Adaptive Clinical Trial Designs with Surrogates: When Should We Bother?

Arielle Anderer, Hamsa Bastani
Wharton School, Operations Information and Decisions, {aanderer, hamsab}@wharton.upenn.edu

John Silberholz
Ross School of Business, Technology and Operations, josilber@umich.edu

The success of a new drug is assessed within a clinical trial using a *primary endpoint*, which is typically the true outcome of interest, *e.g.*, overall survival. However, regulators sometimes allow drugs to be approved using a surrogate outcome — an intermediate indicator that is faster or easier to measure than the true outcome of interest, *e.g.*, progression-free survival — as the primary endpoint when there is demonstrable medical need. While using a surrogate outcome (instead of the true outcome) as the primary endpoint can substantially speed up clinical trials and lower costs, it can also result in poor drug approval decisions since the surrogate is not a perfect predictor of the true outcome. In this paper, we propose *combining* data from both surrogate and true outcomes to improve decision-making within a clinical trial. In contrast to broadly used clinical trial designs that rely on a single primary endpoint, we propose a Bayesian adaptive clinical trial design that simultaneously leverages *both* observed outcomes to inform trial decisions. We perform comparative statics on the relative benefit of our approach, illustrating the types of diseases and surrogates for which our proposed design is particularly advantageous. Finally, we illustrate our proposed design on metastatic breast cancer. We use a large-scale clinical trial database to construct a Bayesian prior, and simulate our design on a subset of clinical trials. We estimate that our proposed design would yield a 5% increase in trial benefits relative to existing clinical trial designs.

Key words: surrogates, Bayesian adaptive clinical trials, metastatic breast cancer

1. Introduction

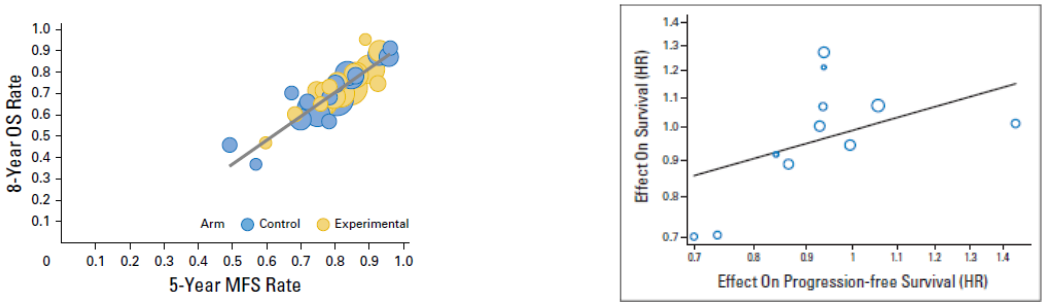
Randomized controlled trials in the medical domain seek to determine a new treatment's efficacy quickly and accurately. Longer clinical trials can delay the release of an effective treatment, incurring significant financial and population health costs. For instance, the average yearly revenue of an approved cancer drug is estimated to be around $600 million; thus, every extra day that an effective drug spends in a clinical trial comes at a cost of millions of dollars in potential revenue to the pharmaceutical company (Prasad and Mailankody 2017). More importantly, the availability of an effective new medical treatment could mean the difference between life and death to patients. Thus, both drug developers and regulatory agencies have incentives to speed up clinical trials. This pressure becomes especially intense in cases where a disease spreads quickly, or when appropriate treatments are lacking. Of the many thousand diseases that are known to affect humans, only

about 500 have a treatment approved by the U.S. Food and Drug Administration (FDA) (Kessler 2016). This mismatch has partly been attributed to the high costs and risks involved in clinical trials, *i.e.*, a novel drug can take 10–15 years and more than $2 billion to develop (NIH 2018). This has spurred the FDA's Accelerated Approval program "for serious conditions that fill an unmet medical need" (FDA 2016).

One key tool to speed up clinical trials is the use of surrogate outcomes. Measuring the true outcome of interest often takes a long time and requires a large patient population (Prentice 1989). Surrogate outcomes are intermediate indicators that are faster or easier to measure than the true outcome, but can reliably predict the efficacy of the treatment with respect to the true outcome of interest. For instance, consider metastatic breast cancer (MBC). The true outcome of interest is typically overall survival duration of each patient compared to a standard therapy. However, this outcome can be difficult to measure because the median overall survival for MBC patients is 21.6 months (Burzykowski et al. 2008); thus, for many patients, it takes 2 years or more after patient recruitment to assess whether the new drug improved overall survival. A common surrogate outcome is progression-free survival, which is the duration of time between starting drug therapy and the progression of the patient's cancer (*i.e.*, an increase in the size or extent of the tumor). In contrast to overall survival, the median duration of progression-free survival is only 7.1 months (Burzykowski et al. 2008). Thus it takes far less time to measure the effect of the new drug on the surrogate outcome, creating an opportunity to significantly reduce the duration of clinical trials.

In general, the success of a treatment is assessed within a clinical trial using a single *primary endpoint* — this can either be the true outcome or the surrogate outcome. The default choice for the primary endpoint is the true outcome, but the FDA's aforementioned Accelerated Approval program allows drugs to be approved using a surrogate outcome as the primary endpoint when there is demonstrable medical need (FDA 2016).

The challenge is that the surrogate outcome is not always a perfect predictor of the true outcome; thus, overreliance on surrogate outcomes may result in poor drug approval decisions. For instance, in the Cardiac Arrhythmia Suppression Trial, investigators approved the drug based on early success on the surrogate (arrhythmia), but the drug actually failed to improve the true outcome of interest (sudden death) (Pratt and Moyé 1995). Indeed, the predictive quality of the surrogate outcome varies greatly depending on the disease and surrogate selected. Figure 1 plots the relationship between surrogate and true outcome pairs across many clinical trials for different drug therapies for prostate cancer (left panel; Xie et al. 2017), and for MBC (right panel; Burzykowski et al. 2008). Evidently, the surrogate outcome is predictive of the true outcome in both cases, but it is a more accurate predictor in the case of prostate cancer compared to metastatic breast cancer.



(a) (b)

Figure 1 Study-level relationship between surrogate and true outcomes published in the medical literature for (a) prostate cancer (Xie et al. 2017); and (b) metastatic breast cancer (Burzykowski et al. 2008).

Thus, when deciding whether to use the surrogate outcome (rather than the true outcome) as the primary endpoint, the FDA must navigate an inherent tension between accelerating clinical trials (by basing approval decisions on quickly observable surrogates) and the risk of approving ineffective or harmful drugs (depending on the likelihood that the surrogate accurately predicts the true outcome). As a result, the FDA has strict criteria for approving the use of a surrogate as the primary endpoint. The main criteria are that the surrogate outcome (i) has a credible clinical relationship with the true outcome, and (ii) is highly predictive of the true outcome (FDA 2018b). While the first criterion is clinical, the second criterion is statistical. Accordingly, a large body of statistical literature has examined whether a surrogate endpoint is a "good enough" predictor of the true outcome of interest to merit being used as the primary endpoint in a clinical trial (see, *e.g.*, Freedman et al. 1992, Buyse and Molenberghs 1998, Burzykowski et al. 2001, Renard et al. 2002, Burzykowski et al. 2004, Daniels and Hughes 1997, Gail et al. 2000, Burzykowski and Buyse 2006, Fleming and DeMets 1996, Weintraub et al. 2015).

We explore the possibility of *combining* data from both surrogate and true outcomes to improve decision-making within a clinical trial. In contrast to existing clinical trial designs that largely rely on a single primary endpoint, we propose a principled Bayesian adaptive clinical trial design that simultaneously leverages both observed outcomes to inform decisions in a clinical trial. We demonstrate that the resulting trial design is more efficient, *i.e.*, the decision-maker can successfully reject ineffective treatments and accept effective treatments more quickly and/or at lower cost. There are several advantages to such an approach. First, the proposed design accounts for uncertainty in the predictive value of the surrogate when making key trial decisions. Instead of simply relying on surrogate outcome data, additionally incorporating a small number of true outcome observations for decision-making can help allay the aforementioned concerns about over-reliance on surrogate outcomes. Second, our design can make use of information from an under-explored source, *i.e.*, surrogates that are only *moderately* predictive of the true outcome (as is the case for MBC, see

Figure 1b). Existing trial designs that use a single primary endpoint fail to leverage moderately predictive surrogates, since a surrogate can only be trusted as a primary endpoint when it is highly predictive. We use metastatic breast cancer as a case study to demonstrate that even moderately predictive surrogates have significant informative value, which can be used to speed up trial decisions when appropriately combined with limited true outcome data. Third, our design improves statistical power by exploiting another untapped source of information: correlations between the surrogate and true outcomes for *individual* patients. In particular, we are likely to have *paired* observations (true and surrogate outcomes) for patients who were recruited early or have higher risk; accounting for such individual-level correlations can reduce the variance of our estimates. Again, existing trial designs fail to leverage correlations between paired outcome observations since all trial decisions are based on a single primary endpoint.

Of course, our proposed design is accompanied by the cost of additional effort and complexity. Our trial requires a Bayesian prior that links the surrogate and true outcomes. We propose constructing this prior from data observed in previous clinical trials for the same disease; estimating such a prior reliably requires significant effort in terms of extracting detailed data from past clinical trials. Furthermore, as with any Bayesian approach, assuming a relationship between the two outcomes introduces a potential source of error in the clinical trial analysis, *i.e.*, a misspecified Bayesian prior may yield incorrect conclusions about the effectiveness of a new therapy. Given these considerations, we study the properties of surrogates and diseases for which our proposed trial design produces a particularly large boost in efficiency. We perform comparative statics and train an interpretable machine learning model to serve as guidelines for when it may be worth adopting our approach despite the additional complexity.

Finally, we illustrate our proposed approach and assess its benefits with a case study on metastatic breast cancer. We use a large-scale MBC clinical trial database to construct a Bayesian prior relating overall survival (true outcome) and progression-free survival (surrogate outcome), and manually collect detailed data on individual patient outcomes from 71 past MBC clinical trials in order to simulate our proposed trial design as well as existing trial designs. Despite the efficiency losses due to the potential misspecification of our learned prior, we estimate a 5% increase in trial benefits (roughly $50 million), which are driven primarily by improved statistical power (*i.e.*, we approve more effective drugs and reject more ineffective drugs). These results suggest that utilizing both true and surrogate outcomes simultaneously in a clinical trial can be valuable compared to relying on a single primary endpoint.

1.1. Contributions

We propose a simple model of a clinical trial under sequential patient recruitment with both surrogate and (delayed) true outcome observations. Following common clinical trial practice, we

model a single intermediate analysis for early stopping; for both the intermediate and final analyses, the respective sample sizes, timing and stopping criteria are determined at the start of the trial. The decision-maker has access to a Bayesian prior that links surrogate and true outcomes both at the *study level* (across clinical trials) and at the *individual level* (for a single patient). Within this framework, we make the following contributions:

1. *Trial Design:* We propose an optimal Bayesian adaptive trial design. In contrast to existing trial designs that determine all trial parameters based on a single primary endpoint (either the true or surrogate outcome), our design leverages observations from *both* outcomes to specify the sample sizes, timing and stopping criteria for the intermediate and final analyses.

2. *Guidelines:* We perform comparative statics and train an interpretable machine learning model on several key properties of diseases and surrogates; these results serve as guidelines for understanding when our proposed design yields significant value relative to traditional designs. We find that our proposed design is particularly advantageous when (i) the surrogate outcome is only moderately predictive of the true outcome across clinical trials, or (ii) there are strong correlations between the surrogate and true outcomes for individual patients.

3. *Case Study:* We illustrate our proposed approach on metastatic breast cancer, a condition where the surrogate is only moderately predictive of the true outcome. We use a large-scale clinical trial database to construct a Bayesian prior linking both outcomes and collect additional detailed data on individual patient outcomes to simulate our proposed trial design. Results suggest that our design would yield a 5% increase in trial benefits relative to traditional designs.

1.2. Related Literature

The design of more efficient clinical trials is well-studied in the healthcare operations literature. We adopt the well-studied Bayesian adaptive clinical trial design (see, *e.g.*, Cheung et al. 2006, Berry et al. 2010, Ahuja and Birge 2016) to improve statistical efficiency and lower costs.

A large literature has developed clinical trial designs that explicitly incorporate salient features of the decision-making process. For example, Chick et al. (2017) study Bayesian trial designs in the presence of delayed outcomes; this is a critical feature in our setting as well, since true outcome observations are delayed relative to surrogate outcome observations. Kouvelis et al. (2017) study a variety of decisions at interim stages of the trial, including the opening of new test centers, setting patient recruitment targets, and stopping decisions. Chick et al. (2018) study trial designs with multiple correlated treatments (*e.g.*, dose-finding trials). Corcoran et al. (2019) study statistical criteria for the drug approval decision based on the results of a clinical trial; while the FDA typically requires clinical trial evidence (based on the primary endpoint) that is statistically significant at the 2.5% level when approving novel drugs, the authors argue that this cutoff should depend on

factors such as disease severity, prevalence, and availability of existing therapies. Bertsimas et al. (2016) build predictive models using historical clinical trial data for advanced gastric cancer drug therapies, with the goal of recommending new combination chemotherapy regimens to be tested in future clinical trials. Our work also relies on historical clinical data to construct our Bayesian prior linking surrogate and true outcomes. Another stream of literature uses bandit algorithms for learning improved treatment regimens. For instance, Negoescu et al. (2017) and Mintz et al. (2017) develop adaptive, personalized treatments by continuously monitoring the patient's state. Lee et al. (2018) develop optimal policies for patient screening when the patient's disease state is continuously evolving but is only partially observable.

The above literature still relies on a single primary endpoint to measure treatment efficacy. In contrast, our paper focuses on incorporating data from multiple outcomes to make key clinical trial decisions. A related literature on co-primary endpoints proposes multiple hypothesis testing procedures to simultaneously examine multiple relevant endpoints within a clinical trial, with the goal of approving the drug if it significantly improves any single endpoint (FDA 2017); we focus on a single true outcome of interest, and surrogates serve only to make quicker or easier inferences about the true outcome.

There is a large literature on defining and evaluating surrogate outcomes. Prentice (1989) first established the statistical definition of a surrogate endpoint, *i.e.*, a surrogate must predict the true outcome well enough to be able to test the null hypothesis on its own. This implies that the surrogate must be *highly* correlated with the true outcome, and be affected by treatment through the same clinical pathway as the true outcome (Fleming and DeMets 1996). Burzykowski and Buyse (2005) refers to surrogate endpoints as "replacement" endpoints, since they are used to replace the true outcome of interest as the primary endpoint of clinical trials; the authors present different validation and evaluation procedures used to determine whether the surrogate is precise enough to act as a replacement. These validation criteria have been further explored by Freedman et al. (1992), Buyse and Molenberghs (1998), Burzykowski et al. (2001), Renard et al. (2002), Burzykowski et al. (2004), Spiegelhalter et al. (2004), and Weintraub et al. (2015). Our proposed trial designs allay many of the concerns raised in this literature since we do not simply use surrogate outcomes as replacement endpoints; rather, we make trial decisions by combining knowledge from a small number of true outcome observations and a large number of surrogate outcome observations, while explicitly accounting for the uncertainty in the predictive value of the surrogate. As a consequence, we can relax some of the criteria proposed above, and derive significant informative value from a much broader class of surrogates that are only *moderately* predictive of the true outcome.

More related to our work, a few papers have studied clinical decision-making strategies that incorporate both surrogate and true outcomes. For example, Pozzi et al. (2016) propose a decision-making aid in multiple sclerosis drug development that simultaneously incorporates both surrogate and true outcome observations in a Bayesian hierarchical model; however, they do not study clinical trial designs. Renfro et al. (2012) suggests a trial design that evaluates whether the correlation observed between surrogate and true outcome observations within some initial phase of the clinical trial matches the expected relationship from the literature; if so, the surrogate outcome can be used as a primary endpoint, and if not, the decision-maker would default to using the true outcome. Such a design still ultimately relies on a single primary endpoint, and does not incorporate the uncertainty of the surrogate's predictive power in a principled way. Closest to our work, Berry (2004) and Han (2005) discuss Bayesian trials where surrogate outcome observations are used to predict not-yet-observed true outcomes within a clinical trial to aid decision-making. However, they do not posit a trial design, *i.e.*, specifying sample sizes, timing and stopping criteria that leverage both surrogate and true outcome observations. Our work bridges this gap, and it furthermore identifies the types of diseases and surrogates where such a design can yield the most value relative to traditional trial designs.

Finally, we rely on the clinical trial meta-analysis literature (see, *e.g.*, Daniels and Hughes 1997, Gail et al. 2000, Burzykowski and Buyse 2006). Meta-analyses provide crucial information for our proposed trial design, including study-level correlations (across many clinical trials on the same disease) and individual-level correlations (across outcomes for a single patient using individual patient data) between surrogate and true outcomes.

2. Model

A huge variety of clinical trial designs have been proposed and implemented in practice, mirroring the diversity of outcomes studied in the medical literature and the variability of trial designer needs. In this section, we establish and provide the rationale for the model of clinical trial design that we study in this work, culminating in an optimal trial design that takes into account both surrogate and true outcomes of enrollees.

2.1. Outcomes and Effect Sizes

Clinical trials outcomes can be of a number of types, including binary valued, continuous valued, count, categorical, or time to event. To simplify the exposition for our proposed clinical trial designs, we limit discussion in Sections 2–3 to continuous-valued true and surrogate outcomes with known sampling variance. Nearly half of recent Bayesian clinical trials used a continuous primary endpoint (Lee and Chu 2012), such as blood counts when treating blood abnormalities, blood pressure when treating heart failure, and forced expiratory volume when treating breathing conditions. Further,

continuous outcomes with known standard variances are frequently studied in the clinical trial design literature (see, *e.g.*, Chick et al. 2017, 2018). Similar clinical trial design ideas can be applied for other types of outcomes; the case of time-to-event surrogate and true outcomes is considered in Appendix C, and we study the effectiveness of the resulting designs for metastatic breast cancer in Section 4.

Keeping with the motivation that a surrogate outcome should be faster to measure than a true outcome, we assume that the surrogate outcome is measured immediately upon patient enrollment in the study,[1] while the true outcome is observed with a delay of Δ time units. As a result, with continuous patient enrollment, we have two patient groups at any point of time: one for which we have observed both surrogate and true outcomes, and one for which we have only observed the surrogate outcome so far.

As is typical in clinical trials with continuous outcomes, we will measure the effect size of a new treatment using the *mean difference*, defined as the mean of the outcome in the treatment group subtracted by the mean of the outcome in the control group. We define the (unknown) effectiveness of the treatment versus the control by vector

$$\boldsymbol{\mu} = [\mu_S \ \mu_T]',$$

where μ_S is the surrogate outcome effect size and μ_T (the value we seek to identify) is the true outcome effect size. For any sufficiently large set of n patients in the trial — who are randomly assigned with equal probability between the control and treatment arms — we model the effect size estimate among those n patients, $\hat{\mathbf{e}} = [\hat{e}_S, \hat{e}_T]'$, as approximately bivariate normally distributed:

$$\begin{aligned} \hat{\mathbf{e}} &\sim \mathcal{N}(\boldsymbol{\mu}, 4\boldsymbol{\Sigma}_I/n) \\ \boldsymbol{\Sigma}_I &= \begin{bmatrix} \sigma_{IS}^2 & \rho_I \sigma_{IS} \sigma_{IT} \\ \rho_I \sigma_{IS} \sigma_{IT} & \sigma_{IT}^2 \end{bmatrix}, \end{aligned}$$

with surrogate sampling variance σ_{IS}^2, true outcome sampling variance σ_{IT}^2, and correlation ρ_I. This approximate bivariate normality follows directly from the central limit theorem.

Key assumption: In this work we adopt a Bayesian framework, and we assume that the effect size $\boldsymbol{\mu}$ for a given study is drawn from a bivariate normal distribution:

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \\ \boldsymbol{\Sigma}_0 &= \begin{bmatrix} \sigma_{0S}^2 & \rho_0 \sigma_{0S} \sigma_{0T} \\ \rho_0 \sigma_{0S} \sigma_{0T} & \sigma_{0T}^2 \end{bmatrix}, \end{aligned}$$

[1] One could additionally model a delay for the surrogate outcome, but this can be absorbed as a constant per-patient waiting cost in our model.

with variance σ_{0S}^2 for the surrogate effect size, variance σ_{0T}^2 for the true outcome effect size, and correlation ρ_0. We collected effect sizes across clinical trials for 30 different true and surrogate outcome pairs from 19 different diseases in the medical literature (for details, see Appendix E), revealing that this assumption is often but not always warranted — the Henze-Zirkler test failed to reject the null hypothesis of bivariate normality for 22 of the 30 pairs ($p > 0.05$), but rejected the null hypothesis in the remaining eight cases. We adopt the bivariate normality assumption both for analytic convenience (it provides a clean characterization of the historical link between the surrogate and true outcome and makes the comparative statics tractable) and because the resulting clinical trial designs often work well even when bivariate normality is violated (we illustrate this on our metastatic breast cancer case study in Section 4).

REMARK 1. We model two different types of correlations between surrogate and true outcomes, *i.e.*, the study-level correlation ρ_0 linking the two effect sizes in a clinical trial, and the individual-level correlation ρ_I linking the two outcomes for an individual patient. ρ_0 determines how much the decision-maker can infer about the true outcome effect size μ_T given surrogate outcome observations, *e.g.*, if $|\rho_0| = 1$, then surrogate outcome observations alone are sufficient to estimate μ_T. In contrast, ρ_I determines whether the decision-maker can leverage *paired* true and surrogate outcome observations for individual patients to reduce the effective variance of true outcome observations. For example, even if $\rho_0 = 0$, if $|\rho_I| = 1$, then paired outcomes for a single patient is sufficient to estimate $\mu_T - \mu_S$. This implies that we can estimate μ_T using a large number of surrogate outcome observations (to estimate μ_S) and a single individual's paired outcome observations (to estimate $\mu_T - \mu_S$). This distinction will be a key feature of our analysis in Section 3. In practice, the magnitudes $|\rho_0|$ and $|\rho_I|$ are somewhat positively correlated, but can also vary significantly; see Figure 4 in Section 3.3 for a scatterplot of both correlations for different true and surrogate outcome pairs that we collected from 65 meta-analyses in the medical literature.

2.2. Clinical Trial Structure and Objectives

The central analytical objective of this work is to compare the efficacy of a new clinical trial design that combines surrogate and true outcome observations against existing clinical trial designs that rely on a single primary endpoint. To do so, we will define three different types of inference for decision-making within clinical trials. The first, **Type A**, makes inference using patient true outcomes only, and represents the standard trial design used in practice. The second, **Type B**, makes inference using patient surrogate outcomes only, and represents a standard design in settings where the surrogate has been deemed of sufficiently high quality to act as the primary endpoint, *e.g.*, clinical trials in FDA's Accelerated Approval program. The third, **Type C**, makes inference using both patient surrogate and true outcomes; this is our proposed design. For simplicity, patients

Outcome Parameters (Exogenous)	
μ_{0S}	Study-level mean surrogate effect size
μ_{0T}	Study-level mean true outcome effect size
σ^2_{0S}	Study-level surrogate effect size variance
σ^2_{0T}	Study-level true outcome effect size variance
ρ_0	Study-level correlation of surrogate and true outcome effect size
σ^2_{IS}	Individual-level surrogate effect size sampling variance
σ^2_{IT}	Individual-level true outcome effect size sampling variance
ρ_I	Individual-level correlation of surrogate and true outcome effect size
Δ	Number of time units between enrolling a patient and observing their true outcome
Economic Parameters (Exogenous)	
a	Population monetary benefit of a unit increase in the true outcome effect size
$-b$	Threshold of population monetary benefit needed to accept a new treatment
c_p	Monetary cost of enrolling one patient in a clinical trial
c_w	Monetary cost of waiting one additional unit of time
n^{max}	Maximum number of patients budgeted for in the clinical trial
Trial Design Parameters (Endogenously Selected by Designer Before Study)	
n	Target number of patients enrolled (one per time unit)
t_1	Intermediate analysis time
t_2	Final analysis time
Algorithm Parameters (Constructed by Algorithm 1)	
δ_t	Parameter used to determine whether to stop early
$\hat{\boldsymbol{\mu}}_{1,t}$	Estimate of $\boldsymbol{\mu}$ at intermediate analysis time for trial type $t \in \{A, B, C\}$
$\hat{\boldsymbol{\mu}}_{2,t}$	Estimate of $\boldsymbol{\mu}$ at final analysis time for trial type $t \in \{A, B, C\}$
Theory Parameters (Used in Theorem 1)	
$\tilde{\sigma}^2_{1,t,T}$	Variance of $\hat{\mu}_{1,t,T}$ (true outcome posterior mean at the intermediate analysis) for trial type $t \in \{A, B, C\}$
$\tilde{\sigma}^2_{2,t,T}$	Variance of $\hat{\mu}_{2,t,T}\|\mathcal{F}_{1,t}$ (true outcome posterior mean at the final analysis, given all information available at time t_1) for trial type $t \in \{A, B, C\}$
$\tilde{\sigma}^2_{3,t,T}$	Variance of $\hat{\mu}_{2,t,T}$ (true outcome posterior mean at the final analysis) for trial type $t \in \{A, B, C\}$
$\mathcal{F}_{1,t}$	Filtration of information available at time t_1 for trial type $t \in \{A, B, C\}$
$\mathcal{F}_{2,t}$	Filtration of information available at time t_2 for trial type $t \in \{A, B, C\}$

Table 2.1 Parameters in model of clinical trials with continuous outcomes

are recruited sequentially. Following standard clinical trial guidelines (Sydes et al. 2004), the trial designer must pre-specify all trial design parameters before the start of the study, *i.e.*, the patient sample size, as well as the timing and early stopping conditions for any intermediate analyses.

Trial Design: We use Bayesian inference for each of our trial designs, beginning with the shared study-level prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. We further use the patient-level distribution $\hat{\mathbf{e}} \sim \mathcal{N}(\boldsymbol{\mu}, 4\boldsymbol{\Sigma}_I/n)$ for the effect size estimate $\hat{\mathbf{e}}$ among n patients to perform Bayesian updates; as noted earlier, we have two patient groups at any point: one for which we have already observed both surrogate and true outcomes, and one for which we have only observed the surrogate outcome so far. Details of the Bayesian updates are provided in Appendix A, and key notation is summarized in Table 2.1.

Furthermore, each of our trial designs is adaptive, in order to distinguish the (well-studied) benefits of early stopping from the benefits of integrating information from both outcomes. In practice, as part of the ethical responsibilities of clinical trial planners, it is recommended that trials be periodically reviewed by a Data and Safety Monitoring Board (DSMB) to determine if the trial should be continued, modified, or stopped early. The DSMB considers many aspects of the trial

when making these determinations, including efficacy (is the drug clearly effective or ineffective?), safety (do toxicities outweigh the potential benefits?), slow accrual, poor data quality or treatment adherence, and results of other studies making this study unnecessary or unethical (Piantadosi 2005). We model DSMBs as performing planned intermediate analyses for drug efficacy. However, such analyses are costly since they require the DSMB members to meet and discuss the trial (Sydes et al. 2004). Thus, we limit ourselves to one planned intermediate analysis. This matches well with the reality of clinical trials today — 65% of cancer randomized clinical trials have no more than one intermediate analysis (Floriani et al. 2008). Our approach straightforwardly generalizes to trial designs with multiple intermediate analyses as well.

To operationalize early stopping in our trial designs, we follow a classical group sequential approach to intermediate analysis, assuming the trial enrolls n patients (one per time unit) with an interim analysis at time t_1 and a final analysis at time t_2. The trial designer selects n, t_1, and t_2 before the study commences. A multitude of other interim analysis designs have been proposed in the literature, including designs with a flexible number and timing of intermediate analyses (Lan and DeMets 1983, Wang and Tsiatis 1987) and designs employing continuous monitoring (Chick et al. 2017, 2018). Though these more advanced designs are attractive due to their ability to improve overall trial outcomes, we use the classical group sequential design in this work both to maintain analytical tractability and because these designs are broadly used in clinical trials today due to their simplicity, fit with the DSMB decision-making process, and ease of maintain blinding in double-blind experiments (Sydes et al. 2004, Tharmanathan et al. 2008).

Objective: When designing a clinical trial, historically planners have most typically planned to enroll sufficient patients to ensure they have some specified power $1-\beta$ to detect a target effect size, while making Type I errors at no more than α rate. However, recently there has been increasing interest in integrating cost-effectiveness into clinical trial design rather than focusing purely on Type I/II errors (Brennan et al. 2006, Hampson and Jennison 2013, NICE 2018). Following Chick et al. (2017) and others, we use a Health Technology Assessment (HTA) objective, which explicitly models the value of trial results (*e.g.*, measured by assigning a monetary value to patient quality-adjusted life years, multiplied by the size of the intended patient population) and the cost of sampling (*e.g.*, per-patient trial enrollment costs) in a Bayesian decision-theoretic framework. In particular, let $a > 0$ be the population or market-level monetary benefit of a unit increase in the true outcome effect size, and $b \leq 0$ be the base cost of accepting a new treatment (*i.e.*, we may not wish to use a new drug unless its benefit compared to an established therapy exceeds some threshold level). Thus, approving a drug with true outcome effect size μ_T yields a value of $a\mu_T + b$ to the trial designer. Next, let c_p be the cost of enrolling an additional patient, and c_w be the cost of waiting an additional unit of time for a trial decision (see Table 2.1). Let the number of enrolled

patients $m_1 = \min(t_1, n)$ if the trial is stopped early, and $m_2 = n$ if the trial is fully completed. Then, the full HTA objective is

$$(a\mu_T + b) \cdot \mathbb{I}[\text{approve drug}] - c_p m_j - c_w t_j \,,$$

where $j = 1$ if the trial stops early and $j = 2$ otherwise. All trial design parameters are chosen to maximize the expected HTA objective of the trial.

2.3. Optimal Clinical Trial Design

With the preliminaries of the model in place, we derive the optimal structure for all three clinical trials in Algorithm 1, which takes the trial type $t \in \{A, B, C\}$ as an input parameter. At a high level, we begin by sequentially enrolling $m_1 = \min(t_1, n)$ patients before the intermediate analysis at time t_1. At this point, we use the true and surrogate outcome observations thus far to obtain posterior estimate $\hat{\boldsymbol{\mu}}_{1,t}$ of $\boldsymbol{\mu}$; see Appendix A for the exact Bayesian update formula. We use $\hat{\boldsymbol{\mu}}_{1,t}$ to decide whether to stop the trial early, and either approve or reject the drug immediately. This decision requires an additional threshold parameter δ_t, which can be computed using root finding (see Appendix B). If our estimate of μ_T is not too large or too small, then we continue the trial by sequentially enrolling additional patients for a total of $m_2 = n$ patients. We conduct our final analysis at time $t_2 \geq n$, where we again use the true and surrogate outcome observations thus far to obtain posterior estimate $\hat{\boldsymbol{\mu}}_{2,t}$ of $\boldsymbol{\mu}$ (see Appendix A) and to decide whether to approve or reject the drug.

Note that the trial design parameters (*i.e.*, target patient recruitment n, and the timing of the intermediate and final analyses t_1 and t_2) have to be specified at the start of the trial. These values should be chosen to maximize the expected HTA objective (see Theorem 1) and vary by trial type.

THEOREM 1. *Consider a clinical trial of type $t \in \{A, B, C\}$. Algorithm 1 maximizes the expected HTA objective, which is given by $b_t = b_t^{trial} + b_t^{eff} + b_t^{fut}$. Here, b_t^{trial} is the expected HTA objective for a trial that is always run to the final analysis, b_t^{eff} is the expected incremental HTA objective improvement from early stopping to make an approval decision for a likely effective drug, and b_t^{fut} is the expected incremental HTA objective improvement from early stopping to make a rejection decision for a likely ineffective drug. These quantities can be computed as*

$$\begin{aligned}
b_t^{trial} &= a\tilde{\sigma}_{3,t,T}[\phi(\tau_t) - \tau_t(1 - \Phi(\tau_t))] - c_p n - c_w t_2 \,, \\
b_t^{eff} &= \int_{-b/a+\delta_t}^{\infty} B_t(\alpha_t(x)) \frac{1}{\sqrt{2\pi\tilde{\sigma}_{1,t,T}^2}} \exp\left(\frac{-(x-\mu_{0,T})^2}{2\tilde{\sigma}_{1,t,T}^2}\right) dx \,, \\
b_t^{fut} &= \int_{-\infty}^{-b/a-\delta_t} B_t(\alpha_t(x)) \frac{1}{\sqrt{2\pi\tilde{\sigma}_{1,t,T}^2}} \exp\left(\frac{-(x-\mu_{0,T})^2}{2\tilde{\sigma}_{1,t,T}^2}\right) dx \,,
\end{aligned}$$

Algorithm 1 Bayesian clinical trial design

Input: Trial type $t \in \{A, B, C\}$; outcome parameters $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_I$, and Δ; economic parameters a, b, c_p, c_w, and n^{max}; trial design parameters n, t_1, and t_2
for $i \in \{1, \ldots, \min(t_1, n)\}$ **do**
  Enroll patient i
end for
$\hat{\boldsymbol{\mu}}_{1,t} \leftarrow$ Posterior mean given observations $\hat{\mathbf{e}}_{1,1}, \hat{e}_{1,2,S}$, and $\hat{e}_{1,3,S}$ for trial type t (see Appendix A)
Compute parameter δ_t using root finding (see Appendix B)
if $\hat{\mu}_{1,t,T} > -b/a + \delta_t$ **then**
  return Approve
else if $\hat{\mu}_{1,t,T} < -b/a - \delta_t$ **then**
  return Reject
end if
for $i \in \{\min(t_1, n) + 1, ..., n\}$ **do**
  Enroll patient i
end for
$\hat{\boldsymbol{\mu}}_{2,t} \leftarrow$ Posterior mean given observations $\hat{\mathbf{e}}_{2,2}, \hat{e}_{1,2,T}$, and $\hat{e}_{2,3,S}$ for trial type t (see Appendix A)
if $\hat{\mu}_{2,t,T} > -b/a$ **then**
  return Approve
else
  return Reject
end if

where we have defined

$$\tau_t = -\frac{a\mu_{0,T} + b}{a\tilde{\sigma}_{3,t,T}}, \quad \alpha_t(x) = -\frac{ax + b}{a\tilde{\sigma}_{2,t,T}},$$
$$B_t(x) = c_p \max(n - t_1, 0) + c_w(t_2 - t_1) - a\tilde{\sigma}_{2,t,T}\left(\phi(x) + x\Phi(x) - x^+\right),$$

$\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF, respectively, and the remaining parameters are defined in Table 2.1.

The proof of Theorem 1 is provided in Appendix B, and characterizes the expected HTA objective of a clinical trial as a function of the trial design parameters n, t_1, and t_2. Since $n \in [0, n^{max}]$, $t_1 \in [0, n + \Delta]$, and $t_2 \in [\max(n, t_1), n + \Delta]$, the number of integer parameter settings is $O(n^{max}\Delta(n^{max} + \Delta))$, and the optimal values for n, t_1, and t_2 can be found by enumeration for typical values of n^{max} and Δ.

3. Analyzing Trial Benefits

As discussed earlier, our proposed trial design requires nontrivial overhead costs for implementation in practice. Thus, we seek to identify the properties of diseases and surrogates where our proposed design can offer the most value relative to existing designs.

3.1. Preliminaries

We begin by studying how the HTA objectives of each of the three trial designs vary as a function of key properties of the disease and surrogate. This will allow us to identify regimes where our proposed design is particularly advantageous relative to existing designs in the next subsection.

Simplification: The full trial design does not yield tractable comparative statics, so we make some simplifying assumptions; we numerically verify that the resulting insights hold for the full designs. First, we consider a setting without an intermediate analysis;[2] *i.e.*, we only analyze the HTA objective given by b_t^{trial} as defined in Theorem 1. Numerical results match the resulting comparative statics closely, suggesting that the benefits of early stopping are somewhat orthogonal to the benefits of combining surrogate and true outcomes. Second, we constrain our trials to observe the key inferential outcome for all enrolled patients; this does not affect Type B trials (since surrogate outcomes are observed immediately upon study enrollment), but constrains Type A and C trials to wait Δ time steps after the final recruited patient (to observe the resulting true outcomes) before making a decision.[3] Finally, we maintain an unbiased expectation of the surrogate and true outcome effect sizes in our prior ($\boldsymbol{\mu}_0 = 0$) and normalize the variances of the true and surrogate outcome effect sizes ($\sigma_{0S} = \sigma_{0T} = 1$);[4] we also take $b = 0$, implying that any positive true outcome effect size is sufficient for an approval decision.

Parameters of Interest: We are interested in comparative statics with respect to the market size a, the delay Δ, the study-level correlation ρ_0, and the individual-level correlation ρ_I. We additionally define the following parameters governing key statistical properties of the disease and the surrogate:

$$R_S = \sigma_{IS}/\sigma_{0S}\,, \quad R_T = \sigma_{IT}/\sigma_{0T}\,.$$

The ratios R_S and R_T capture the ease of measuring the surrogate and true outcomes respectively, *i.e.*, smaller values of R_S and R_T imply that fewer observations are needed to accurately estimate the respective effect sizes.

LEMMA 1 **(Type A Trials)**. *The following comparative statics are for the HTA objective b_A^{trial} of a trial on n patients that performs inference on only true outcomes.*

1. *The objective does not depend on any of the surrogate properties, including the individual-level correlation ρ_I, study-level correlation ρ_0, or the surrogate ratio R_S:*

$$\frac{db_A^{trial}}{d\rho_I} = \frac{db_A^{trial}}{d\rho_0} = \frac{db_A^{trial}}{dR_S} = 0\,.$$

2. *The objective is monotonically decreasing in the delay Δ and the true outcome ratio R_T:*

$$\frac{db_A^{trial}}{d\Delta} = -c_w \quad \textit{and} \quad \frac{db_A^{trial}}{dR_T} = -\sqrt{\frac{2}{\pi}} \cdot \frac{4anR_T}{\left(n + 4R_T^2\right)^2}\,.$$

[2] It is worth noting that 43% of cancer randomized clinical trials have no intermediate analyses, making this setting relevant to practice as well (Floriani et al. 2008).

[3] The same assumption is made in Chick et al. (2017), *i.e.*, the decision-maker must wait to observe all outcomes for the "pipeline subjects" who have been treated but whose outcomes have yet to be observed.

[4] These two assumptions are without loss of generality and can be achieved by scaling and shifting the effect sizes.

3. *The objective is monotonically increasing in the market size a:*

$$\frac{db_A^{trial}}{da} = \sqrt{\frac{1}{2\pi}} \cdot \frac{n}{n + 4R_T^2}.$$

Lemma 1 follows directly from the definition of b_A^{trial} provided in Theorem 1 and the fact that $\tilde{\sigma}_{3,A,T} = \sigma_{0,T}^4/(\sigma_{0,T}^2 + 4\sigma_{I,T}^2/o)$, where o is the total number of observed outcomes in the trial. Since Type A trials do not utilize surrogate outcomes, surrogate properties do not play any role in determining the HTA objective of the trial. Larger values of the delay Δ imply larger waiting costs to achieve the same statistical power; similarly, larger values of the true outcome ratio R_T imply that one requires greater patient enrollment to achieve the same statistical power. As a result, both Δ and R_T result in more costly trials. A larger market a implies more profits to be reaped from a successful trial.

LEMMA 2 **(Type B Trials)**. *The following comparative statics are for the HTA objective b_B^{trial} of a trial on n patients that performs inference on only surrogate outcomes.*

1. *The objective does not depend on the individual-level correlation ρ_I, the delay Δ, or the true outcome ratio R_T:*

$$\frac{db_B^{trial}}{d\rho_I} = \frac{db_B^{trial}}{d\Delta} = \frac{db_B^{trial}}{dR_T} = 0.$$

2. *The objective is monotonically decreasing in the surrogate outcome ratio R_S:*

$$\frac{db_B^{trial}}{dR_S} = -\sqrt{\frac{2}{\pi}} \cdot \frac{4an\rho_0^2 R_S}{\left(n + 4R_S^2\right)^2}.$$

3. *The objective is monotonically increasing in the magnitude of the study-level correlation $|\rho_0|$ and in the market size a:*

$$\frac{db_B^{trial}}{d\rho_0} = \sqrt{\frac{2}{\pi}} \cdot \frac{an\rho_0}{n + 4R_S^2} \quad \textit{and} \quad \frac{db_B^{trial}}{da} = \sqrt{\frac{1}{2\pi}} \cdot \frac{n\rho_0^2}{n + 4R_S^2}.$$

Lemma 2 follows directly from the definition of b_B^{trial} provided in Theorem 1 and the fact that $\tilde{\sigma}_{3,B,T} = (\rho_0^2\sigma_{0,S}^2\sigma_{0,T}^2)/(\sigma_{0,S}^2 + 4\sigma_{I,S}^2/n)$. Since Type B trials do not utilize true outcomes, true outcome properties do not play any role in determining the HTA objective of the trial. Larger values of the surrogate outcome ratio R_S imply that one requires greater patient enrollment to achieve the same statistical power. Larger values of the study-level correlation $|\rho_0|$ imply lower Type I/II errors when using only surrogates, resulting in better trial decisions. A larger market a implies more profits to be reaped from a successful trial.

Next, we consider Type C trials. Since these trials are especially complex, we study comparative statics for clinical trials where patient enrollment n is large relative to the other trial parameters.

REMARK 2. More specifically, as detailed in Appendix D, we consider the regime where the maximal eigenvalue of the inverse study-level Bayesian prior $\lambda_{\max}(\mathbf{\Sigma}_0^{-1})$ is small relative to n. First, this rules out the regime where the study-level prior is very informative, *i.e.*, σ_{0S} and σ_{0T} are very small. This choice matches practice since the FDA favors uninformative or objective priors in order to avoid "assuming the result" (see, *e.g.*, Lee and Chu 2012). Second, this rules out the regime with extremely high study-level correlations, since $\lambda_{\max}(\mathbf{\Sigma}_0^{-1}) \to \infty$ when $|\rho_0| \to 1$. However, a decision-maker can safely rely on Type B trials in this setting; in contrast, we are interested in the regime where the surrogate outcome is not a perfect predictor of the true outcome, and accurate inference requires a moderate number of true outcome observations.

LEMMA 3 **(Type C Trials)**. *The following comparative statics are for the HTA objective b_C^{trial} of a large trial on n patients that performs inference on both outcomes.*

1. *The objective is monotonically increasing in the market size a to the highest order in the sample size n:*

$$\frac{db_C^{trial}}{da} = \sqrt{\frac{1}{2\pi}} + \mathcal{O}\left(\frac{1}{n}\right).$$

2. *The objective is monotonically decreasing in the delay Δ and the true outcome ratio R_T to the highest order in the sample size n:*

$$\frac{db_C^{trial}}{d\Delta} = -c_w \quad \textit{and} \quad \frac{db_C^{trial}}{dR_T} = -\sqrt{\frac{2}{\pi}} \cdot \frac{4aR_T}{n} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

3. *The objective is non-monotonic in the individual-level correlation $|\rho_I|$, study-level correlation $|\rho_0|$, and the surrogate outcome ratio R_S to the highest order in the sample size n:*

$$\begin{aligned}
\frac{db_C^{trial}}{d\rho_I} &= \sqrt{\frac{2}{\pi}} \frac{16aR_S R_T^2(\rho_I R_S - \rho_0 R_T)}{n^2(1-\rho_0^2)} + \mathcal{O}\left(\frac{1}{n^3}\right) \\
\frac{db_C^{trial}}{d\rho_0} &= -\sqrt{\frac{2}{\pi}} \frac{16aR_T^2(\rho_I R_S - \rho_0 R_T)(R_T - \rho_0\rho_I R_S)}{n^2(1-\rho_0^2)^2} + \mathcal{O}\left(\frac{1}{n^3}\right) \\
\frac{db_C^{trial}}{dR_S} &= \sqrt{\frac{2}{\pi}} \frac{16a\rho_I R_T^2(\rho_I R_S - \rho_0 R_T)}{n^2(1-\rho_0^2)} + \mathcal{O}\left(\frac{1}{n^3}\right)
\end{aligned}$$

The proof of Lemma 3 is provided in Appendix D, and follows from performing a Taylor expansion around n on b_C^{trial} defined in Theorem 1, yielding $b_C^{trial} = a/\sqrt{2\pi}(1 - 4R_T^2/n + 16R_T^2(\rho_I^2 R_S^2 - 2\rho_0\rho_I R_S R_T + R_T^2)/(n^2(1-\rho_0^2))) - c_p n - c_w(n+\Delta) + \mathcal{O}(1/n^3)$. We discuss this result in more detail in the next subsection.

3.2. Absolute Benefits of Type C Trials

Several of the results in Lemma 3 are unsurprising. Like Type A trials, Type C trials require some true outcome observations; as a result, larger values of the delay Δ and the true outcome ratio R_T result in more costly trials. Similarly, a larger market a implies more profits to be reaped from a successful trial.

However, we find all the non-monotone relationships identified in Lemma 3 to be unexpected and of great interest. Intuitively, a stronger study-level correlation $|\rho_0|$ and a stronger individual-level correlation $|\rho_I|$ would lead to a more informative surrogate, and a lower surrogate ratio R_S would lead to improved statistical power for measuring μ_S — thus, *a priori* one might assume that the HTA objective would be monotonically increasing in $|\rho_0|$ and $|\rho_I|$, and monotonically decreasing in R_S. Instead, we see that the sign of these three quantities' relationships with the HTA objective are controlled by the signs of expressions $\rho_I R_S - \rho_0 R_T$ and $R_T - \rho_0 \rho_I R_S$. For fixed R_S and R_T, this in turn partitions the (ρ_0, ρ_I) space into regions displaying different relationships between the exogenous parameters and the HTA objective of a Type C trial. Limiting our attention to the case of $\rho_0, \rho_I \geq 0$ without loss of generality, Figure 2 shows that we obtain either three regions of interest (in the case when $R_T < R_S$) or only two (when $R_T \geq R_S$).

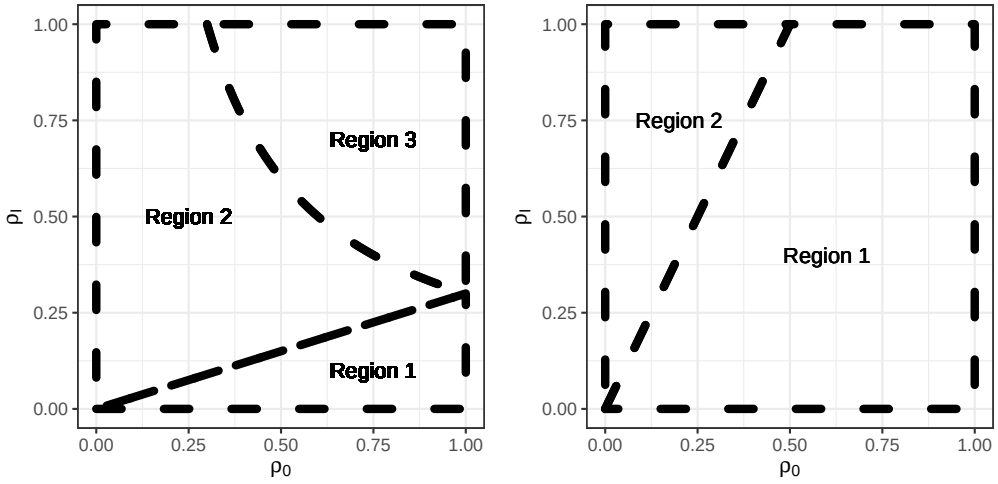


Figure 2 Regions exhibiting different relationships between the exogenous outcome parameters ρ_0, ρ_I, and R_S and the HTA objective of a Type C trial. Three regions emerge when $R_T < R_S$ (left), while only two emerge when $R_T \geq R_S$ (right).

In **Region 1**, $\rho_I R_S \leq \rho_0 R_T$ (and therefore $R_T - \rho_0 \rho_I R_S \geq 0$), meaning the benefit of a type C trial is monotone decreasing in ρ_I and R_S and monotone increasing in ρ_0. In **Region 2**, $\rho_I R_S \geq \rho_0 R_T$ and $R_T - \rho_0 \rho_I R_S \geq 0$, meaning the benefit of a type C trial is monotone increasing in ρ_I and R_S and monotone decreasing in ρ_0. Finally, in **Region 3**, $\rho_I R_S \geq \rho_0 R_T$ and $R_T - \rho_0 \rho_I R_S \leq 0$, meaning

the benefit of a type C trial is monotone increasing in ρ_0, ρ_I, and R_S. Region 3 only exists when $R_T < R_S$ (the true outcome requires fewer samples to measure accurately than the surrogate).

To build intuition about the behavior exhibited in the three identified regions, it is helpful to think of several simple cases. First, regardless of R_S, the HTA objective is strictly increasing in individual-level correlation ρ_I when the study-level correlation $\rho_0 = 0$; similarly, the HTA objective of a Type C trial is strictly increasing in study-level correlation ρ_0 when the individual-level correlation $\rho_I = 0$. We expect these relationships because the surrogate provides no information at all when $\rho_0 = \rho_I = 0$, so increasing either correlation from this point improves the HTA objective by strictly increasing the informative value of the surrogate.

On the other hand, when both correlations are nonzero, it is possible that increasing the magnitude of one correlation can *lower* the objective. The key driver behind this result is that less predictive surrogates introduce an element of randomness, thereby increasing our effective sample size. For instance, consider the case where the study-level correlation $\rho_0 = 1$, *i.e.*, the surrogate effect size is perfectly predictive of the true outcome effect size. Now, a high individual-level correlation $\rho_I = 1$ implies that the surrogate and true outcomes for each patient are perfectly correlated as well, making a patient's true outcome observation uninformative. On the other hand, a low individual-level correlation $\rho_I = 0$ allows us to observe two independent outcome observations per patient, effectively doubling our sample size for the same number of patients n. In other words, increasing $|\rho_I|$ can reduce benefit by removing a source of statistical independence when $|\rho_0|$ is high. A similar argument holds vice-versa, *i.e.*, when $\rho_I = 1$, increasing $|\rho_0|$ can reduce benefit. In particular, there are always local optima at the points $(\rho_0 = 1, \rho_I = 0)$ and $(\rho_0 = 0, \rho_I = 1)$, *i.e.*, in Regions 1 and 2.

Thus, the traditional wisdom that more predictive surrogates are always more valuable is not true from a statistical viewpoint — Type C trials can improve power with surrogate outcomes that are moderately rather than highly predictive of true outcomes. In contrast, moderately predictive surrogates are currently not used to inform trial decisions, while highly predictive surrogates are used widely in Type B trials.

However, in the case $R_S > R_T$, we additionally see Region 3, where simultaneously increasing both ρ_0 and ρ_I is useful; thus, there is an additional local optimum at $(\rho_0 = 1, \rho_I = 1)$. In this regime, the benefit of additional statistical independence is smaller than the benefit of increased individual-level correlation of the surrogate outcomes (see Remark 1 in Section 2.1).

3.3. Relative Benefits of Type C Trials

We now use the intermediate results from Section 3.1 to characterize parameter regimes where Type C trials are particularly advantageous.

THEOREM 2 **(When to Bother?)**. *The HTA benefit of Type C trials relative to the best of Type A and B trials is given by* $b_C^{trial} - \max\{b_A^{trial}, b_B^{trial}\}$. *For large n, Type A will be the best comparator when* $c_w\Delta < a(1-\rho_0^2)/\sqrt{2\pi}$, *and the benefit of Type C over Type A is increasing in a, is unaffected by* Δ, *and is non-monotone in* $|\rho_0|$, $|\rho_I|$, R_S, *and* R_T. *Meanwhile, Type B will be the best comparator when* $c_w\Delta > a(1-\rho_0^2)/\sqrt{2\pi}$, *and the benefit of Type C over type B is increasing in* R_S *and a, is decreasing in* $|\rho_0|$, R_T, *and* Δ, *and is non-monotone in* $|\rho_I|$.

The proof of Theorem 2 is provided in Appendix D; we omit the exact expressions since they follow directly from Lemmas 1–3.

Theorem 2 and Lemmas 1–3 allow us to reason about the types of diseases and surrogates for which our proposed design is particularly advantageous. As established in Lemma 3 and Section 3.2, we can identify three regions within the (ρ_0, ρ_I) space where a Type C trial might have particularly high benefit — low ρ_I and moderate/high ρ_0 (the most promising part of Region 1), low ρ_0 and high ρ_I (the most promising part of Region 2), and high ρ_0 and ρ_I (the most promising part of Region 3). However, from Theorem 2, we know that the Type B trial will likely be the comparator when ρ_0 is large, and the incremental benefit of Type C over Type B decreases as ρ_0 gets larger. As a result, we might conclude that two regions of the (ρ_0, ρ_I) space that are most likely to yield large benefits of using Type C trials versus the best comparator are: the region with low ρ_I and moderate ρ_0, and the region with low ρ_0 and high ρ_I.

To test this hypothesis, in Figure 3, we numerically simulate the relative improvement in the full HTA objective by Type C trials,

$$\frac{b_C - \max\{b_A, b_B\}}{\max\{b_A, b_B\}},$$

as a function of ρ_0 and ρ_I, using a grid of exogenous parameter values (see Section 3.4 for details). Importantly, unlike our comparative statics above, these results do not make any approximations, include an intermediate analysis, and follow the optimal designs prescribed in Theorem 1. Indeed, the two identified regions are the ones that provide the most relative benefit for a Type C trial compared to the best-performing single-endpoint design (see Figure 3).

We achieve the most relative benefit when ρ_I is large and ρ_0 is low. When ρ_0 is large, we achieve little improvement relative to surrogate-only designs. However, when ρ_0 is moderate, we observe that the improvement is non-monotone in ρ_I — as predicted by Lemma 3, a small value of ρ_I provides benefits through increased statistical independence while a large value of ρ_I reduces the effective variance of true outcome observations; however, in this regime, moderate values of ρ_I result in less benefit.

Identifying unexpected regions where Type C trials are promising has the potential to greatly expand the use of surrogates in clinical trials, to the overall benefit of population health. Figure 4

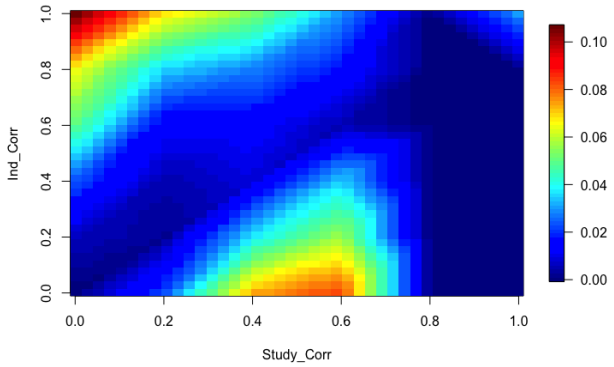


Figure 3 **Relative improvement in HTA objective by Type C trials relative to existing trials as a function of the study-level correlation ρ_0 (x-axis) and the individual-level correlation ρ_I (y-axis).**

shows a scatterplot of ρ_0 and ρ_I for different true and surrogate outcome pairs that we collected from 65 meta-analyses in the medical literature. Historically only pairs with very high ρ_0 values would be considered viable surrogates for use as the primary endpoint, but our proposed design allows the trial designer to tap into a richer variety of surrogates, particularly deriving benefits for surrogates within the two regions we identified. Figure 4 shows that there indeed exist diseases and surrogates that lie in these regions, *i.e.*, these trials can significantly benefit from incorporating surrogates, but currently do not utilize them. In Section 4, we illustrate the benefits of our design for metastatic breast cancer, where both correlations are moderate.

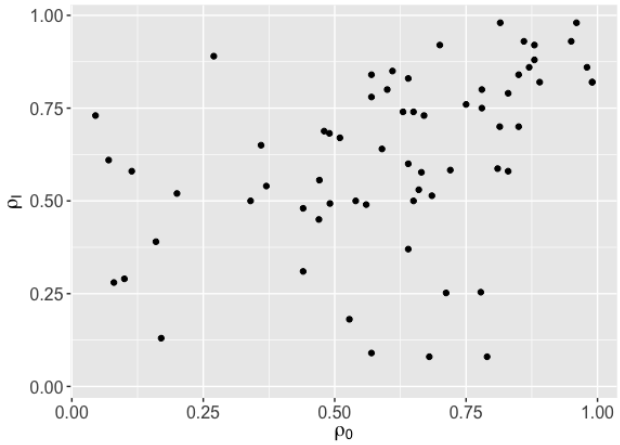


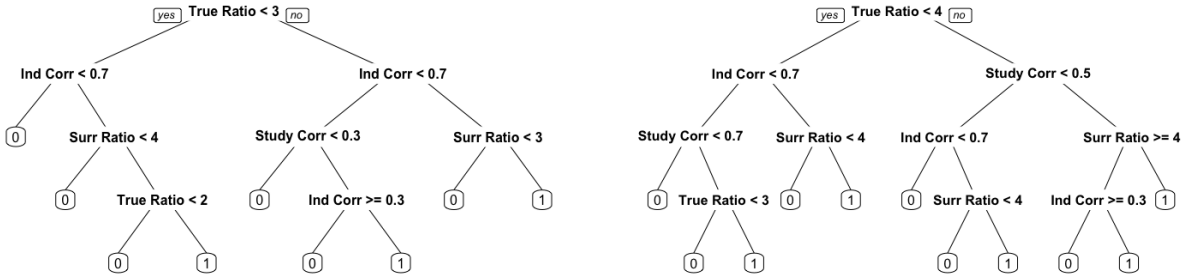
Figure 4 **Scatterplot of study-level correlation ρ_0 (x-axis) and individual-level correlation ρ_I (y-axis) for different true and surrogate outcome pairs collected from 65 meta-analyses in the medical literature.**

3.4. Interpretable Machine Learning Model

We now complement our analytical insights with an interpretable decision tree model that identifies parameter regimes where our proposed Type C trial design offers significantly higher HTA benefit relative to existing Type A and B trial designs.

In order to train a decision tree, we generate a dataset by numerically simulating the benefits of all three trial types for a large set of parameter instances. Again, unlike our comparative statics, these results do not make any approximations, include an intermediate analysis, and follow the optimal designs prescribed in Theorem 1. Once again, we maintain an unbiased expectation of the benefit of the surrogate and true outcome effect sizes in our prior ($\boldsymbol{\mu}_0 = 0$), and normalize the variances of the true and surrogate outcome effect sizes ($\sigma_{0S} = \sigma_{0T} = 1$); we also take $b = 0$, implying that any positive true outcome effect size is sufficient for an approval decision. We choose $c_p = 500,000$ and $c_w = 250,000$. We consider two market sizes: $a = 5 \times 10^8$ (small market) and $a = 5 \times 10^9$ (large market). The former captures settings where the trial costs are significant compared to the benefit of an effective drug, while the latter captures the opposite. We then run all three trial designs, sweeping over all combinations of the following parameters: $\Delta \in \{25, 50, 75, 100, 125\}$, $R_S \in \{1, 2, 3, 4, 5\}$, $R_T \in \{1, 2, 3, 4, 5\}$, $\rho_I \in \{0, .2, .4, .6, .8, .99\}$, and $\rho_0 \in \{0, .2, .4, .6, .8, .99\}$. For simplicity, we choose our outcome for the decision trees as an indicator variable for whether our trial produces relatively large benefits for this particular market size, *i.e.*, the benefit is in the top quartile[5] of simulated instances.

Figure 5 shows the resulting decision trees for both small and large markets.



(a) Small Market (b) Large Market

Figure 5 Decision trees identifying disease and surrogate parameter regimes where Type C trials offer the most HTA benefit relative to the benefit of existing trial designs.

[5] This choice is arbitrary, and we find that our results are robust to the cutoff choice.

In both small and large markets, our results match our earlier analytical findings. In line with Theorem 2, Type C trials produce sizeable benefits relative to existing trial designs when R_T is moderate and ρ_I is high. Lower ρ_I can be compensated by a moderate ρ_0. As noted earlier, the relative benefit of Type C trials is low when the surrogate has a high ρ_0 or a low R_S, since Type B trials are already very efficient in these settings; similarly, the relative benefit of Type C trials is low when the disease has a low R_T, since Type A trials are already very efficient in this setting.

4. Simulation Based on Large-Scale Clinical Trial Database

While the clinical trial models in Section 2 (for continuous outcomes) and Appendix C (for time-to-event outcomes) were built to resemble real-world trials, they still make some assumptions that may be violated in practice. As described in Section 2.1 and Appendix E, these models assume study effect sizes are drawn from a bivariate normal prior, while the time-to-event designs further assume exponentially distributed event times. However, these assumptions may not always hold. It is therefore important to confirm that these designs still provide benefit even when their assumptions are not exactly met. To this end, we provide a detailed evaluation of the performance of our designs for the case of metastatic breast cancer (MBC) drugs. We chose MBC because it represents a disease with enormous global burden — it is the most deadly cancer among women globally (Bray et al. 2018). MBC drugs have also been extensively studied in the medical literature, allowing us to calibrate and evaluate our trial designs on a wealth of past clinical trial data.

The true outcome of interest here is generally overall survival (OS; survival time from study enrollment), and a popular surrogate outcome is progression-free survival (PFS; time from study entry to disease progression). Importantly, both outcomes are time-to-event rather than continuous; thus, we use the time-to-event trial designs described in Appendix C rather than the continuous event designs from Section 2.

In the remainder of this section, we use a large-scale database of MBC clinical trial results and perform additional literature review to estimate the parameters needed for our time-to-event trial design. Then we use individual patient outcomes from a subset of trials to simulate how our proposed Type C trial design would have performed compared to Type A (true outcomes only) and Type B (surrogate outcomes only) designs; simulation of individual patient outcomes required a significant data collection effort from the MBC literature (see Section 4.2).

4.1. Clinical Trial Design

As described in Appendix C, the effect size in the time-to-event setting is the log hazard ratio. Under exponentially distributed hazard times and the proportional hazards assumption, a trial with true log hazard ratio $\boldsymbol{\mu} = [\mu_S, \mu_T]'$ will have exponentially distributed surrogate and true outcome times in the treatment arm with rates $e^{\mu_S}\lambda_S$ and $e^{\mu_T}\lambda_T$ respectively, where λ_S and λ_T

are the corresponding rates in the control arm. Effective drugs have a negative value of μ_T; thus, a (benefit of a unit increase in μ_T) is also negative. Another key difference is that we define the intermediate and final analyses to occur after a fixed number of events (*i.e.*, when the posterior variance reaches a target level of v_1 for the intermediate analysis and v_2 for the final analysis) rather than at fixed times t_1 and t_2 as we did for continuous outcomes; this is a common trial design choice for time-to-event outcomes.

The time-to-event trial design takes as input eight exogenous outcome parameters (μ_{0S}, μ_{0T}, σ^2_{0S}, σ^2_{0T}, ρ_0, ρ_I, λ_S, and λ_T) and five exogenous economic parameters (a, b, c_p, c_w, and n^{max}), and further computes three endogenous trial design parameters (n, v_1, and v_2); these parameters are summarized in Table C.1 in Appendix C. We now describe how each parameter was chosen to obtain the final trial design, and also introduce a new parameter λ_E, which captures the patient enrollment rate in the clinical trial.

We chose our individual-level correlation ρ_I, and our control group surrogate and true outcome arrival rates λ_S and λ_T based on the individual patient data meta-analysis of Burzykowski et al. (2008). The authors report the rank correlation coefficient between PFS and OS to be 0.688 with 95% confidence interval $[0.686, 0.690]$, so we set $\rho_I = 0.688$. They also report a median PFS and OS of 7.05 months and 21.55 months respectively across all studies. Since an exponential distribution with rate λ has median $(\ln 2)/\lambda$, we infer surrogate and true outcome exponential distribution event rates to be $\lambda_S = (\ln 2)/7.05 \approx 0.098$ and $\lambda_T = (\ln 2)/21.55 \approx 0.032$. All remaining outcome parameters are estimated using a repository of 1,865 studies of MBC drug therapies collected by Silberholz et al. (2019), which is publicly available at `http://www.cancertrials.info`.

We perform our trial simulations on a subset of 71 studies from this repository — two-arm RCTs (one treatment and one control) that provide surrogate and true outcome effect size estimates. We manually extract Kaplan-Meier curves for both outcomes for each of these studies, providing us with the detailed patient time-to-event data needed for simulating our trial designs (see Section 4.2). Table 4.1 summarizes key features of the patients and their outcomes in these 71 RCTs.

For study $i \in \{1, \ldots, 71\}$, let $\hat{e}_{iT}$ and σ^2_{iT} denote the effect size estimate and its sampling variance for the true outcome (log OS hazard ratio), and let $\hat{e}_{iS}$ and σ^2_{iS} denote the same quantities for the surrogate outcome (log PFS hazard ratio). Assuming individual-level effect size correlation ρ_I, the sampling variance for $\hat{\mathbf{e}}_i := [\hat{e}_{iS}, \hat{e}_{iT}]'$ is

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2_{iS} & \rho_I \sigma_{iS} \sigma_{iT} \\ \rho_I \sigma_{iS} \sigma_{iT} & \sigma^2_{iT} \end{bmatrix}.$$

Assuming each study's effect size estimate is independently drawn from the prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, the log likelihood of a given $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ is

$$LL(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \propto -\sum_{i=1}^{71} \log\det(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_i) + (\hat{\mathbf{e}}_i - \boldsymbol{\mu}_0)'(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_i)^{-1}(\hat{\mathbf{e}}_i - \boldsymbol{\mu}_0).$$

	Average	Range
Publication year	2005	[1984, 2017]
Number of patients	292	[51, 1198]
Proportion female	1.00	[0.99, 1.00]
Median age	56	[48, 72]
Mean ECOG performance status	0.59	[0.27, 1.05]
Proportion with visceral disease	0.63	[0.18, 0.87]
Median OS (months)	21.3	[9.0, 45.6]
Log(OS hazard ratio)	-0.06	[-0.80, 0.59]
Median PFS (months)	7.2	[2.2, 17.9]
Log(PFS hazard ratio)	-0.14	[-0.85, 0.63]

Table 4.1 Aggregate patient characteristics of the 71 RCTs of metastatic breast cancer drug therapies used for the simulation-based evaluation. Eastern Cooperative Oncology Group (ECOG) performance status measures patient level of functioning on a scale from 0 (fully active) to 5 (dead). Visceral disease represents particularly severe MBC that has spread to internal organs such as the liver and lungs.

Maximizing the log likelihood with the Nelder-Mead simplex yields parameter values $\mu_{0S} = -0.117$, $\mu_{0T} = -0.058$, $\sigma_{0S} = 0.251$, $\sigma_{0T} = 0.120$, and $\rho_0 = 0.646$. Figure 6 plots the study effect size estimates along with the fitted 95% confidence ellipse for study true effect sizes $\boldsymbol{\mu}$.[6] While a bivariate normal distribution may visually appear to be a reasonable assumption, the effect sizes reject the null hypothesis of multivariate normality ($p = 0.037$) under the Henze-Zirkler test. This highlights that MBC drug therapies are well suited for the purposes of this section — to test the performance of our proposed designs even when the assumptions are not exactly met.

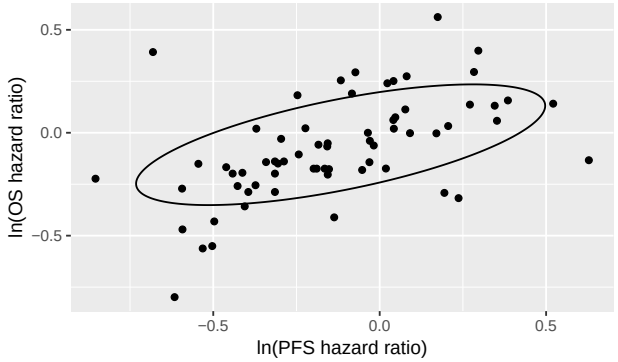


Figure 6 Effect size estimates of the 71 MBC RCTs from Silberholz et al. (2019) used to parameterize the prior, along with the 95% confidence ellipse of the true study effect sizes from the fitted prior.

We estimate the MBC economic parameters from the perspective of a social planner attempting to extend length of life among MBC patients in the USA; we discuss how other decision-makers might make different choices in selecting these parameters in Section 5. From published estimates that per-patient costs average $5.2 million in Phase III trials (Sertkaya et al. 2016) and an average

[6] Note that we would not expect 95% of study effect size estimates to fall into this ellipse because it is a 95% confidence ellipse for $\boldsymbol{\mu}$, the true effect size, while study effect size estimates also have sampling errors.

patient enrollment of 370 in Phase III MBC trials (Silberholz et al. 2019), we arrive at an estimated $c_p = \$14,000$ for the incremental cost of enrolling one additional patient in a clinical trial. Next, the parameter $-a$ captures the rate at which population health improves as the log hazard ratio improves (decreases). The true outcome exponential distribution rate $\lambda_T = 0.032$ yields expected survival of 31 months, so a new drug with log hazard ratio μ_T yields an expected survival $31e^{-\mu_T} \approx 31 - 31\mu_T$ for small μ_T. Combining the facts that (i) the United States MBC incidence rate is roughly 40,000 new cases per year (National Cancer Institute 2019), (ii) a new drug has an estimated market share of 10% based on the number of MBC drug classes on the market (Corcoran et al. 2019), (iii) the estimated time to obsolescence is 30 years since roughly a third of FDA-approved MBC drugs were approved in the last decade (FDA 2019), and (iv) an assumed willingness-to-pay of \$50,000 per year for an extra year of life, we obtain an estimate of $-a = \$16$ billion. The parameter $-b$ captures the threshold of population benefit needed to justify using a new drug; it would vary significantly based on the price of the new drug and on the price of the new drug's competitors. We set $-b = \$800$ million by assumption. The parameter c_w captures the expected incremental population health cost of delaying an approval decision by a month. There is no cost if the drug is ineffective ($\mu_T \geq 0$) and otherwise the cost scales with μ_T. From our outcome parameter estimates, $\mu_T \sim \mathcal{N}(-0.058, 0.120^2)$ so $\mathbb{E}[\min(\mu_T, 0)] = -0.082$. The same logic that we used for estimating a yields an estimate of $c_w = \$3.5$ million. Finally, we set $n^{max} = 688$, which is the 90th percentile of trial enrollments observed in Phase III MBC trials (Silberholz et al. 2019).

Although our trial design in Appendix C assumes one patient is enrolled every time period, all other time-based parameters estimated thus far have been measured in months. To convert from months to the inter-enrollment period, we estimate λ_E, the enrollment rate per month for Phase III MBC trials. We manually collected enrollment periods and patient counts for 59 Phase III trials from the aforementioned MBC clinical trial repository, yielding an estimate of $\lambda_E = 8$.

Given the values of the exogenous parameters, we then found the optimal trial design for each trial type $t \in \{A, B, C\}$. Following the approach detailed in Appendix C, we ran multi-start Nelder-Mead simplex from randomly selected initial values to identify high-quality values for n_t (the target enrollment), v_{1t} (the posterior variance at which the intermediate analysis occurs), and v_{2t} (the posterior variance at which the final analysis occurs). Once these values were selected, we again followed Appendix C to obtain $(\underline{\mu_t}, \overline{\mu_t})$, the range of intermediate analysis posterior mean values at which the trial will be continued.

Table 4.2 summarizes the full set of parameter values used in the MBC trial design, along with the sources of the parameters.

Param.	Value	Source
μ_{0S}	-0.117	Estimated from data in Silberholz et al. (2019)
μ_{0T}	-0.058	"
σ_{0S}	0.251	"
σ_{0T}	0.120	"
ρ_0	0.646	"
ρ_I	0.688	Burzykowski et al. (2008)
λ_S	0.098	"
λ_T	0.032	"
a	-1.6×10^{10}	National Cancer Institute (2019), Corcoran et al. (2019), and FDA (2019)
c_w	3.5×10^6	"
b	-8.0×10^8	Assumption
c_p	1.4×10^4	Sertkaya et al. (2016) and Silberholz et al. (2019)
n^{max}	688	Silberholz et al. (2019)
λ_E	8	Silberholz et al. (2019)
n_A	688	Optimized based on the other parameters (see Appendix C)
v_{1A}	0.0093	"
v_{2A}	0.0041	"
$\underline{\mu_A}$	-0.1185	"
$\overline{\mu_A}$	0.0170	"
n_B	378	"
v_{1B}	0.0110	"
v_{2B}	0.0099	"
$\underline{\mu_B}$	-0.0799	"
$\overline{\mu_B}$	-0.0197	"
n_C	688	"
v_{1C}	0.0086	"
v_{2C}	0.0040	"
$\underline{\mu_C}$	-0.1023	"
$\overline{\mu_C}$	0.0016	"

Table 4.2 Summary of all chosen parameters for the MBC trial simulation.

4.2. Simulation Model

To simulate clinical trials, we need to be able to construct random sets of control- and treatment-group patients, labeled by their surrogate and true outcome event times. To this end, we first obtained empirical cdfs of both event times from Kaplan-Meier curves published in the 71 identified RCTs. Kaplan-Meier curves display estimates of the proportion of patients who are event-free after t time has elapsed from study entry, for varying values of t. We manually extracted Kaplan-Meier curves in digital form for the treatment and control group for both OS and PFS from clinical trial reports using the Java-based plotdigitizer utility. While the Kaplan-Meier curves yield the marginal distributions of OS and PFS outcomes in the control and treatment group, we still need joint distributions of OS and PFS outcomes to simulate random control and treatment groups. For a given study i, we approximate the joint distribution using a Gaussian copula, which controls the strength of dependence between OS and PFS with a single parameter $\rho \in [-1, 1]$. For the control (experiment) arm in each study i, we define $\rho_{ctl,i}$ ($\rho_{exp,i}$) to be the value of ρ that yields a correlation

between OS and PFS event times that is closest to $\rho_I = 0.688$. We numerically compute $\rho_{ctl,i}$ and $\rho_{exp,i}$ values for each study i using a grid search on the interval $[-1, 1]$.

We simulated each trial type t for each study i for 1,000 replicates. For each replicate r, we assume patients are enrolled in evenly-spaced intervals, *i.e.*, joining the study at times $0,\ 1/\lambda_E,\ \cdots,\ (n_t - 1)/\lambda_E$, and then getting allocated randomly with 50/50 probability to an arm. Each patient allocated to the experiment (control) arm has their surrogate and true outcome event time drawn i.i.d. from the experiment (control) group empirical survival distributions for study i, joined by a Gaussian copula with parameter $\rho_{exp,i}$ ($\rho_{ctl,i}$). Based on these event times, we compute t_{1tr} (t_{2tr}) and n_{1tr} (n_{2tr}), the time and number of patients enrolled when the posterior variance for the Bayesian inference first reaches v_{1t} (v_{2t}), indicating it is time for the intermediate (final) analysis. We compute effect sizes $\hat{\mathbf{e}}_{ij}$ using the logrank test to estimate the log hazard ratio, performing the Bayesian updates given in Appendix C to obtain the intermediate (final) analysis posterior mean $\hat{\mu}_{1tT}$ ($\hat{\mu}_{2tT}$). If $\hat{\mu}_{1tT} > \overline{\mu_t}$, we stop early ($e_{tr} = 1$) and reject the drug ($a_{tr} = 0$). If $\hat{\mu}_{1tT} < \underline{\mu_t}$, we stop early ($e_{tr} = 1$) and accept the drug ($a_{tr} = 1$). Otherwise, we continue to the final analysis ($e_{tr} = 0$), accepting the drug if $\hat{\mu}_{2tT} < -b/a$ ($a_{tr} = 1$) and rejecting it otherwise ($a_{tr} = 0$).

For each trial type t, we then average over all trials and replicates to estimate the average cost of patient enrollment C_t^{pe}, the average cost of waiting C_t^w, and the average HTA objective of the trial decision B_t (here we assume that the actual true outcome effect size equals $\hat{e}_{iT}$, the published effect size in study i). We then obtain:

$$C_t^{pe} = \frac{1}{71000} \sum_{i=1}^{71} \sum_{r=1}^{1000} c_p(n_{1,t,r} e_{t,r} + n_{2,t,r}(1 - e_{t,r}))$$

$$C_t^{w} = \frac{1}{71000} \sum_{i=1}^{71} \sum_{r=1}^{1000} c_w(t_{1,t,r} e_{t,r} + t_{2,t,r}(1 - e_{t,r}))$$

$$B_t = \frac{1}{71000} \sum_{i=1}^{71} \sum_{r=1}^{1000} (a\hat{e}_{iT} + b) a_{t,r}$$

Detailed pseudocode of the simulation model for a study i and trial type $t \in \{A, B, C\}$ is provided in Algorithm 2 in Appendix F.

4.3. Simulation Results

On average, we found that running a type A trial yielded a HTA objective of \$1.01 billion, running a type B trial yielded a HTA objective of \$1.06 billion, and running a type C trial yielded a HTA objective of \$1.11 billion. Thus, a type C trial resulted in a \$55 million (5.2%) higher average trial benefit than the best of existing trial types. Table 4.3 breaks down the HTA objective results by patient enrollment costs, waiting costs, and population benefits of trial decisions for each trial type.

	Trial Type A: True Outcome Only	Trial Type B: Surrogate Outcome Only	Trial Type C: Combined Outcomes
Avg. Cost of Patient Enrollment	7	3	6
Avg. Cost of Waiting	302	84	262
Avg. Benefit of Trial Decision	1316	1141	1378
Avg. Overall Benefit	1006	1055	1110

Table 4.3 Breakdown of different components of the HTA objective from our simulation results (in millions).

All of the advantage of Type C trials compared to Type A trials arose from reduced waiting times (we save an average of \$40 million, or 13% of waiting costs), and better approve/reject decisions (we gain an average of \$62 million, or 5% of population benefits). This matches our expectation, since incorporating surrogate outcomes should yield faster trial decisions and improve statistical power compared to Type A trials that rely only on true outcomes. We note that despite potential prior misspecification, Type C trials achieve faster *and* more accurate trial decisions.

Most of the advantage of Type C trials compared to Type B trials arose from making better approve/reject decisions (we gain an average of \$237 million, or 21% of population benefits). In contrast, Type B trials actually significantly reduced waiting times relative to Type C trials (average savings of \$178 million, or 68% of waiting costs). Again, this is to be expected since one can estimate the surrogate effect size μ_S much faster than the true outcome effect μ_T. However, as we see here, a key concern with relying on surrogate outcomes alone is that they may result in poor trial decisions (since the surrogate outcome is not perfectly predictive of the true outcome).

We note that the average cost of patient enrollment did not play a large role for any of the three trial types. This is likely because the cost of patient enrollment c_p is much smaller relative to the cost of waiting c_w and the benefit of a unit improvement in effect size a.

5. Discussion and Conclusions

In this work, we proposed and studied the properties of clinical trial designs that take into account both surrogate and true outcome information when making design and stopping decisions (Section 2), identifying situations when these designs are particularly advantageous compared to current approaches (Section 3), and establishing that these designs could have a significant benefit in actual clinical trials even when the assumptions of the model are not exactly met (Section 4). In this final section, we discuss remaining barriers to the implementation of the proposed trial designs, as well as limitations of the current study and directions for future work.

A key barrier to the widespread implementation of the designs presented in this work is their acceptance by regulators such as the FDA in the United States. The FDA has a vested interest in

the implementation of modern clinical trial designs in order to (i) make drug development more efficient and less costly in order to improve patient health and increase competition in the drug market, and to (ii) increase the amount of information we can learn about a new drug's benefits through the use of multiple arms, adaptive randomization, personalization, etc. (FDA 2018c). To support these efforts, the FDA began a Complex Innovative Trial Design pilot meeting program in 2018 to facilitate the use of complex adaptive, Bayesian, and other novel trial designs (FDA 2018a). Bayesian trial designs are also increasingly popular in practice (Lee and Chu 2012); the BATTLE trial was a particularly successful instance, which employed a Bayesian adaptive design for personalizing treatment allocation using patient biomarker profiles (Kim et al. 2011).

As a result, there may be cause for optimism towards piloting designs inspired by this work. First, our proposed design has the potential to accelerate the approval process for drugs, which matches FDA's key objectives (FDA 2018c,a). Furthermore, the FDA already has formal mechanisms (*e.g.*, the Accelerated Approval Program, FDA 2016) to recognize surrogates when their benefits (speed or ease of measurement) outweigh their costs (not exactly measuring the true outcome of interest). Designs that combine surrogate and true outcomes represent a natural extension of these currently approved designs. Lastly, just as with surrogate-only designs, the FDA could require long-term post-approval follow-up to show whether the drug actually provides the anticipated clinical benefit for the true outcome of interest (FDA 2016). If it later becomes clear that the drug does not improve the true outcome, the FDA has existing regulatory procedures for removing the drug from the market.

In general, Bayesian clinical trial designs have several advantages, including the ability to directly maximize the designer's utility function (Chick et al. 2017), allow for more frequent monitoring and interim decision-making (Kim et al. 2011), and account for uncertainty and prior information systematically (Lee and Chu 2012). However, a significant disadvantage is that prior misspecification (either due to inadequate historical data or manipulation by a study designer) may lead to unwarranted conclusions. The Bayesian community has proposed several solutions to this challenge, including pre-specifying the prior, performing sensitivity analyses, using objective priors, and modeling the uncertainty in the distribution parameters through a hierarchical model (Lee and Chu 2012). Such approaches can be applied to our proposed designs as well to ensure that our trial decisions are robust.

Along the lines of the concerns about prior misspecification, one particularly significant concern in defining the prior comes from publication bias, *i.e.*, some study authors choose not to publish the results of their studies (typically smaller studies with negative results). This phenomenon has been documented extensively in the context of meta-analyses (Rothstein et al. 2005) and specifically in the medical literature (Easterbrook et al. 1991), and could reasonably be expected to cause

overly optimistic priors for Bayesian clinical trials that are parameterized using the published medical literature, as they would be expected to introduce positive bias in the prior mean $\boldsymbol{\mu}_0$. While publication bias is a concern in any medical literature review, several steps can be taken to limit the impact of this phenomenon on the designs proposed in this study. First, a number of approaches exist to identify and eliminate publication bias. Funnel plots (Light and Pillemer 1984) are a popular visual tool for detecting publication bias, plotting effect size against study precision in a scatterplot; asymmetric effect sizes for small studies can readily be identified from these plots. Going a step further, approaches like the trim and fill method (Duval and Tweedie 2000) can impute the missing smaller studies. However, publication bias may not be a significant concern in our setting, since such a bias would likely cause researchers to *underestimate* the study-level correlation $|\rho_0|$. In particular, consider a bivariate normally distributed $\boldsymbol{\mu} = [\mu_S, \mu_T]'$ with a positive correlation $\rho_0 > 0$. Under publication bias, we would observe a truncated bivariate normal distribution $\boldsymbol{\mu}|\mu_T \geq t$ for some constant t; using this data, we would estimate a correlation that is strictly smaller in magnitude than ρ_0 (see, *e.g.*, Rao et al. 1968). This implies that employing a prior with $\boldsymbol{\mu}_0 = 0$ and estimating parameters $\{\rho_0, \sigma_{0S}^2, \sigma_{0T}^2\}$ from data that is collected in the presence of publication bias would actually lead to a *conservative* trial design due to an underestimated value of $|\rho_0|$.

Our proposed clinical trial model and resulting designs could be extended in a number of ways. We focused on continuous-valued outcomes with known sampling variance, with a surrogate outcome that is observed immediately upon patient enrollment and a true outcome that is observed after a fixed delay; Appendix C further considered time-to-event outcomes with exponentially distributed wait times to the surrogate and true outcomes. Our trial designs can be straightforwardly extended to many other types of outcomes, such as continuous outcomes with unknown variance, or binary or categorical outcomes. The effect sizes considered could be correspondingly expanded, *e.g.*, to odds ratios or risk ratios for binary outcomes. Also, we maximized a HTA objective that balances trial costs and population benefit, but one could consider other objectives. An objective that may be of particular interest to pharmaceutical company decision makers would be cost minimization with a target power $1 - \beta$ to detect a specified effect size while committing no more than α rate of Type I errors. Our approach can also be numerically extended to allow for multiple intermediate analyses or continuous monitoring.

Finally, we note that surrogates, more broadly interpreted, have found wide use outside of clinical trials as well. We define a surrogate outcome to be any outcome that is closely related to the true outcome of interest, but is much faster or easier to measure. For example, in revenue management, e-commerce platforms use abundant customer click data (surrogate outcome) to make product recommendations rather than the relatively sparse customer purchase data (true outcome

of interest); alternatively, Medicare measures hospital quality of care through patient readmission rates (surrogate outcome) rather than patient mortality rates (true outcome of interest). Such examples are abundant throughout revenue management and healthcare. In these settings, Bastani (2018) demonstrates that combining surrogate and true outcomes can improve predictive power in supervised learning problems. Analogously, our proposed trial designs can be used for adaptive A/B testing, performing efficient inference by combining surrogate and true outcomes.

Acknowledgments

The authors gratefully acknowledge Sahil Gupta, Munashe Mandizwidza and Jacob Newsham for their significant data collection efforts, as well as various seminar participants for helpful feedback.

References

Ahuja, Vishal, John R Birge. 2016. Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European Journal of Operational Research* **248**(2) 619–633.

Bastani, Hamsa. 2018. Predicting with proxies. *arXiv preprint arXiv:1812.11097* .

Berry, Donald A. 2004. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science* **19**(1) 175–187.

Berry, Scott M, Bradley P Carlin, J Jack Lee, Peter Muller. 2010. *Bayesian adaptive methods for clinical trials*. CRC press.

Bertsimas, Dimitris, Allison O'Hair, Stephen Relyea, John Silberholz. 2016. An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* **62**(5) 1511–1531.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, Ahmedid Jemal. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**(6) 394–424.

Brennan, Alan, Stephen E Chick, Ruth Davies. 2006. A taxonomy of model structures for economic evaluation of health technologies. *Health economics* **15**(12) 1295–1310.

Burzykowski, Tomasz, Marc Buyse. 2005. *The Evaluation of Surrogate Endpoints*. Springer.

Burzykowski, Tomasz, Marc Buyse. 2006. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* **5**(3) 173–186.

Burzykowski, Tomasz, Marc Buyse, Martine J Piccart-Gebhart, George Sledge, James Carmichael, Hans-Joachim Lück, John R Mackey, Jean-Marc Nabholtz, Robert Paridaens, Laura Biganzoli, et al. 2008. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *Journal of Clinical Oncology* .

Burzykowski, Tomasz, Geert Molenberghs, Marc Buyse. 2004. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **167**(1) 103–124.

Burzykowski, Tomasz, Geert Molenberghs, Marc Buyse, Helena Geys, Didier Renard. 2001. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(4) 405–422.

Buyse, Marc, Geert Molenberghs. 1998. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1014–1029.

Cheung, Ying Kuen, Lurdes YT Inoue, J Kyle Wathen, Peter F Thall. 2006. Continuous bayesian adaptive randomization based on event times with covariates. *Statistics in medicine* **25**(1) 55–70.

Chick, Stephen, Martin Forster, Paolo Pertile. 2017. A bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(5) 1439–1462.

Chick, Stephen E, Noah Gans, Ozge Yapar. 2018. Bayesian sequential learning for clinical trials of multiple correlated medical interventions. Available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3184758.

Corcoran, Taylor C, Fernanda Bravo, Elisa F Long. 2019. Flexible fda approval policies .

Daniels, Michael J, Michael D Hughes. 1997. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine* **16**(17) 1965–1982.

Duval, S., R. Tweedie. 2000. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**(2) 455–463.

Easterbrook, P. J., R. Gopalan, J. A. Berlin, D. R. Matthews. 1991. Publication bias in clinical research. *The Lancet* **337**(8746) 867–872.

FDA. 2016. Accelerated approval program. Online. URL `https://www.fda.gov/drugs/information-healthcare-professionals-drugs/accelerated-approval-program`.

FDA. 2017. Multiple endpoints in clinical trials: Guidance for industry. Online. URL `https://www.fda.gov/media/102657/download`.

FDA. 2018a. Complex innovative trial designs pilot program. Online. URL `https://www.fda.gov/drugs/development-resources/complex-innovative-trial-designs-pilot-program`.

FDA. 2018b. *Considerations for Discussion of a New Surrogate Endpoint(s) at a Type C PDUFA Meeting Request*. FDA. URL `https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DevelopmentResources/UCM614581.pdf`.

FDA. 2018c. Fda in brief: Fda modernizes clinical trial designs and approaches for drug development, proposing new guidance on the use of adaptive designs and master protocols. Online. URL `https://www.fda.gov/news-events/fda-brief/fda-brief-fda-modernizes-clinical-trial-designs-and-approaches-drug-development-proposing-new`.

FDA. 2019. Drugs fda. Online. URL `https://www.accessdata.fda.gov/scripts/cder/daf/`.

Fleming, Thomas R, David L DeMets. 1996. Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine* **125**(7) 605–613.

Floriani, Irene, Nicole Rotmensz, Elena Albertazzi, Valter Torri, Marisa De Rosa, Carlo Tomino, Fillipo de Braud. 2008. Approaches to interim analysis of cancer randomised clinical trials with time to event endpoints: A survey from the italian national monitoring centre for clinical trials. *Trials* **9**(1) 46.

Freedman, Laurence S, Barry I Graubard, Arthur Schatzkin. 1992. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine* **11**(2) 167–178.

Gail, Mitchell H, Ruth Pfeiffer, Hans C Van Houwelingen, Raymond J Carroll. 2000. On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**(3) 231–246.

Hampson, Lisa V, Christopher Jennison. 2013. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1) 3–54.

Han, Shu. 2005. Modeling auxiliary information in clinical trials. Ph.D. thesis, Rice University.

Henze, N, B Zirkler. 1990. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods* **19**(10) 3595–3617.

Kessler, Glenn. 2016. Are there really 10,000 diseases and just 500 'cures'? *Washington Post* URL `https://www.washingtonpost.com/news/fact-checker/wp/2016/11/17/are-there-really-10000-diseases-and-500-cures/`.

Kim, Edward S, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. 2011. The battle trial: personalizing therapy for lung cancer. *Cancer discovery* **1**(1) 44–53.

Kouvelis, Panos, Joseph Milner, Zhili Tian. 2017. Clinical trials for new drug development: Optimal investment and application. *Manufacturing & Service Operations Management* **19**(3) 437–452.

Lan, K. K. Gordon, David L. DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3) 659–663.

Lee, Elliot, Mariel S Lavieri, Michael Volk. 2018. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management* .

Lee, Jack, Caleb Chu. 2012. Bayesian clinical trials in action. *Statistics in medicine* **31**(25) 2955–2972.

Light, Richard J., David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Harvard University Press.

Mintz, Yonatan, Anil Aswani, Philip Kaminsky, Elena Flowers, Yoshimi Fukuoka. 2017. Non-stationary bandits with habituation and recovery dynamics. *arXiv preprint arXiv:1707.08423* .

National Cancer Institute. 2019. Surveillance epidemiology and end results. Online. URL `http://seer.cancer.gov`.

Negoescu, Diana M, Kostas Bimpikis, Margaret L Brandeau, Dan A Iancu. 2017. Dynamic learning of patient response types: An application to treating chronic diseases. *Management science* **64**(8) 3469–3488.

NICE, UK. 2018. Developing nice guidelines: the manual. UK National Institute for Health and Care Excellence. URL `https://www.nice.org.uk/process/pmg20/chapter/incorporating-economic-evaluation`.

NIH. 2018. National center for advancing translational sciences. Online. URL `https://ncats.nih.gov/about`.

Piantadosi, Steven. 2005. *Clinical Trials: A Methodologic Perspective*. 2nd ed. Wiley Series in Probability and Statistics, Wiley-Interscience.

Pozzi, Luca, Heinz Schmidli, David I Ohlssen. 2016. A bayesian hierarchical surrogate outcome model for multiple sclerosis. *Pharmaceutical statistics* **15**(4) 341–348.

Prasad, Vinay, Sham Mailankody. 2017. Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA internal medicine* **177**(11) 1569–1575.

Pratt, Craig M, Lemuel A Moyé. 1995. The cardiac arrhythmia suppression trial: casting suppression in a different light. *Circulation* **91**(1) 245–247.

Prentice, Ross L. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* **8**(4) 431–440.

Rao, B. Raja, Mohan L. Garg, C. C. Li. 1968. Correlation between the sample variances in a singly truncated bivariate normal distribution. *Biometrika* **55**(2) 433–436.

Renard, Didier, Helena Geys, Geert Molenberghs, Tomasz Burzykowski, Marc Buyse. 2002. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **44**(8) 921–935.

Renfro, Lindsay A, Bradley P Carlin, Daniel J Sargent. 2012. Bayesian adaptive trial design for a newly validated surrogate endpoint. *Biometrics* **68**(1) 258–267.

Rothstein, Hannah R., Alexander J. Sutton, Michael Borenstein, eds. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Wiley.

Sertkaya, Aylin, Hui-Hsing Wong, Amber Jessup, Trinidad Beleche. 2016. Key cost drivers of pharmaceutical clinical trials in the united states. *Clinical Trials* **13**(2) 117–126.

Silberholz, John, Dimitris Bertsimas, Linda Vahdat. 2019. Clinical benefit, toxicity and cost of metastatic breast cancer therapies: Systematic review and meta-analysis. *Breast Cancer Research and Treatment* .

Spiegelhalter, David J, Keith R Abrams, Jonathan P Myles. 2004. *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13. John Wiley & Sons.

Sydes, Matthew R, David J Spiegelhalter, Douglas G Altman, Abdel B Babiker, Mahesh KB Parmar, DAMOCLES Group. 2004. Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clinical trials* **1**(1) 60–79.

Tharmanathan, Puvan, Melanie Calvert, John Hampton, Nick Freemantle. 2008. The use of interim data and data monitoring committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. *BMC medical research methodology* **8**(1) 12.

Wang, S. K., A. A. Tsiatis. 1987. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**(1) 193–199.

Weintraub, William S, Thomas F Lüscher, Stuart Pocock. 2015. The perils of surrogate endpoints. *European heart journal* **36**(33) 2212–2218.

Xie, Wanling, M Regan, Marc Buyse, Susan Halabi, P Kantoff, Oliver Sartor, Howard Soule, N Clarke, Laurence Collette, J Dignam, et al. 2017. Metastasis-free survival is a strong surrogate of overall survival in localized prostate cancer. *Journal of Clinical Oncology* **35**(27) 3097–3114.

Appendix A: Bayesian Updates for Clinical Trial Model with Continuous Outcomes

A.1. Notation

For inference in a Type A trial, define $\mathcal{F}_{1,A}$ as the filtration of all true outcome information available through time t_1 and $\mathcal{F}_{2,A}$ as the filtration of all true outcome information available through time t_2. Similarly define $\mathcal{F}_{1,B}$ and $\mathcal{F}_{2,B}$ for Type B trials (filtrations of all surrogate outcome information) and $\mathcal{F}_{1,C}$ and $\mathcal{F}_{2,C}$ for Type C trials (filtrations of all surrogate and true outcome information).

Let $n_1 := \min(t_1, n)$ be the number of patients enrolled before the intermediate analysis and $n_2 := (n - t_1)^+$ be the number of patients enrolled after the intermediate analysis. Note that this implies that we observe n_1 surrogate outcomes by the intermediate analysis and an additional n_2 surrogate outcomes between the intermediate and final analysis. Let period 1 be the time between the start of the study and the intermediate analysis, period 2 be the time between the intermediate and final analyses, and period 3 be the time after the final analysis (we observe no information from period 3). Define n_{ij} to be the number of patients enrolled in period $i \in \{1,2\}$ who had their true outcome observed in period $j \in \{1,2,3\}$. Simple algebra yields

$$\begin{aligned}
n_{1,1} &= \min((t_1 - \Delta)^+, n) \\
n_{1,3} &= (n_1 - (t_2 - \Delta)^+)^+ \\
n_{1,2} &= n_1 - n_{1,1} - n_{1,3} \\
n_{2,3} &= (n - (t_2 - \Delta)^+)^+ - n_{1,3} \\
n_{2,2} &= n_2 - n_{2,3}
\end{aligned}$$

Denote $o_2 := n_{1,2} + n_{2,2}$ to be the number of true outcomes observed in period 2.

Let $\hat{\mathbf{e}}_{i,j}$ be the full effect size estimate vector among the $n_{i,j}$ patients enrolled in period i and with a true outcome observed in period j; for Bayesian updates we assume that either $n_{i,j} = 0$ (in which case we fix $\hat{\mathbf{e}}_{i,j} = 0$), or otherwise $n_{i,j}$ is sufficiently large that approximate bivariate normality is expected to hold.

Lastly, we will denote several matrices that we will use during Bayesian updates (note that $\mathbf{S} + \mathbf{U} = (4\boldsymbol{\Sigma}_I)^{-1}$):

$$\begin{aligned}
\mathbf{S} &:= \begin{bmatrix} 1/(4\sigma_{I,S}^2) & 0 \\ 0 & 0 \end{bmatrix} \\
\mathbf{T} &:= \begin{bmatrix} 0 & 0 \\ 0 & 1/(4\sigma_{I,T}^2) \end{bmatrix} \\
\mathbf{U} &:= \frac{1}{4\sigma_{I,T}^2(1-\rho_I^2)} \begin{bmatrix} \rho_I^2\sigma_{I,T}^2/\sigma_{I,S}^2 & -\rho_I\sigma_{I,T}/\sigma_{I,S} \\ -\rho_I\sigma_{I,T}/\sigma_{I,S} & 1 \end{bmatrix}
\end{aligned}$$

A.2. Bayesian Updates

For Bayesian updates, we use the fact that with prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and observation $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma})$, that $\boldsymbol{\mu}|\mathbf{y} \sim \mathcal{N}(\tilde{\boldsymbol{\Sigma}}[\mathbf{A}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0], \tilde{\boldsymbol{\Sigma}})$ for $\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$ (Bishop 2006). This result can be directly applied to obtain the posterior distribution at the intermediate analysis in a type A trial ($n_{1,1}$ patients with a true outcome observed), in a type B trial (n_1 patients with a surrogate outcome observed),

and in a type C trial ($n_{1,1}$ patients with both outcomes observed and an additional $n_{1,2}+n_{1,3}$ patients with only a surrogate outcome observed).

$$
\begin{aligned}
\boldsymbol{\mu}|\mathcal{F}_{1,A} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1,A}, \boldsymbol{\Sigma}_{1,A}) \\
\hat{\boldsymbol{\mu}}_{1,A} &:= \boldsymbol{\Sigma}_{1,A}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n_{1,1}\mathbf{T}\hat{\mathbf{e}}_{1,1}) \\
\boldsymbol{\Sigma}_{1,A} &:= (\boldsymbol{\Sigma}_0^{-1} + n_{1,1}\mathbf{T})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{1,B} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1,B}, \boldsymbol{\Sigma}_{1,B}) \\
\hat{\boldsymbol{\mu}}_{1,B} &:= \boldsymbol{\Sigma}_{1,B}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n_{1,1}\mathbf{S}\hat{\mathbf{e}}_{1,1} + n_{1,2}\mathbf{S}\hat{\mathbf{e}}_{1,2} + n_{1,3}\mathbf{S}\hat{\mathbf{e}}_{1,3}) \\
\boldsymbol{\Sigma}_{1,B} &:= (\boldsymbol{\Sigma}_0^{-1} + n_1\mathbf{S})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{1,C} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1,C}, \boldsymbol{\Sigma}_{1,C}) \\
\hat{\boldsymbol{\mu}}_{1,C} &:= \boldsymbol{\Sigma}_{1,C}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n_{1,1}(\mathbf{S}+\mathbf{U})\hat{\mathbf{e}}_{1,1} + n_{1,2}\mathbf{S}\hat{\mathbf{e}}_{1,2} + n_{1,3}\mathbf{S}\hat{\mathbf{e}}_{1,3}) \\
\boldsymbol{\Sigma}_{1,C} &:= (\Sigma_0^{-1} + n_1\mathbf{S} + n_{1,1}\mathbf{U})^{-1}
\end{aligned}
$$

A straightforward application of the same identity yields the posterior distribution at the final analysis for a type A trial (after o_2 additional true outcome observations) and a type B trial (after n_2 additional surrogate observations). For type C trials, it can be used to perform the update for the $n_{2,3}$ new patients with only a surrogate observed and the $n_{2,2}$ new patients with both outcomes observed. An additional $n_{1,2}$ patients observe a true outcome in the second period after having a surrogate observed in the first period. From the conditional distribution of a multivariate normal distribution, $\hat{e}_{1,2,T}|\hat{e}_{1,2,S} = x \sim \mathcal{N}(\mu_T + \sigma_{I,T}\rho_I(x - \mu_S)/\sigma_{I,S}, 4\sigma_{I,T}^2(1-\rho_I^2)/n_{1,2})$, so again the identity from Bishop (2006) can be applied. All together, we obtain the following posterior distributions at the final analysis:

$$
\begin{aligned}
\boldsymbol{\mu}|\mathcal{F}_{2,A} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2,A}, \boldsymbol{\Sigma}_{2,A}) \\
\hat{\boldsymbol{\mu}}_{2,A} &:= \boldsymbol{\Sigma}_{2,A}(\boldsymbol{\Sigma}_{1,A}^{-1}\hat{\boldsymbol{\mu}}_{1,A} + n_{1,2}\mathbf{T}\hat{\mathbf{e}}_{1,2} + n_{2,2}\mathbf{T}\hat{\mathbf{e}}_{2,2}) \\
\boldsymbol{\Sigma}_{2,A} &:= (\boldsymbol{\Sigma}_{1,A}^{-1} + o_2\mathbf{T})^{-1} \\
&= (\boldsymbol{\Sigma}_0^{-1} + o\mathbf{T})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{2,B} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2,B}, \boldsymbol{\Sigma}_{2,B}) \\
\hat{\boldsymbol{\mu}}_{2,B} &:= \boldsymbol{\Sigma}_{2,B}(\boldsymbol{\Sigma}_{1,B}^{-1}\hat{\boldsymbol{\mu}}_{1,B} + n_{2,2}\mathbf{S}\hat{\mathbf{e}}_{2,2} + n_{2,3}\mathbf{S}\hat{\mathbf{e}}_{2,3}) \\
\boldsymbol{\Sigma}_{2,B} &:= (\boldsymbol{\Sigma}_{1,B}^{-1} + n_2\mathbf{S})^{-1} \\
&= (\boldsymbol{\Sigma}_0^{-1} + n\mathbf{S})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{2,C} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2,C}, \boldsymbol{\Sigma}_{2,C}) \\
\hat{\boldsymbol{\mu}}_{2,C} &:= \boldsymbol{\Sigma}_{2,C}(\boldsymbol{\Sigma}_{1,C}^{-1}\hat{\boldsymbol{\mu}}_{1,C} + n_{2,2}(\mathbf{S}+\mathbf{U})\hat{\mathbf{e}}_{2,2} + n_{2,3}\mathbf{S}\hat{\mathbf{e}}_{2,3} + n_{1,2}\mathbf{U}\hat{\mathbf{e}}_{1,2}) \\
\boldsymbol{\Sigma}_{2,C} &:= (\boldsymbol{\Sigma}_{1,C}^{-1} + n_2\mathbf{S} + o_2\mathbf{U})^{-1} \\
&= (\boldsymbol{\Sigma}_0^{-1} + n\mathbf{S} + o\mathbf{U})^{-1}
\end{aligned}
$$

A.3. Variance of Estimates

Finally, we can read the distribution of the estimates $\hat{\boldsymbol{\mu}}_{1,t}$, $\hat{\boldsymbol{\mu}}_{2,t}|\mathcal{F}_{1,t}$, and $\hat{\boldsymbol{\mu}}_{2,t}$:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{1,t} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \tilde{\boldsymbol{\Sigma}}_{1,t}) \\ \hat{\boldsymbol{\mu}}_{2,t}|\mathcal{F}_{1,t} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1,t}|\mathcal{F}_{1,t}, \tilde{\boldsymbol{\Sigma}}_{2,t}) \\ \hat{\boldsymbol{\mu}}_{2,t} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \tilde{\boldsymbol{\Sigma}}_{3,t})\,,\end{aligned}$$

where

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{1,t} &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{1,t} \\ \tilde{\boldsymbol{\Sigma}}_{2,t} &= \boldsymbol{\Sigma}_{1,t} - \boldsymbol{\Sigma}_{2,t} \\ \tilde{\boldsymbol{\Sigma}}_{3,t} &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{2,t}\,.\end{aligned}$$

For notational simplicity we denote $\tilde{\sigma}_{1,t,T}$ to be the bottom-right element in $\tilde{\boldsymbol{\Sigma}}_{1,t}$ (the variance of $\hat{\mu}_{1,t,T}$). We similarly define $\tilde{\sigma}^2_{2,t,T}$ and $\sigma^2_{3,t,T}$ from $\tilde{\boldsymbol{\Sigma}}_{2,t}$ and $\tilde{\boldsymbol{\Sigma}}_{3,t}$, respectively.

Appendix B: Proof of Theorem 1

Proof of Theorem 1 Consider a fixed trial type $t \in \{A, B, C\}$. Under Bayesian statistical inference for the trial (see Appendix A for details of the updates), the optimal adopt / no adopt decision if we terminate at the intermediate analysis is based on posterior mean $\hat{\mu}_{1,t,T}$ — in particular, we adopt whenever $a\hat{\mu}_{1,t,T} + b > 0$, yielding expected monetary benefit

$$\mathbb{E}\left[(a\mu_T + b) \cdot \mathbb{I}[a\hat{\mu}_{1,t,T} + b > 0] - c_p n_1 - c_w t_1\right] = (a\hat{\mu}_{1,t,T} + b)^+ - c_p n_1 - c_w t_1\,.$$

Similarly, the optimal adopt / no adopt decision at the final analysis is based on posterior mean $\hat{\mu}_{2,t,T}$ — we adopt whenever $a\hat{\mu}_{2,t,T} + b > 0$, yielding expected monetary benefit $(a\hat{\mu}_{2,t,T} + b)^+ - c_p n - c_w t_2$. The key decision, then, is whether to continue or terminate the trial at the intermediate analysis.

At the intermediate analysis, $\hat{\mu}_{2,t,T}$ is a random quantity — $\hat{\boldsymbol{\mu}}_{2,t}|\mathcal{F}_{1,t} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1,t}|\mathcal{F}_{1,t}, \tilde{\boldsymbol{\Sigma}}_{2,t})$, so $\hat{\mu}_{2,t,T}|\mathcal{F}_{1,t} \sim \mathcal{N}(\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t}, \tilde{\sigma}^2_{2,t,T})$. From the expectation of the rectified Gaussian distribution, the expected monetary benefit of the trial if it is continued all the way to completion is $\mathbb{E}[(a\hat{\mu}_{2,t,T} + b)^+|\mathcal{F}_{1,t}] - c_p n - c_w t_2 = a\tilde{\sigma}_{2,t,T}[\phi(\alpha_{1,t}) - \alpha_{1,t}(1 - \Phi(\alpha_{1,t}))] - c_p n - c_w t_2$, where $\alpha_t(x) = -\frac{ax+b}{a\tilde{\sigma}_{2,t,T}}$, $\alpha_{1,t} = \alpha_t(\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t})$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf, respectively.

The expected incremental benefit of stopping the trial early is

$$B_t(\alpha_{1,t}) = \begin{cases} c_p n_2 + c_w(t_2 - t_1) - a\tilde{\sigma}_{2,t,T}(\phi(\alpha_{1,t}) + \alpha_{1,t}\Phi(\alpha_{1,t})) & \text{if } \alpha_{1,t} < 0 \\ c_p n_2 + c_w(t_2 - t_1) - a\tilde{\sigma}_{2,t,T}[\phi(\alpha_{1,t}) - \alpha_{1,t}(1 - \Phi(\alpha_{1,t}))] & \text{otherwise}\,. \end{cases}$$

Note that $B_t(\alpha_{1,t})$ is a continuous function that satisfies $B_t(\alpha_{1,t}) = B_t(-\alpha_{1,t})$ and $\lim_{\alpha_{1,t}\to\infty} B_t(\alpha_{1,t}) = c_p n_2 + c_w(t_2 - t_1) > 0$. The derivative of $B_t(\alpha_{1,t})$ with respect to $\alpha_{1,t}$ is $-a\tilde{\sigma}_{2,t,T}(\Phi(\alpha_{1,t}) - \mathbb{I}[\alpha_{1,t} > 0])$, which is negative for $\alpha_{1,t} < 0$ and positive for $\alpha_{1,t} > 0$, meaning the minimum benefit of early stopping occurs when $\alpha_{1,t} = 0$ ($\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t} = -b/a$), with benefit $c_p n_2 + c_w(t_2 - t_1) - a\tilde{\sigma}_{2,t,T}/\sqrt{2\pi}$.

We can now fully characterize the optimal strategy at the intermediate analysis. If $c_p n_2 + c_w(t_2 - t_1) - a\tilde{\sigma}_{2,t,T}/\sqrt{2\pi} \geq 0$, then set $\delta_t = 0$. Otherwise, define $\alpha^*_{1,t}$ to be the unique solution to $B_t(\alpha_{1,t}) = 0$ for $\alpha_{1,t} > 0$;

by symmetry, $B_t(-\alpha^*_{1,t}) = 0$. $B_t(\alpha_{1,t}) \geq 0$ whenever $\alpha_{1,t} \geq \alpha^*_{1,t}$ $(\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t} \leq -b/a - \tilde{\sigma}_{2,t,T}\alpha^*_{1,t})$ and whenever $\alpha_{1,t} \leq -\alpha^*_{1,t}$ $(\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t} \geq -b/a + \tilde{\sigma}_{2,t,T}\alpha^*_{1,t})$; otherwise $B_t(\alpha_{1,t}) < 0$. Setting $\delta_t = \tilde{\sigma}_{2,t,T}\alpha^*_{1,t}$, it is optimal to continue the trial to the final analysis whenever $\hat{\mu}_{1,t,T}|\mathcal{F}_{1,t} \in (-b/a - \delta_t, -b/a + \delta_t)$ and to stop at the intermediate analysis otherwise.

At the beginning of the study, $\hat{\mu}_{1,t,T}$ is a random quantity — $\hat{\boldsymbol{\mu}}_{1,t} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \tilde{\boldsymbol{\Sigma}}_{1,t})$, so $\hat{\mu}_{1,t,T} \sim \mathcal{N}(\mu_{0,T}, \tilde{\sigma}^2_{1,t,T})$. Integration over $\hat{\mu}_{1,t,T}$ yields b^{eff}_t, the expected incremental benefit of early stopping due to efficacy, and b^{fut}_t, the expected incremental benefit of early stopping due to futility.

$$b^{fut}_t = \int_{-\infty}^{-b/a-\delta_t} B_t(\alpha_t(x)) \frac{1}{\sqrt{2\pi\tilde{\sigma}^2_{1,t,T}}} \exp\left(\frac{-(x-\mu_{0,T})^2}{2\tilde{\sigma}^2_{1,t,T}}\right) dx$$

$$b^{eff}_t = \int_{-b/a+\delta_t}^{\infty} B_t(\alpha_t(x)) \frac{1}{\sqrt{2\pi\tilde{\sigma}^2_{1,t,T}}} \exp\left(\frac{-(x-\mu_{0,T})^2}{2\tilde{\sigma}^2_{1,t,T}}\right) dx$$

Lastly, we establish b^{trial}_t, the expected benefit of a trial that never stops at the intermediate analysis, accepts the treatment if $a\hat{\mu}_{2,t,T} + b > 0$, and yields expected benefit $(a\hat{\mu}_{2,t,T} + b)^+ - c_p n - c_w t_2$. At the beginning of the study, $\hat{\mu}_{2,t,T}$ is a random quantity — $\hat{\boldsymbol{\mu}}_{2,t} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_0, \tilde{\boldsymbol{\Sigma}}_{3,t})$, so $\hat{\mu}_{2,t,T} \sim \mathcal{N}(\mu_{0,T}, \tilde{\sigma}^2_{3,t,T})$. The expectation of the rectified Gaussian distribution yields $b^{trial}_t = a\tilde{\sigma}_{3,t,T}[\phi(\tau_t) - \tau_t(1 - \Phi(\tau_t))] - c_p n - c_w t_2$, for $\tau_t := -\frac{a\mu_{0,T}+b}{a\tilde{\sigma}_{3,t,T}}$. □

Appendix C: Trial Designs with Time-to-Event Outcomes

In this appendix, we modify the model with continuous outcomes, no delay for the surrogate outcome, and a fixed time delay Δ for the true outcome to instead capture time-to-event outcomes — the surrogate is observed with an uncertain delay from the time of enrollment, and the true outcome is observed with some additional uncertain delay after the surrogate is observed. This setting is ubiquitous in the clinical trial literature (for instance, it appears in the setting considered in Section 4), but has a number of complexities that makes it less suitable for the comparative statics approach taken in Section 3 of this work.

Unlike in the setting with continuous outcomes with known variance, the distribution of the surrogate and true outcome survival times will play a role in optimal trial designs. Here, we assume the surrogate and true outcome times are exponentially distributed, with known rates λ_S and λ_T for the surrogate and true outcome, respectively, in the control arm of the trial. Since the control group has often been tested extensively in previous clinical trials, assuming these rates are known is not an unreasonable assumption. The effect size in the time-to-event setting is log hazard ratio; under the proportion hazards assumption, a trial with true log hazard ratio $\boldsymbol{\mu} = [\mu_S, \mu_T]'$ will have exponentially distributed treatment arm surrogate and true outcome times with rates $e^{\mu_S}\lambda_S$ and $e^{\mu_T}\lambda_T$, respectively.

Consider a group of n patients randomly allocated with equal probabilities to a treatment and control group, and assume all patients have observed both their surrogate and true outcome event. If we define $\hat{\mathbf{e}}$ to be the logrank test estimates of the surrogate and true outcome log hazard ratios in a trial with true log hazard ratio vector $\boldsymbol{\mu}$, then asymptotic analysis of the logrank test with the Martingale Central Limit Theorem yields approximate normality: $\hat{\mathbf{e}} \sim \mathcal{N}(\boldsymbol{\mu}, (4/n)\boldsymbol{\Sigma}_I)$ for $\boldsymbol{\Sigma}_I$ with sampling variance 1 for both outcomes and correlation ρ_I. Since the variance of the effect size estimate depends on event counts (a random quantity),

Outcome Parameters (Exogenous)	
μ_{0S}	Study-level mean surrogate effect size
μ_{0T}	Study-level mean true outcome effect size
σ^2_{0S}	Study-level surrogate effect size variance
σ^2_{0T}	Study-level true outcome effect size variance
ρ_0	Study-level correlation of surrogate and true outcome effect size
ρ_I	Individual-level correlation of surrogate and true outcome effect size
λ_S	Exponential distribution rate for the surrogate outcome in the control arm
λ_T	Exponential distribution rate for the true outcome in the control arm
Economic Parameters (Exogenous)	
$-a$	Population monetary benefit of a unit decrease in the true outcome effect size
$-b$	Threshold of population monetary benefit needed to accept a new treatment
c_p	Monetary cost of enrolling one patient in a clinical trial
c_w	Monetary cost of waiting one additional unit of time
n^{max}	Maximum number of patients budgeted for in the clinical trial
Trial Design Parameters (Endogenously Selected by Designer Before Study)	
n	Target number of patients enrolled (one per time unit)
v_1	Posterior variance at which the intermediate analysis will occur
v_2	Posterior variance at which the final analysis will occur

Table C.1 Parameters in model of clinical trials with time-to-event outcomes

we take a common trial design choice, defining the intermediate and final analysis to occur after a fixed number of events (more generally, once the posterior variance reaches a target level v_1 for the intermediate analysis and v_2 for the final analysis). Table C.1 summarizes the parameters of our clinical trial design with time-to-event outcomes.

For a given trial arm (treatment or control) and outcome (surrogate or true outcome), if we assume patient enrollments that are uniform at random through an enrollment period $[0, T]$, that event times are exponentially distributed with arrival rate λ, and that there is no censoring, then the probability a patient will have had an observed event by time w, $E(w, \lambda, T)$, satisfies

$$\begin{aligned} E(w,\lambda,T) &= \int_0^{\min(w,T)} \frac{1-e^{-\lambda(w-s)}}{T} ds \\ &= \frac{1}{\lambda T} \cdot \begin{cases} \lambda w + e^{-\lambda w} - 1 & \text{if } w \le T \\ \lambda T + e^{-\lambda w}(1-e^{\lambda T}) & \text{otherwise} \end{cases} \end{aligned}$$

Considering an arbitrary patient in the study population (randomly assigned with probability 0.5 to either the control or treatment group), the probability their surrogate and true event times occur before time w under surrogate hazard ratio μ_S and true outcome hazard ratio μ_T are:

$$\begin{aligned} A_S(w,\mu_S) &= (E(w,\lambda_S,T) + E(w,e^{\mu_S}\lambda_S,T))/2 \\ A_T(w,\mu_T) &= (E(w,\lambda_T,T) + E(w,e^{\mu_T}\lambda_T,T))/2 \end{aligned}$$

For each trial type $t \in \{A, B, C\}$ with fixed trial design parameters n, v_1, and v_2, let $\sigma^2_{t,T}(w, \boldsymbol{\mu})$ be the posterior variance of the true outcome estimate after $nA_S(w, \mu_S)$ surrogate events are observed and $nA_T(w, \mu_T)$ true outcome events are observed. As established in Appendix A, this is a function of w and μ_T for type A trials, a function of w and μ_S for type B trials, and a function of w, μ_S, and μ_T for type C trials. Define $w^t_1(\boldsymbol{\mu})$ and $w^t_2(\boldsymbol{\mu})$ to be the unique solutions to $\sigma^2_{t,T}(w^t_1(\boldsymbol{\mu}), \boldsymbol{\mu}) = v_1$ and $\sigma^2_{t,T}(w^t_2(\boldsymbol{\mu}), \boldsymbol{\mu}) = v_2$, respectively. For sufficiently large n, the time until the intermediate (final) analysis is normally distributed with

mean $w_1^t(\boldsymbol{\mu})$ ($w_2^t(\boldsymbol{\mu})$). Similarly, the expected proportion of patients enrolled at the intermediate analysis is $p_1^t(\boldsymbol{\mu}) = \max(w_1^t(\boldsymbol{\mu})/T, 1)$ and at the final analysis is $p_2^t(\boldsymbol{\mu}) = \max(w_2^t(\boldsymbol{\mu})/T, 1)$.

We are now prepared to update our optimal clinical trial design with a single intermediate analysis. The development closely follows the proof of Theorem 1, noting that $\tilde{\sigma}_{1,t,T} = \sigma_{0,T}^2 - v_1$, $\tilde{\sigma}_{2,t,T} = v_1 - v_2$, and $\tilde{\sigma}_{3,t,T} = \sigma_{0,T}^2 - v_2$. First, consider the benefit of a full trial without any intermediate analysis for trial types $t \in \{A, B, C\}$:

$$b_t^{trial} = |a|\tilde{\sigma}_{3,t,T}[\phi(\tau_t) - \tau_t(1 - \Phi(\tau_t))] - P_t^{trial} - W_t^{trial}$$
$$\tau_t := -\frac{a\mu_{0,T} + b}{|a|\tilde{\sigma}_{3,t,T}}$$

Here, P_t^{trial} represents the expected cost of patient enrollment and W_t^{trial} represents the expected cost of waiting. Under the continuous outcome model, we have $P_t^{trial} = c_p n$ and $W_t^{trial} = c_w t_2$. For time-to-event outcomes, we could for instance obtain W_t^{trial} by integrating $c_w w_2^t(\boldsymbol{\mu})$ over the prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. However, for simplicity we use the approximations $W_t^{trial} \approx c_w w_2^t(\boldsymbol{\mu}_0)$ and $P_t^{trial} \approx c_p p_2^t(\boldsymbol{\mu}_0)$, which perform well in practice.

We now turn our attention to the other piece of optimal trial design: optimal stopping at the intermediate analysis. We limit our attention to threshold-based stopping rules based on the posterior true outcome effect size mean $\hat{\mu}_{1,t,T}$. Again following the proof of Theorem 1, the expected incremental benefit of early stopping when $\hat{\mu}_{1,t,T} = x$ is $B_t(x) = P_t^{int}(x) + W_t^{int}(x) - |a|\tilde{\sigma}_{2,t,T}(\phi(\alpha_t(x)) + \alpha_t(x)\Phi(\alpha_t(x)) - \alpha_t(x)^+)$, where $\alpha_t(x) := -\frac{ax+b}{|a|\tilde{\sigma}_{2,t,T}}$. Here, $P_t^{int}(x)$ represents the expected patient enrollment costs saved by early stopping at the intermediate analysis when the true outcome posterior mean is x. Similarly, $W_t^{int}(x)$ represents the expected waiting costs saved by early stopping.

The following procedure provides good-quality approximations to the optimal stopping rule. Noting that $\hat{\mu}_{1,t} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{1,t})$, let $s(x)$ be the mean of $\hat{\mu}_{1,t,s}|\hat{\mu}_{1,t,T} = x$ and use approximations $P_t^{int}(x) \approx c_p(p_2^t([s(x), x]') - p_1^t([s(x), x]'))$ and $W_t^{int}(x) \approx c_w(w_2^t([s(x), x]') - w_1^t([s(x), x]'))$. If $B_t(-b/a) \geq 0$ then set $\underline{\mu_t} = \overline{\mu_t} = -b/a$, and otherwise use 1-d root finding to identify $\underline{\mu_t} < -b/a$ such that $B_t(\underline{\mu_t}) = 0$ and $\overline{\mu_t} > -b/a$ such that $B_t(\overline{\mu_t}) = 0$. We continue to the final analysis if $\hat{\mu}_{1,t,T} \in (\underline{\mu_t}, \overline{\mu_t})$ and otherwise we stop the trial at the intermediate analysis. Unlike in the case of continuous outcomes, $\underline{\mu_t}$ and $\overline{\mu_t}$ are not necessarily symmetric around $-b/a$.

Good quality values of n, v_1, and v_2 can be obtained by selecting random n, v_1, and v_2 and then maximizing $b_t^{trial} + b_t^{eff} + b_t^{fut}$ over the (n, v_1, v_2) space from that starting point using the Nelder-Mead simplex. This procedure can then be repeated as many times as desired, with the best-quality solution identified across the replicates selected as the final design.

Appendix D: Comparative Statics

We start with the following lemma for deriving our first- and second-order approximations.

LEMMA 4. *Given a symmetric matrix $X \in \mathbb{R}^{2\times 2}$ and the identity matrix I,*

$$(I + X)^{-1} = I - X + X^2 + \mathcal{O}\left(\lambda_{\max}(X)^3\right).$$

Proof of Lemma 4 Since X is a symmetric matrix, there exists a unitary matrix U and a diagonal matrix $\Sigma = \begin{bmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{bmatrix}$ such that $X = U\Sigma U'$. Furthermore, the eigenvalues of X are σ_0 and σ_1. We can then write

$$\begin{aligned}
(I+X)^{-1} &= (I+U\Sigma U')^{-1} \\
&= U'(I+\Sigma)^{-1}U \\
&= U'\left(\begin{bmatrix} 1+\sigma_0 & 0 \\ 0 & 1+\sigma_1 \end{bmatrix}\right)^{-1} U \\
&= U'\left(\begin{bmatrix} \frac{1}{1+\sigma_0} & 0 \\ 0 & \frac{1}{1+\sigma_1} \end{bmatrix}\right) U \\
&= U'\left(\begin{bmatrix} 1-\sigma_0+\sigma_0^2 & 0 \\ 0 & 1-\sigma_1+\sigma_1^2 \end{bmatrix}\right)^{-1} U \;+\; \mathcal{O}(\sigma_0^3+\sigma_1^3) \\
&= U'(I-\Sigma+\Sigma^2)U + \mathcal{O}(\sigma_0^3+\sigma_1^3) \\
&= I - X + X^2 + \mathcal{O}(\sigma_0^3+\sigma_1^3)\,.
\end{aligned}$$

□

Proof of Lemma 3 We seek to approximate b_C^{trial} to second order in n so that we can compute tractable comparative statics. Recall from Theorem 1 that

$$b_C^{trial} = a\tilde{\sigma}_{3,C,T}[\phi(\tau_C) - \tau_C(1-\Phi(\tau_C))] - c_p n - c_w t_2\,,$$

where $\tau_C = -(a\mu_{0,T}+b)/(a\tilde{\sigma}_{3,C,T})$. Recall further from Section 3.1 that we are considering the case where $\mu_{0,T} = b = 0$ and we constrain $t_2 = n+\Delta$, *i.e.*, we wait to observe all true outcomes for all enrolled patients. Thus, we can simplify

$$b_C^{trial} = \frac{a}{\sqrt{2\pi}} \cdot \tilde{\sigma}_{3,C,T} - c_p n - c_w(n+\Delta)\,.$$

Then, it suffices to approximate $\tilde{\sigma}_{3,C,T}$. For $\Sigma = 4\Sigma_I$, we can write

$$\begin{aligned}
\tilde{\Sigma}_{3,C} &= \Sigma_{2,C}(n^2\Sigma^{-1}\Sigma_0\Sigma^{-1} + n\Sigma^{-1})\Sigma_{2,C} \\
\Sigma_{2,C} &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}\,.
\end{aligned}$$

Since n is large relative to $\lambda_{\max}\left(\Sigma_0^{-1}\right)$, we can apply Lemma 4 to separate higher-order terms in n, *i.e.*,

$$\begin{aligned}
\Sigma_{2,C} &= \frac{1}{n}\left(\frac{1}{n}\Sigma\Sigma_0^{-1} + I\right)^{-1}\Sigma \\
&= \frac{1}{n}\left(I - \frac{1}{n}\Sigma\Sigma_0^{-1} + \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma\Sigma_0^{-1} + \mathcal{O}\left(\frac{1}{n^3}\right)\right)\Sigma \\
&= \frac{1}{n}\Sigma - \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma + \frac{1}{n^3}\Sigma\Sigma_0^{-1}\Sigma\Sigma_0^{-1}\Sigma + \mathcal{O}\left(\frac{1}{n^4}\right)\,.
\end{aligned}$$

Next, applying the above approximation to the definition of $\tilde{\Sigma}_{3,C}$, we have

$$\begin{aligned}
\tilde{\Sigma}_{3,C} = &\left(\frac{1}{n}\Sigma - \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma + \frac{1}{n^3}\Sigma\Sigma_0^{-1}\Sigma\Sigma_0^{-1}\Sigma + \mathcal{O}\left(\frac{1}{n^4}\right)\right)\cdot\left(n^2\Sigma^{-1}\Sigma_0\Sigma^{-1} + n\Sigma^{-1}\right) \\
&\cdot\left(\frac{1}{n}\Sigma - \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma + \frac{1}{n^3}\Sigma\Sigma_0^{-1}\Sigma\Sigma_0^{-1}\Sigma + \mathcal{O}\left(\frac{1}{n^4}\right)\right) \\
= &\,\Sigma_0 - \frac{1}{n}\Sigma + \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma + \mathcal{O}\left(\frac{1}{n^3}\right)\,.
\end{aligned}$$

Comparative statics in Lemma 3 follow from taking derivatives of the expression given above for b_C^{trial}, where $\tilde{\sigma}_{3,C,T} = \tilde{\Sigma}_{3,C}[2,2]$ is evaluated using the second-order approximation

$$\tilde{\Sigma}_{3,C} = \Sigma_0 - \frac{1}{n}\Sigma + \frac{1}{n^2}\Sigma\Sigma_0^{-1}\Sigma + \mathcal{O}\left(\frac{1}{n^3}\right).$$

□

Proof of Theorem 2 Using the second-order n approximation for $\tilde{\Sigma}_{3,C}$ prescribed in the proof of Lemma 3, we can compute

$$b_C^{trial} - b_A^{trial} = \frac{a}{\sqrt{2\pi}} \cdot \left(1 - \frac{4R_T^2}{n} + \frac{16R_T^2(\rho_I^2 R_S^2 - 2\rho_0\rho_I R_S R_T + R_T^2)}{(1-\rho_0^2)n^2} - \frac{n}{n+4R_T^2}\right) + \mathcal{O}\left(\frac{1}{n^3}\right)$$
$$b_C^{trial} - b_B^{trial} = \frac{a}{\sqrt{2\pi}} \cdot \left(1 - \frac{4R_T^2}{n} + \frac{16R_T^2(\rho_I^2 R_S^2 - 2\rho_0\rho_I R_S R_T + R_T^2)}{(1-\rho_0^2)n^2} - \frac{n\rho_0^2}{n+4R_S^2}\right) - c_w\Delta + \mathcal{O}\left(\frac{1}{n^3}\right).$$

We are interested in comparative statics on $b_C^{trial} - \max\{b_A^{trial}, b_B^{trial}\}$. From $\lim_{n\to\infty} b_A^{trial} - b_B^{trial} = a(1-\rho_0^2)/\sqrt{2\pi} - c_w\Delta$, we conclude that, for large n, trial type A will be selected as the comparator when $a(1-\rho_0^2)/\sqrt{2\pi} > c_w\Delta$, and otherwise trial type B will be selected as the comparator.

Several of the cases are straightforward for $b_C^{trial} - b_A^{trial}$: the incremental benefit of type C is monotone increasing in a and unaffected by Δ. Since the benefit of trial type A does not depend on parameters ρ_0, ρ_I, or R_S, we conclude from Lemma 3 that $b_C^{trial} - b_A^{trial}$ is non-monotone in these three parameters. Algebra indicates that the derivative with respect to R_T of the incremental difference is $16a\sqrt{2/\pi}R_T(R_T\rho_0 - \rho_I R_S)(2R_T\rho_0 - \rho_I R_S)/[(n+4R_T^2)^2(1-\rho_0^2)]$, which is non-monotone.

Several of the cases are also straightforward for $b_C^{trial} - b_B^{trial}$: the incremental benefit of type C is monotone increasing in a and monotone decreasing in Δ. Since the benefit of trial type B does not depend on parameters ρ_I or R_T, we conclude from Lemma 3 that $b_C^{trial} - b_B^{trial}$ is non-monotone in $|\rho_I|$ and monotone decreasing in R_T. Algebra indicates that the derivative with respect to R_S is $8\rho_0^2 R_S/n + \mathcal{O}(1/n^2)$, so the incremental benefit is monotone increasing in R_S. Finally, the derivative with respect to ρ_0 is $-2\rho_0 + \mathcal{O}(1/n)$, so the incremental benefit is monotone decreasing in $|\rho_0|$. □

Appendix E: Bivariate Normality Assumption

In order to see whether our assumption that the effect sizes of surrogate and true outcomes follow a bivariate normal distribution, we collected effect sizes across clinical trials for 30 different true and surrogate outcome pairs from 19 different cancers. The effect sizes across clinical trials for each surrogate/true outcome pair is plotted in Figure 7.

To collect these effect sizes, we searched PubMed for meta-analyses of surrogate endpoints. Our initial search returned 80 results. Of these, 21 reported effect sizes for both true and surrogate endpoints from each of the trials included in the meta-analysis; some of these studies looked at multiple potential surrogate endpoints. From these 21 studies we were able to gather information on the effect sizes of 30 true and surrogate outcome pairs across 19 different cancer subtypes. In all cases, the true outcome was the log overall survival hazard ratio. The surrogates varied; time-to-event surrogates included log hazard ratios for progression-free survival, time to progression, disease-free survival, and failure-free survival, while binary surrogates included log odds ratios for overall response rate and disease control rate.

Next, for each surrogate/true outcome pair, we tested the study-level effect sizes for bivariate normality using the Henze-Zirkler test, a test of the null hypothesis that the data are distributed according to a multivariate normal distribution (Henze and Zirkler 1990). Out of the 30 pairs, the test rejected the null hypothesis that the effect sizes were multivariate normally distributed for eight pairs and failed to reject the null hypothesis for the remaining 22 pairs.

Appendix F: Pseudocode for MBC Trial

Algorithm 2 provides detailed pseudocode for simulating the metastatic breast cancer clinical trial.

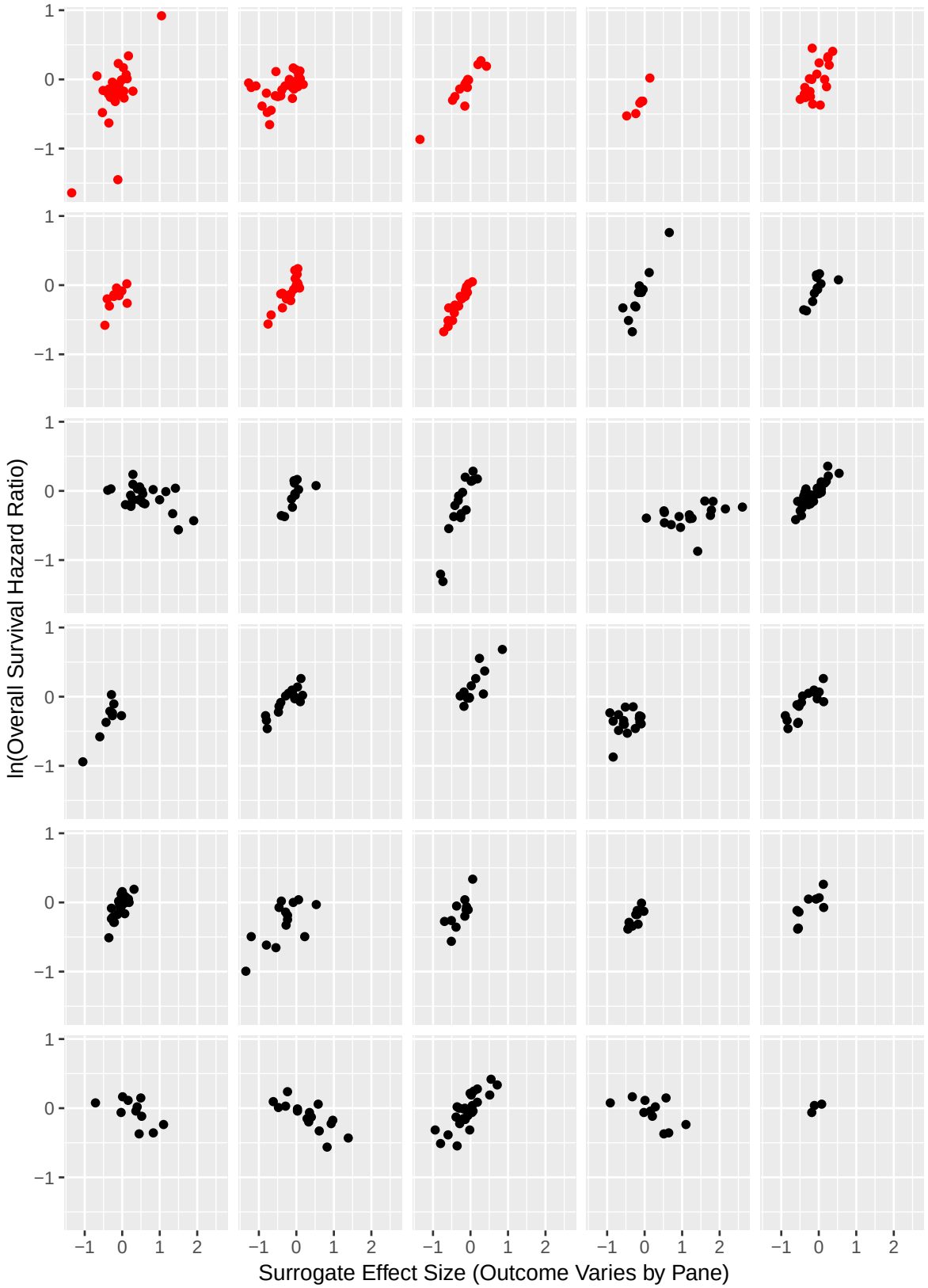


Figure 7 **Surrogate and true outcome effect sizes of cancer studies across 30 surrogate/true outcome pairs from 19 cancer subtypes. The eight surrogate/true outcome pairs with red points reject the null hypothesis of bivariate normality ($p \leq 0.05$), while the remaining 22 with black points do not ($p > 0.05$).**

Algorithm 2 Detailed Pseudocode for Simulating a Clinical Trial for MBC

Input: Study i; trial type t; MBC trial design parameters (see Table 4.2); and $F_C(s,t)$ ($F_E(s,t)$), the empirical distribution of control (experiment) group event times for the surrogate and true outcome, joined by a Gaussian copula with parameter $\rho_{ctl,i}$ ($\rho_{exp,i}$)
for $r = 1, \ldots, 1000$ **do**
  for $p = 1, \ldots, n_t$ **do**
    $enrl_p \leftarrow (p-1)/\lambda_E$
    if $rand() \leq 0.5$ **then**
      $ctl_p \leftarrow 1$
      $(surr_p, true_p) \leftarrow$ random draw from $F_C(s,t)$
    else
      $ctl_p \leftarrow 0$
      $(surr_p, true_p) \leftarrow$ random draw from $F_E(s,t)$
    end if
    $surr_p \leftarrow \min(surr_p, true_p)$
  end for
  $sTimes \leftarrow \{enrl_p + surr_p \ \forall p = 1, \ldots, n_t\}$
  $tTimes \leftarrow \{enrl_p + true_p \ \forall p = 1, \ldots, n_t\}$
  for $c \in sTimes \cup tTimes$ **do**
    $n_1 \leftarrow |\{p \in \{1, \ldots, n_t\} | enrl_p + surr_p \leq c\}|$
    $n_{11} \leftarrow |\{p \in \{1, \ldots, n_t\} | enrl_p + true_p \leq c\}|$
    $v_c \leftarrow$ the bottom-right element of $\mathbf{\Sigma}_{1,t}$ computed using n_1 and n_{11}, as in Appendix A
  end for
  $t_{1tr} \leftarrow \min\{c \in sTimes \cup tTimes | v_c \leq v_{1t}\}$
  $t_{2tr} \leftarrow \min\{c \in sTimes \cup tTimes | v_c \leq v_{2t}\}$
  $n_{1tr} \leftarrow |\{p \in \{1, \ldots, n_t\} | enrl_p \leq t_{1tr}\}|$
  $n_{2tr} \leftarrow |\{p \in \{1, \ldots, n_t\} | enrl_p \leq t_{2tr}\}|$
  $O_{11} \leftarrow \{p \in \{1, \ldots, n_t\} | enrl_p + true_p \leq t_{1tr}\}$
  $O_{12} \leftarrow \{p \in \{1, \ldots, n_t\} | enrl_p + surr_p \leq t_{1tr}$ and $t_{1tr} < enrl_p + true_p \leq t_{2tr}\}$
  $O_{13} \leftarrow \{p \in \{1, \ldots, n_t\} | enrl_p + surr_p \leq t_{1tr}$ and $enrl_p + true_p > t_{2tr}\}$
  $O_{22} \leftarrow \{p \in \{1, \ldots, n_t\} | t_{1tr} < enrl_p + surr_p \leq t_{2tr}$ and $t_{1tr} < enrl_p + true_p \leq t_{2tr}\}$
  $O_{23} \leftarrow \{p \in \{1, \ldots, n_t\} | t_{1tr} < enrl_p + surr_p \leq t_{2tr}$ and $enrl_p + true_p > t_{2tr}\}$
  for $(i,j) \in \{(1,1), (1,2), (1,3), (2,2), (2,3)\}$ **do**
    $n_{ij} \leftarrow |O_{ij}|$
    $\hat{e}_{ijS} \leftarrow$ logrank test statistic from ctl_p and $surr_p$ for $p \in O_{ij}$, multiplied by $2/\sqrt{n_{ij}}$
    $\hat{e}_{ijT} \leftarrow$ logrank test statistic from ctl_p and $true_p$ for $p \in O_{ij}$, multiplied by $2/\sqrt{n_{ij}}$
  end for
  $\hat{\mu}_{1tT}, \hat{\mu}_{2tT} \leftarrow$ posterior means computed using $\hat{\mathbf{e}}_{ij}$ and n_{ij} (see Appendix C)
  if $\hat{\mu}_{1tT} \leq \underline{\mu_t}$ **then**
    $e_{tr} \leftarrow 1;\ a_{tr} \leftarrow 1$
  else if $\hat{\mu}_{1tT} > \overline{\mu_t}$ **then**
    $e_{tr} \leftarrow 1;\ a_{tr} \leftarrow 0$
  else if $\hat{\mu}_{2tT} < -b/a$ **then**
    $e_{tr} \leftarrow 0;\ a_{tr} \leftarrow 1$
  else
    $e_{tr} \leftarrow 0;\ a_{tr} \leftarrow 0$
  end if
end for
return all t_{1tr}, n_{1tr}, e_{tr} and a_{tr} values