\,h + (1 - h\rho) \\ & \times\mathbb{E}[B_{\text{CT}}(y_t, t) + U_t B_{\text{CT},y}(y_t, t) + hB_{\text{CT},t}(y_t, t) + U_t^2 B_{\text{CT},yy}(y_t, t)/2] + o(h) \end{aligned} \tag{32}$$

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:

$$0 = -c - \rho B_{\text{CT}} + B_{\text{CT},t} + (B_{\text{CT},y} + \delta_{\text{on}})(y_t/n_t) + \sigma_X^2 B_{\text{CT},yy}/2. \tag{33}$$

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. (31a) are equal and are smoothly matched (Chernoff, 1961). Here, these conditions are:

$$B_{\text{CT}}(y, t) = G(y, t) \text{ on } \partial\mathcal{C} \text{ (free boundary)}; \tag{34a}$$
$$B_{\text{CT},y}(y, t) = G_y(y, t) \text{ on } \partial\mathcal{C} \text{ (smooth pasting)}. \tag{34b}$$

Equations (33) and (34) 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.2.3 Continuous time approximation during stage I

The first term in the maximand of Bellman's equation in Eq. (9) for discrete time is extended to continuous time by using Eq. (30). 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:

$$B_{\text{CT}}(y_0, 0) = \max\left\{ \sup_{u\in[0,\tau]} \left\{H(u) + e^{-u\rho}G(y_0, u)\right\},\ H(\tau) + e^{-\tau\rho}B_{\text{CT}}(y_0, \tau)\right\}. \tag{35}$$

This determines the continuation set on $\mathbb{R} \times [0, \tau)$: let u_y be the smallest u which maximises the supremum in Eq. (35) 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.2.4 Analysis and computation of the PDE for stage II

The analysis for the discrete time stopping problem in Eq. (5) 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. (29) can be reduced to the solution of the continuous time version of Bellman's equation, the free boundary problem in Eq. (33) subject to the implicit boundary conditions in Eq. (34), 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.3 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. Our stage II and stage III analysis constitutes an optimal stopping and free boundary problem which falls 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. (35) and Eq. (28) 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.2.1. Thus, $V_{\text{CT}}^{\pi^*_{\text{CT}}}(\mu_0, n_0)$ in Eq. (29) equals $B_{\text{CT}}(\mu_0 n_0, 0)$ in Eq. (35) for the special case

$\rho = 0$, $c > 0$, $\delta_{\text{on}} = 0$. Moreover, the optimal stopping time is $T_{\text{CT}} \leq \tau$ if the first term in the maximand in Eq. (35) 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. (33) and (34) may be solved using numerical methods.

S.4 Numerical solution of the 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. (29).

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 (with reverse time proportional to $1/n_t$) which take advantage of some standardizations which are more difficult in our context due to the generality (and 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 methods 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 $\partial\mathcal{C}$.

For the special case $\tau = 0$, $T_{\max}$ arbitrarily large, $c = 0$, $\rho > 0$, Chick and Gans (2009) show a close structural relationship between the boundary of the continuation sets for the offline stoppable bandit and the optimal continuation set and Gittins index for the online bandit of Brezzi and Lai (2002). Chick and Gans (2009, Online Companion) also 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.5 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 optimal stopping problem in Eq. (5) 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 (2016) 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 as they are allocated (as is the case here) and compute indices for a beta-binomial model.

S.6 Further simulation results

We present additional analysis of the illustration of section 3, extensions of the sensitivity analysis for the application of section 5, as well as an additional application.

S.6.1 Changing the delay τ for the illustrative simulation of section 3

Figure 2 of the illustrative simulation is plotted assuming that the delay τ (1000 patient pairs), is relatively high compared with $T_{\max}$ (2000 patient pairs). Figure 7 shows the result of halving τ to 500 patient pairs by halving the rate of recruitment, R, leaving all other parameters unchanged.

Comparing Figure 7 with Figure 2, halving the delay means that stage III starts earlier, at an effective sample size of 600. Further, the stage I regions between A and C and D and B that characterise Figure 2 are eliminated, implying that there are no values of μ_0 such that a fixed sample $u \in (0, \tau)$ is an optimal policy. Points A and B remain similar in Figure 7 to their positions in Figure 2, whereas C shifts up to A and D shifts down to B in Figure 7 as compared to their positions in Figure 2. The reason is that, for $t < \tau$, the expected value of information is smaller with fewer pipeline subjects, so that entering Stage II is more advantageous. Above A

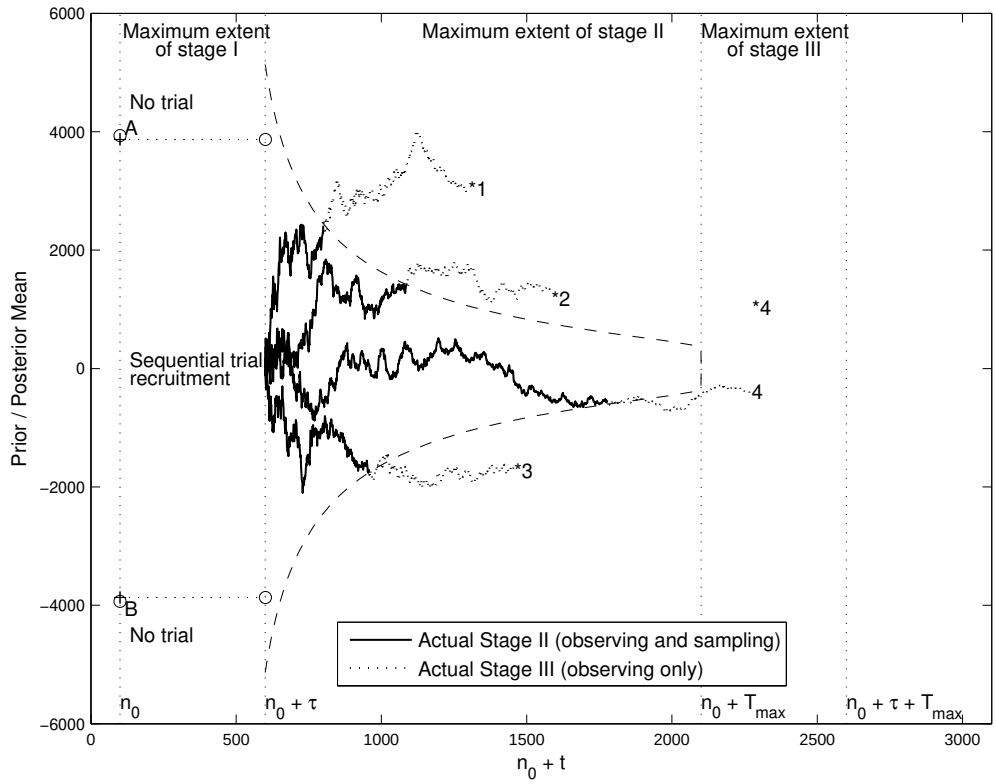


Figure 7: Comparator scenario to that of Figure 2, showing the impact of a smaller τ relative to $T_{\max}$. KEY: '*' value of the sampling mean w_i for each path i; '—-' path of posterior mean when in stage II; '$\cdots$' path of posterior mean when in stage III.

and below B in Figure 7 the number of samples for the Optimal Bayes Sequential and Optimal Bayes One Stage policies coincide at 0.

S.6.2 Further sensitivity analysis for the application of section 5 (known sampling variance)

Figure 8 completes the analysis of the impact of changes in the recruitment rate carried out in section 5.1, by showing how this parameter affects the expected sample size of the trial. In particular, it shows that the greater expected reward of the trial with the largest recruitment rate among those considered ($R = 907$, see Figure 5) is associated with a larger expected sample size over most of the range of μ_0 considered. This is due to the greater value of information implied by a larger number of subjects in the pipeline.

Figure 9 shows the result of increasing c from \$200 to \$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.

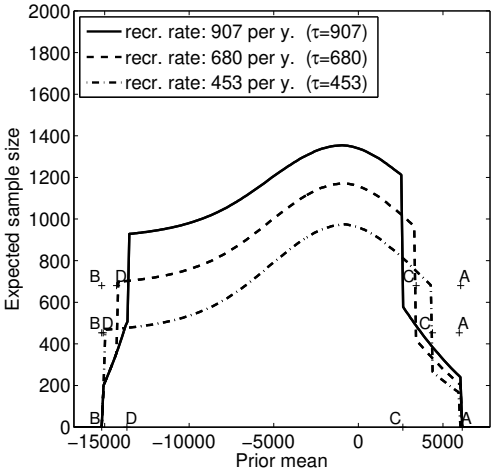


Figure 8: Expected sample sizes for Optimal Bayes Sequential policy for several recruitment rates.

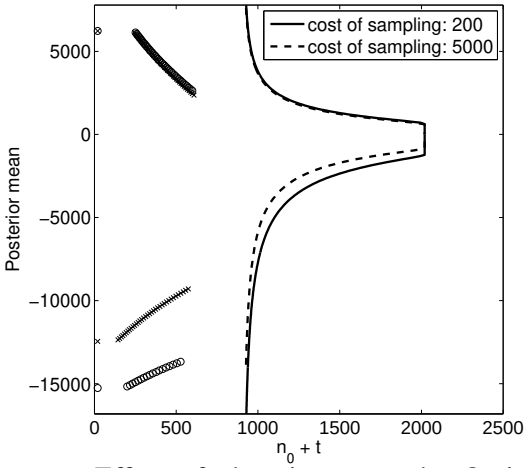


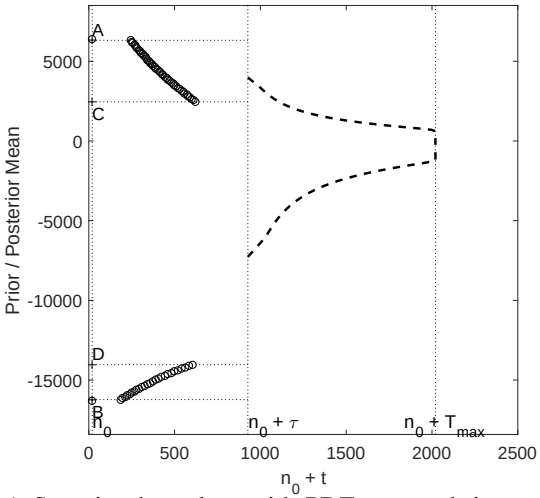
Figure 9: Effect of changing c on the Optimal Bayes Sequential policy. KEY: '$\circ$' c =\$200; '$\times$' c =\$5000.

S.6.3 Further results for the application of section 5 (PDE approach adapted for unknown sampling variance)

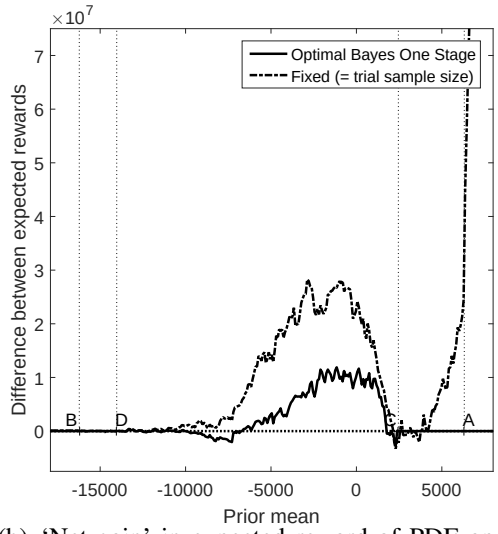
This subsection provides some additional illustrations for section 5.2. Figure 10 replicates Figure 6 in the main paper for a different approach to the computation of the value of continuing during Stage II, among those proposed in Chick et al. (2015). Unlike for KG_*, which is based on a one-stage lookahead estimate of the value of continuation, the solution is based on the PDE approach to Stage II that was described in Appendix S.4. The probabilities associated with the trinomial tree are rescaled according to Eq. (16) to account for the fact that σ_X^2 is unknown. Thus, the trinomial tree allows for greater variation when sampling variances are unknown as compared to when they are known.

We observe that the Stage I stopping boundaries are relatively similar for the KG_* approximation and the PDE approach (compare Figure 6(a) with Figure 10(a)). The upper boundary for stage II sampling is also similar. The PDE approach estimates the value of continuing to sample differently to the KG_* approach, and this leads to a near-vertical line for the optimal stopping boundary upon entering Stage II in this example. This can be seen from the vertical estimated optimal stopping boundary in Figure 10(a) at $n_0 + \tau$ on the horizontal axis, and for prior means in the range $\mu_0 \in [-14000, -7000]$. In that range, one samples τ samples, stops sampling, and selects an alternative after all pipeline data have been observed. The expected number of samples for that range of values of μ_0 may be seen in Figure 10(c). The performance metrics ('net gain' and 'proportion of correct adoption decisions') of the PDE approach are slightly better than those of the KG_* approach for some values of the prior mean, but are otherwise very similar (compare Figure 10(b) and (d) with Figure 6(b) and (d)).

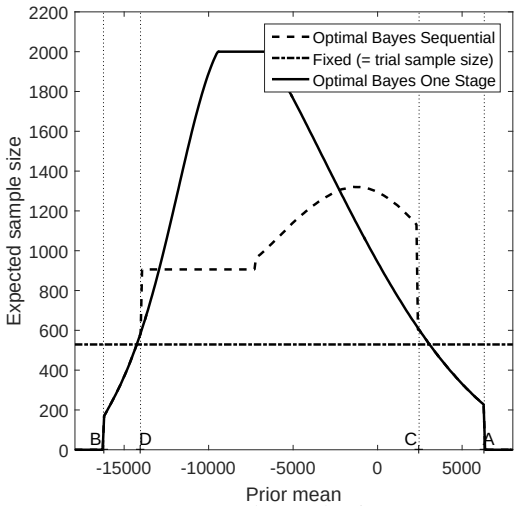
Plug-in estimates for the unknown sampling variance can be used to rescale sample paths of data as in section 5.2. This rescaling is illustrated in Figure 11. The stopping boundaries delimit the KG_* continuation set when computed such that χ_t/ξ_t equals 17358^2, the value of the variance reported in the relevant medical studies, and assuming $\tilde{\rho} = 0$ so as to illustrate the effect of zero discounting in this application. The solid lines plot sample means for several sample paths. They were generated assuming that the unknown means and variances were sampled according to



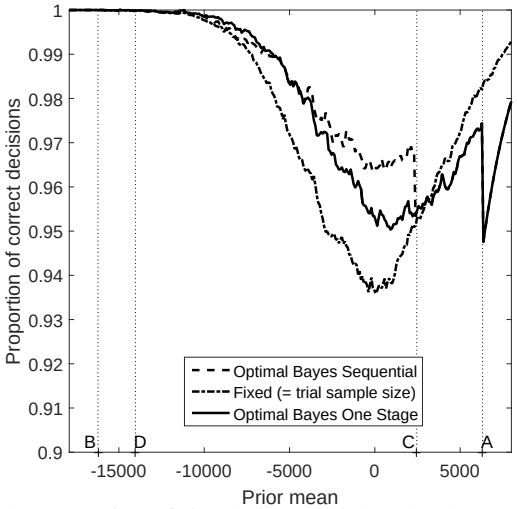
(a) Stopping boundary with PDE approach in section S.6.3 for unknown sampling variance.



(b) 'Net gain' in expected reward of PDE approach over comparator policies.



(c) Expected sample size.



(d) Proportion of simulations which make the correct adoption decision.

Figure 10: Operating characteristics for the stents application of section 5 with PDE approach in section S.6.3 adapted to approximate the Optimal Bayes Sequential policy with unknown sampling variance.

the prior distribution described in section 4. Observations for patient pairs were then generated according to the statistics for each of the 5 sample paths plotted. The circles represent the actual sample means for the 5 simulated studies. The dashed lines represent sample paths scaled by $17358\sqrt{\xi_{\eta t}/\chi_{\eta t}}$. Figure 11 demonstrates that the rescaling of sample paths does not have a huge effect. The larger the shape parameter of the unknown variances, the more certain is the value of the unknown sampling variance and hence the smaller is the rescaling effect on the sample paths. Similarly, the earlier the stopping time, the greater the potential rescaling effect on the sample paths.

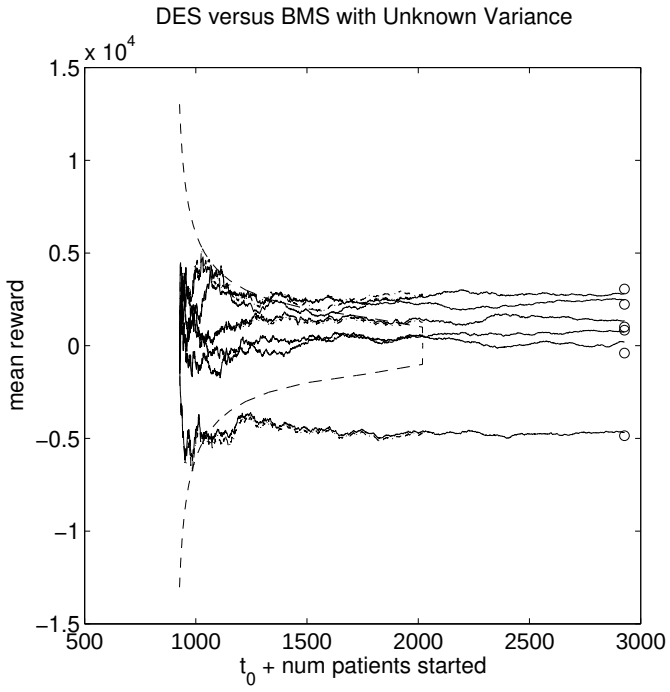


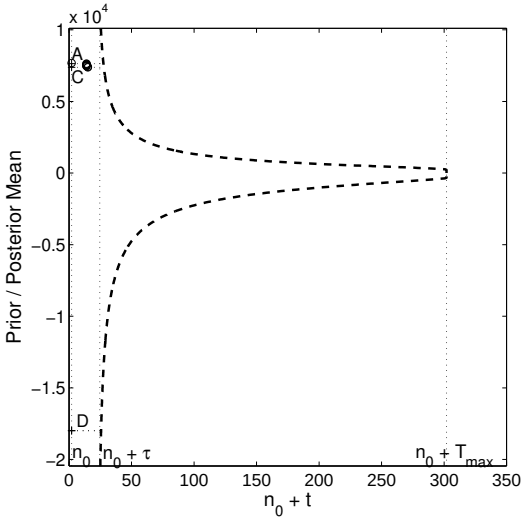
Figure 11: Several sample paths in Stage II for the stents application of section 5 (with discount rate $\tilde{\rho} = 0$) and rescaling of those paths to account for unknown variance.

Parameter	*Value*	*Source*
n_0	2	Assumption
c	£2000	Assumption
λ	£20000/QALY	Assumption
σ_X	£7420.00	Derived from Edlin et al. (2012)
Annual discount rate	0.01	Assumption
P	135000	Assumption
I	£0	Assumption
$T_{\max}$	300	Assumption
End point	QALY	Edlin et al. (2012)
Delay in observing the primary end point	1 year	Costa et al. (2012)
δ_{on}	0	Assumption
Study size (number of pairs)	62	Costa et al. (2012)
Duration of recruitment period	33 months	Costa et al. (2012)
Equivalent annual rate of recruitment R	23	Derived from Costa et al. (2012)
τ	23	Derived from Costa et al. (2012)

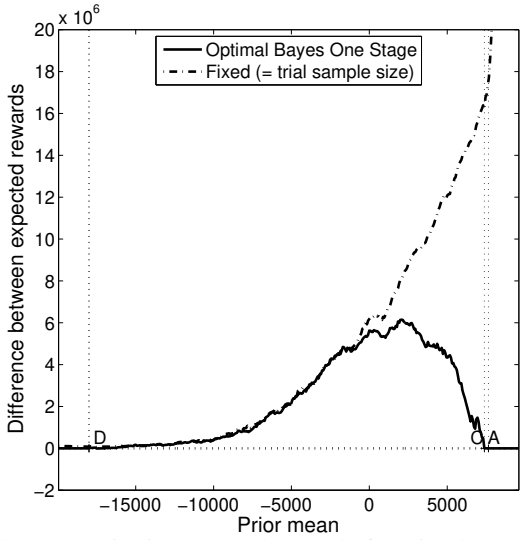
Table 2: Parameter values used for the hip arthroplasty application of section S.6.4.

S.6.4 Additional application: hip arthroplasty

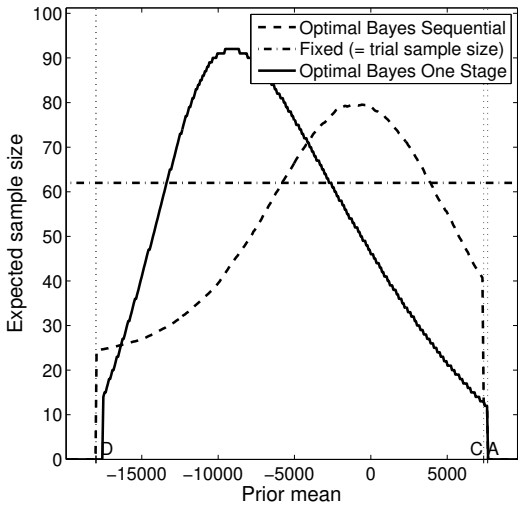
This additional application is based on data from a trial for the comparison of two surgical procedures for hip arthroplasty. The existing technology is total hip anthroplasty (THA), the new technology is resurfacing anthroplasty (RSA). Trial design and clinical outcomes are described in Costa et al. (2012); Edlin et al. (2012) present a cost-effectiveness analysis based on the same



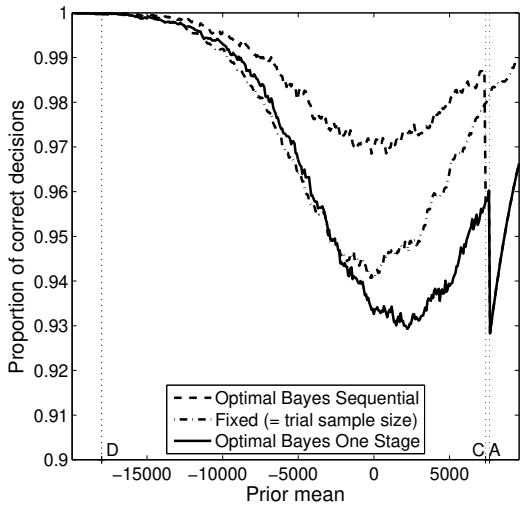
(a) Optimal Bayes Sequential policy.



(b) 'Net gain' in expected reward of Optimal Bayes Sequential over comparator policies.



(c) Expected sample size.



(d) Proportion of simulations which make the correct adoption decision.

Figure 12: Optimal Bayes Sequential policy and operating characteristics for the hip arthroplasty application of section S.6.4.

data. The estimated parameters are summarised in Table 2. The cost-effectiveness study was based on data from 58 patients in the treatment group and 64 in the control group. We approximate the number of pairs to 62. The recruitment period was between May 2007 and February 2010, which implies a recruitment rate of 23 pairs per year. The primary outcome was observed after 1 year. Hence, although the delay, in terms of calendar time, is the same for this application as for the one in section 5, the recruitment rate, and hence the delay in terms of the number of observations, τ, are substantially smaller.

Table 4 of Edlin et al. (2012) reports the mean and 95% confidence interval for incremental

net monetary benefit, which we use to estimate σ_X = £7420. The value of P is estimated assuming a European perspective and is obtained using the following data reported in Edlin et al. (2012): the annual number of hip arthroplasty interventions in the UK is around 45000, of which 6% are RSA. Assuming that an adoption decision today implies that the new technology will be used for the next 5 years, and taking into account that the UK population is about 10% of the European population, we come to the estimate of $P = 45000 \cdot 0.06 \cdot 5 \cdot 10 = 135000$ for the total European market. The sampling cost is assumed equal to £2000. The parameter values are summarised in Table 2.

Results are presented in Figure 12. Figure 12(a) shows that the comparatively small recruitment rate has a large impact on the optimal stopping policy. Since only 23 patients are recruited before the first outcome is observed, the range of values for μ_0 such that the Optimal Bayes One Stage policy is optimal is very narrow in the positive region of the prior mean (distance AC on the vertical axis of Figure 12(a)) and absent in the negative region.

For a value of the prior mean close to zero, Figure 12(b) shows that the expected net gain of the Optimal Bayes Sequential policy over the two alternative policies is approximately £6m. This is a smaller gain, in absolute terms, in comparison with the application of section 5, but it is larger in terms of gain per patient, due to the smaller value of P. This is consistent with what is shown in Figure 5(a), that is, keeping the other parameters fixed, a comparatively small recruitment rate leads to larger gains of the Optimal Bayes Sequential policy over the Optimal Bayes One Stage policy. In terms of the proportion of correct decisions, Figure 12(d) shows that the results are very similar to those of the stents application.

References

Ahuja, V. and Birge, J. (2016). Response-adaptive designs for clinical trials: simultaneous learning from multiple patients. *European Journal of Operations Research*, 248(2):619–633.

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., Forster, M., and Pertile, P. (2015). Optimal sequential sampling with delayed observations and unknown variance. In Yilmaz, L., Chan, W., Moon, I., Roeder, T., Macal, C., and Rossetti, M., editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 3789–3800, Piscataway, NJ. IEEE, Inc.

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.

Costa, M. L., Achten, J., Parsons, N. R., Edlin, R. P., Foguet, P., Prakash, U., Griffin, D. R., et al. (2012). Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial. *BMJ*, 344:e2147.

Edlin, Tubeuf, S., Achten, J., Parsons, N., and Costa, M. (2012). Cost-effectiveness of total hip arthroplasty versus resurfacing arthroplasty: economic evaluation alongside a clinical trial. *BMJ Open*, 2(5):e001162.

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. *JRSS, Series A*, 177(2):419–438.

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