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 approve drugs 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 late-phase 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 design would yield a 16% decrease in trial costs relative to existing clinical trial designs, while maintaining the same Type I/II error rates.

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

1. Introduction

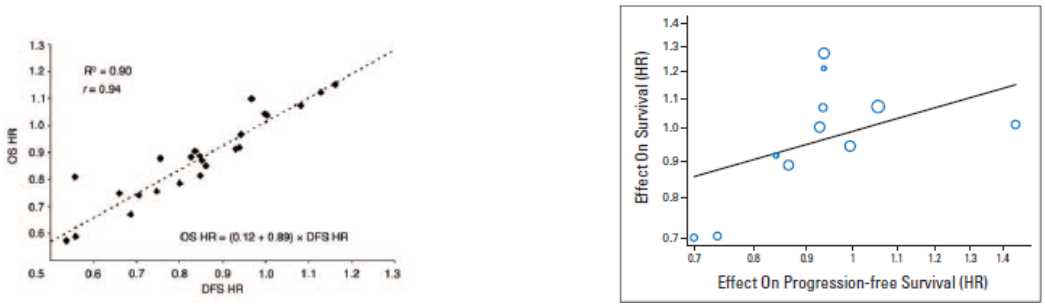
Randomized controlled trials in the medical domain seek to determine a new treatment's efficacy quickly and accurately. Longer clinical trials can delay the release of an effective treatment, incurring significant financial and population health costs. For instance, the average yearly revenue of an approved cancer drug is estimated to be around $550 million; thus, every extra day that an effective drug spends in a clinical trial comes at a cost of more than a million 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, as has prominently been the case with SARS-CoV-2 vaccine development and COVID-19 treatments. This has spurred the United States Food and Drug Administration (FDA), which handles US regulatory approvals for drugs, to create the 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 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) or patient death. In contrast to overall survival, the median duration of progression-free survival is only 7 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). In recent years, 66% of oncology drugs have been approved on the basis of a surrogate outcome (Kemp and Prasad 2017).

The challenge is that the surrogate outcome is generally not 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 colon cancer (left panel; Sargent et al. 2005), and for metastatic breast cancer (MBC) (right panel; Burzykowski et al. 2008). Evidently, the surrogate outcome is predictive of



(a) (b)

Figure 1 Study-level relationship between a surrogate (x-axis) and true outcome (y-axis) for (a) colon cancer (Sargent et al. 2005) and (b) metastatic breast cancer (Burzykowski et al. 2008).

the true outcome in both cases, but it is a more accurate predictor in the case of colon cancer compared to MBC.

Thus, when deciding whether to approve a drug based on results for a surrogate outcome (rather than the true outcome), 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). While drugs approved on the basis of a surrogate outcome must still undergo post-approval experimentation to verify effectiveness on the true outcome, this process takes many years, *e.g.,* out of 134 accelerated approvals in oncology from 1992-2014, only 55% were confirmed or withdrawn within 5 years, and 73% of withdrawals took over 8 years (Kemp and Prasad 2017). This long turnaround time amplifies the importance of an accurate initial drug approval decision.

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 less costly for a target Type I/II error rate, *i.e.*, the

decision-maker can successfully reject ineffective treatments and accept effective treatments with shorter trials and/or lower patient enrollments. There are several advantages to such an approach. First, the proposed design accounts for the imperfect relationship between the surrogate and true outcome when making key trial decisions. Incorporating a small number of true outcome observations in addition to surrogate outcome data 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 often 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 to illustrate 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 93 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 16% reduction in trial costs (roughly $29.2 million); this improvement is largely because our design can meet target

Type I/II error rates with significantly shorter trials than standard designs. 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 (typically Phase III) clinical trial under sequential patient recruitment with both surrogate and true outcome observations. Following common clinical trial practice, we model a single interim analysis for early stopping; for both the interim 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 interim and final analyses.

2. *Guidelines:* We perform comparative statics on several key properties of diseases and surrogates; these results illustrate 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 16% decrease in trial costs relative to traditional designs.

1.2. Related Literature

The design of efficient clinical trials is well-studied in the statistics and healthcare operations literatures. 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.

Our model builds on a large literature that incorporates salient features of clinical trial decision-making. Some examples include analysis of the benefits and impacts of interim analyses (Rojas-Cordova and Bish 2018, Rojas-Cordova and Hosseinichimeh 2018), patient recruitment and stopping decisions within these interim analyses (Kouvelis et al. 2017), evaluating the downstream

impact of different statistical criteria for drug approval decisions (Corcoran et al. 2019), incorporating delayed outcome observations in continuous monitoring designs (Chick et al. 2017), and leveraging historical clinical trial data to design new therapies (Bertsimas et al. 2016).

The above literature all 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 or across all the endpoints (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. Several papers have since further explored these criteria (Freedman et al. 1992, Buyse and Molenberghs 1998, Burzykowski et al. 2001, Renard et al. 2002, Burzykowski et al. 2004, Spiegelhalter et al. 2004, 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 imperfect relationship between the surrogate and true outcome. 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.

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 imperfect relationship between the surrogate and true outcome in a principled way when analyzing a trial that ultimately selects the surrogate as the primary endpoint. 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 furthermore identifies the types of diseases and surrogates where such a design can yield the most value.

Surrogates have also found use in other fields such as development economics (Athey et al. 2019) and personalized decision-making (Bastani 2020). We focus on clinical trial designs, but future work could explore how our methods can be used for adaptive A/B testing in other contexts.

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 trial outcomes can be of a number of types, including binary-valued, continuous-valued, count, categorical, or time-to-event. To simplify the exposition, we focus on time-to-event true and surrogate outcomes. Here, patients in the treatment and control groups experience events at some random time after enrollment, and the clinical trial is run to establish whether the rates of events differ between the two groups. The use of surrogates is well motivated when dealing with a time-to-event true outcome for which events take a long time to observe, since waiting to observe true outcome events can extend the length of the clinical trial and slow drug approval decisions. As the examples in Figure 1 capture, surrogacy relationships with time-to-event surrogate and true outcomes are ubiquitous in oncology clinical trials, generally with time to disease recurrence or time to disease progression serving as a surrogate for the true outcome of a patient's overall survival (time to death).

In clinical trials, the performance of one treatment versus another on some outcome is measured via an *effect size*. Effect sizes vary depending on the outcome type; for time-to-event outcomes, the predominant effect size is the *(log) hazard ratio*. The (time-dependent) hazard ratio

$$HR(t) = \frac{h_T(t)}{h_C(t)}$$

is the ratio of treatment group hazard function $h_T(t)$ to control group hazard function $h_C(t)$. These hazard functions capture the instantaneous rate of events among individuals who have not yet had an event. For a continuous time model where the control (treatment) group event time is a random variable with pdf f_C (f_T) and cdf F_C (F_T), the hazard functions are defined as

$$h_C(t) = \frac{f_C(t)}{1 - F_C(t)} \quad \text{and} \quad h_T(t) = \frac{f_T(t)}{1 - F_T(t)}.$$

Since most time-to-event outcomes used in clinical trials measure the time to some undesirable event (*e.g.*, disease progression or death), we generally desire $HR(t) < 1$, implying that the treatment decreases the hazard rate compared to the control.

Numerous methods such as the Cox proportional hazards model (Cox 1972) and various calculations derived from the logrank test (Machin et al. 2006) can be used to estimate the hazard ratio between two treatments based on the timing of events observed in patients in the control and treatment group, taking into account the fact that some patients may have been lost to follow-up, or may not have had their event time observed by the time of the analysis (these patients are referred to as having been *censored* for the outcome). Under the *proportional hazards* assumption where $\ln(HR(t)) = \mu$ for some time-invariant constant μ, these methods return estimates of μ.[1]

The asymptotics of various methods for estimating the log hazard ratio are well studied (see, *e.g.*, Schoenfeld 1981 and Sugimoto et al. 2013). We rely on this theory to study clinical trials with a sufficiently large set of n patients who are randomly assigned with equal probability between the control and treatment arms, under a setting where the proportional hazards assumption holds and patients in the control and treatment arms have the same patterns of censoring. 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 log hazard ratio and μ_T (the value we seek to identify) is the true outcome log hazard ratio. Then the Mantel-Haenszel hazard ratio estimate (Machin et al. 2006) of the surrogate and true outcome log hazard ratios among the p_J proportion of the trial's n patients who have observed both a surrogate and true outcome, $\hat{\mathbf{e}}^{\mathbf{J}} = [\hat{e}_S^J \ \hat{e}_T^J]'$, along with its error term $\boldsymbol{\epsilon}^{\mathbf{J}} = [\epsilon_S^J \ \epsilon_T^J]'$, are approximately bivariate normally distributed for sufficiently large n:

$$\hat{\mathbf{e}}^{\mathbf{J}} = \boldsymbol{\mu} + \boldsymbol{\epsilon}^{\mathbf{J}} \quad \text{and} \quad \boldsymbol{\epsilon}^{\mathbf{J}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_I/(p_J n)),$$
$$\boldsymbol{\Sigma}_I = \begin{bmatrix} 4 & 4\rho_I \\ 4\rho_I & 4 \end{bmatrix}.$$

[1] Though we do not consider the situation in this work, these methods are also widely used when the proportional hazards assumption is violated. In that case, these methods return $\ln(HR(t))$ averaged across time t, and different methods apply different patterns of time averaging.

Furthermore, the surrogate outcome effect size estimate $\hat{e}_S^N$ among the p_N proportion of the trial's patients who have observed only the surrogate outcome is also approximately normally distributed for a sufficiently large trial:

$$\hat{e}_S^N = \mu_S + \epsilon_S^N \quad \text{and} \quad \epsilon_S^N \sim \mathcal{N}(0, 4/(p_N n)),$$

where ϵ_S^N is independent of $\boldsymbol{\epsilon}^{\mathbf{J}}$.

ASSUMPTION 1. *We adopt a Bayesian framework, and assume that the effect size* $\boldsymbol{\mu}$ *for a given study is drawn from a zero-mean bivariate normal distribution:*

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

with variance σ_{0S}^2 *for the surrogate effect size, variance* σ_{0T}^2 *for the true outcome effect size, and correlation* ρ_0*.*

The ethical concept of *equipoise* argues that average clinical trial effect sizes should be near 0, validating the assumed mean of the prior distribution. We collected effect sizes across clinical trials for 30 different time-to-event true outcomes and their corresponding surrogate outcomes from 19 different diseases in the medical literature (for details, see Appendix F), revealing that average effect sizes are indeed close to 0 in practice. The assumption of bivariate normality 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 in practice (we illustrate this on our metastatic breast cancer case study in Section 4).

REMARK 1. Note that we have modeled two different types of correlations between surrogate and true outcomes: 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 an accurate estimate of the surrogate effect size μ_S — do drugs that improve the surrogate outcome tend to also improve the true outcome? For large values of $|\rho_0|$, an accurate estimate of surrogate effect size μ_S will yield an accurate estimate of true outcome effect size μ_T, even if we have observed very few or no true outcome events. In contrast, ρ_I determines the degree to which we can learn about a patient's true outcome event time from their surrogate event time — are the two event times

highly correlated? Large $|\rho_I|$ unlocks learning from *paired* true and surrogate outcome observations for individual patients. For instance, consider a study where many patients have had a surrogate event but no true outcome event, yielding an accurate estimate of μ_S. If the study also has a small number of patients with both event types, then this smaller group of patients yields (highly uncertain) estimates $\hat{e}_S^J$ and $\hat{e}_T^J$ of the effect sizes among those patients. However, our accurate estimate of μ_S gives us an accurate estimate of surrogate error term ϵ_S^J for the paired group, and a large $|\rho_I|$ gives us an accurate estimate of true outcome error term ϵ_T^J for the paired group, ultimately yielding an accurate estimate of μ_T despite having very few true outcome observations.

The distinction between ρ_0 and ρ_I will be a key feature of our analysis in Section 3. Although both capture the relationship between the two outcomes, they need not take similar values — in real clinical trial data, the magnitudes $|\rho_0|$ and $|\rho_I|$ are positively correlated but can vary significantly (see Figure 4 in Section 3.3 for a scatterplot of both correlations for different true and surrogate outcome pairs collected from 65 meta-analyses in the medical literature). Appendix A posits a causal model of the impact of drugs on disease progression that can be used to identify situations where we may expect large $|\rho_0|$ paired with small $|\rho_I|$ or vice versa. For instance, we may get large ρ_0 and small ρ_I when survival post-progression has a much larger variance than time-to-progression; conversely, we may get small ρ_0 and large ρ_I when time-to-progression has a much larger variance but much smaller mean than survival post-progression.

2.2. Clinical Trial Structure

Our central analytical objective 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 clinical trial designs. 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. The third, **Type C**, makes inference using both patient surrogate and true outcomes; this is our proposed design. 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 interim analyses.

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 distributions $\hat{\mathbf{e}}^{\mathbf{J}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_I/(p_J n))$ and $\hat{e}^N \sim \mathcal{N}(\mu_S, 4/(p_N n))$ for the effect size estimates among n patients to perform Bayesian updates. Details of the Bayesian updates are provided in Appendix B, and key notation is summarized in Table 2.1.

Outcome Parameters (Exogenous)	
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	Control-arm exponential distribution rate for the surrogate outcome
λ_T	Control-arm exponential distribution rate for the true outcome
Economic Parameters (Exogenous)	
α	Type I error control — the trial must reject the null hypothesis when it holds with $\leq \alpha$ probability
β, δ	Type II error control — the trial must have power $\geq 1-\beta$ to detect effect size magnitude δ
λ_E	Rate of patient enrollment into the trial
c_n	Monetary cost of enrolling one patient in a clinical trial
c_w	Monetary cost of waiting one additional unit of time
r	Discount rate applied to future costs
Trial Design Parameters (Endogenously Selected by Designer Before Study)	
n	Target number of patients enrolled (uniformly at rate λ_E from times 0 through n/λ_E)
v_1	Effect size variance triggering an interim analysis
m_1	Effect size estimate magnitude that triggers us to reject the null hypothesis at the interim analysis
v_2	Effect size variance triggering a final analysis
m_2	Effect size estimate magnitude that triggers us to reject the null hypothesis at the final analysis
Algorithm Parameters (Constructed by Algorithm 1)	
$\hat{\boldsymbol{\mu}}_{1t}$	Estimate of $\boldsymbol{\mu}$ at interim analysis time for trial type $t \in \{A, B, C\}$
$\hat{\boldsymbol{\mu}}_{2t}$	Estimate of $\boldsymbol{\mu}$ at final analysis time for trial type $t \in \{A, B, C\}$

Table 2.1 Parameters in model of clinical trials with continuous outcomes

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 interim 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 interim analysis. This matches well with the reality of clinical trials today — 65% of cancer RCTs have no more than one interim analysis (Floriani et al. 2008). Our approach straightforwardly generalizes to trial designs with multiple interim analyses as well (see discussion in Section 5).

To operationalize early stopping in our trial designs, we follow a classical group sequential approach to interim analysis, assuming the trial enrolls n patients uniformly at random between times 0 and n/λ_E (achieving enrollment rate λ_E) and that the interim and final analyses occur once the posterior variance of the trial's effect size estimate for the true outcome reaches thresholds v_1 and v_2, respectively. For trial type A (B), this is identical to performing these analyses once a

fixed number of true outcome events q_T (surrogate outcome events q_S) have been observed across all patients in the trial, which is a standard approach to prescribing analysis times in the literature. For a type C trial, the posterior variance of the effect size estimate is a function of both q_S and q_T (see Appendix B for details). Note that q_S and q_T can both be observed without unblinding trial participants. The trial designer selects n, v_1, and v_2 before the study commences, along with the effect size posterior mean magnitudes m_1 and m_2 that will trigger the trial to reject the null hypothesis of $\mu_T = 0$ at the interim and final analysis, respectively. Algorithm 1 formalizes the trial design.

Algorithm 1 Bayesian clinical trial design

Input: Trial type $t \in \{A, B, C\}$; outcome parameters $\sigma_{0S}^2,\ \sigma_{0T}^2,\ \rho_0,$ and ρ_I; trial design parameters $n,\ v_1,\ v_2,\ m_1$, and m_2
Initialize: $i = 1$, $V_t = \sigma_{0T}^2$, $q_S = 0$, $q_T = 0$
while $V_t > v_1$ **do**
  $event \leftarrow$ Next event to occur
  if $event ==$ Patient arrival **then**
    Enroll patient i, randomly allocating them to treatment or control with equal probability
    $i \leftarrow i + 1$
  else if $event ==$ Surrogate outcome observed **then**
    $q_S \leftarrow q_S + 1$, update posterior variance V_t based on q_S, q_T following Appendix B.4
  else if $event ==$ True outcome observed **then**
    $q_T \leftarrow q_T + 1$, update posterior variance V_t based on q_S, q_T following Appendix B.4
  end if
end while
$\hat{\boldsymbol{\mu}}_{1t} \leftarrow$ Posterior mean given observations $\hat{\mathbf{e}}^{\mathbf{J}}$ and $\hat{e}^N$, for trial type t (see Appendix B)
if $|\hat{\mu}_{1tT}| > m_1$ **then**
  return Reject null hypothesis and terminate trial at interim analysis
end if
while $V_t > v_2$ **do**
  $event \leftarrow$ Next event to occur
  if $event ==$ Patient arrival **then**
    Enroll patient i, randomly allocating them to treatment or control with equal probability
    $i \leftarrow i + 1$
  else if $event ==$ Surrogate outcome observed **then**
    $q_S \leftarrow q_S + 1$, update posterior variance V_t based on q_S, q_T following Appendix B.4
  else if $event ==$ True outcome observed **then**
    $q_T \leftarrow q_T + 1$, update posterior variance V_t based on q_S, q_T following Appendix B.4
  end if
end while
$\hat{\boldsymbol{\mu}}_{2t} \leftarrow$ Posterior mean given observations $\hat{\mathbf{e}}^{\mathbf{J}}$ and $\hat{e}^N$, for trial type t (see Appendix B)
if $|\hat{\mu}_{2tT}| > m_2$ **then**
  return Reject null hypothesis and terminate trial at final analysis
else
  return Fail to reject null hypothesis and terminate trial at final analysis
end if

A multitude of other designs have been proposed in the literature, including designs with a flexible number and timing of interim analyses (Lan and DeMets 1983, Wang and Tsiatis 1987) and designs employing continuous monitoring (Chick et al. 2017, 2018). Although these advanced designs are attractive due to their ability to improve trial outcomes, we use the classical group sequential design 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 maintaining double-blinding; see *e.g.*, Sydes et al. 2004, Tharmanathan et al. 2008).

2.3. Type I/II Error Control

The overall goal of all three trial designs is to identify whether the rate of true outcome events significantly differs between the control and treatment arms. To do so, the trial rejects or fails to reject the null hypothesis that the treatment has no impact on the true outcome ($\mu_T = 0$). To ensure comparability between the three trial types, we require that each achieve the same quality of inference about the true outcome of interest, using the ubiquitous clinical trial concept of Type I/II error control. To control the Type I error rate, we require the trial to not reject the null hypothesis more than α proportion of the time when the null hypothesis holds: $\mu_T = 0$, which implies via the study-level prior distribution that $\mu_S \sim \mathcal{N}(0, (1-\rho_0^2)\sigma_{0S}^2)$. To control the Type II error rate, we require that the trial have the power to reject the null hypothesis at least $1-\beta$ proportion of the time with true outcome effect size magnitude δ: $\mu_T = \pm\delta$, which implies that $\mu_S \sim \mathcal{N}(\pm\sigma_{0S}\rho_0\delta/\sigma_{0T}, (1-\rho_0^2)\sigma_{0S}^2)$.

Type I/II error control requires the trial designs to collect a sufficient amount of data to confidently differentiate between the $\mu_T = 0$ and $\mu_T = \pm\delta$ cases. Since the posterior variance of the effect size estimate indicates the confidence of a trial in that estimate (a small variance means more confidence), Type I/II error control corresponds to requiring sufficiently small posterior variances v_1 and v_2, coupled with appropriate thresholds m_1 and m_2, in the trial design.

Although Type I/II error control with interim analyses is well studied in the literature (see, *e.g.*, Demets and Lan (1994)), for completeness we derive the constraints on v_1, v_2, m_1, and m_2 that are imposed by Type I/II error control in Appendix C. In the simple scenario of no interim analysis (equivalent to $m_1 = \infty$), the single optimal v_2 can be easily obtained by solving an equation in one variable:

$$\xi(v_2) = \Phi(\Phi^{-1}(1-\alpha/2) - \delta\sqrt{1/v_2 - 1/\sigma_{0T}^2}) - \Phi(\Phi^{-1}(\alpha/2) - \delta\sqrt{1/v_2 - 1/\sigma_{0T}^2}) = \beta, \qquad (1)$$

and then setting

$$m_2 = \Phi^{-1}(1-\alpha/2)\sqrt{v_2(1 - v_2/\sigma_{0T}^2)}.$$

By selecting very large target patient enrollment counts n, type A and C trials can achieve arbitrarily small posterior variances if run for a sufficiently long time, so these designs can provide Type I/II error control for any selected α, β, and δ. However, because type B trials only rely on the surrogate outcome, their best achievable posterior variance is only $(1-\rho_0^2)\sigma_{0T}^2$. This implies that the type B trial design will not be able to obtain the desired error control for some particularly stringent values of α, β, and δ.

When an interim analysis is added, many different sets of v_1, v_2, m_1, and m_2 can satisfy the Type I/II requirements. We will proceed to select the most attractive set of these parameters based on how they compare in terms of expected trial costs.

While all of the results in the main body of the paper are derived under an expected cost objective subject to constraints on the Type I and II error rates, the key ideas of our trial design could be applied to other commonly used objective functions. For instance, it would be straightforward to instead focus on a Health Technology Assessment objective, which integrates 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). Further, in Appendix E we adapt our trial design to the expected-success framework like that described by Berry (1972), and explore the effect of our design under this framework.

2.4. Trial Costs

Subject to achieving the desired Type I and II error rates, all trial design parameters are chosen to minimize the expected discounted cost, where c_n is the cost of enrolling an additional patient, c_w is the cost of waiting an additional unit of time for a trial decision, and r is the discount rate (see Table 2.1). Let P and W be the random variables indicating the number of patients enrolled and the waiting cost respectively. Then, the expected discounted cost is

$$c_n\lambda_E\mathbb{E}\left[\frac{1-e^{-rP/\lambda_E}}{r}\right]+c_w\mathbb{E}\left[\frac{1-e^{-rW}}{r}\right].$$

The cost of a trial depends centrally on the speed at which patients experience surrogate and true outcome events — faster event rates correspond to smaller waiting costs as well as potentially smaller patient enrollment costs. To estimate these costs, we will assume that patients have exponentially distributed surrogate and true outcome times. Importantly, these assumptions are only used for cost estimation in our trial design — they are not used in the inference and will therefore have no impact on the trial's Type I/II error control.

Assumption 2. *When estimating trial costs, we assume that surrogate event times in the control arm are exponentially distributed with rate λ_S, and that true outcome event times in the control arm are exponentially distributed with rate λ_T.*

Given that the control therapy will typically have been previously tested in a similar patient population, we consider it reasonable to assume that decision makers would have good estimates of λ_S and λ_T when designing the study. Under the proportional hazards assumption and true effect sizes $\boldsymbol{\mu} = [\mu_S \ \mu_T]'$, the above assumption implies that the treatment arm's surrogate and true outcome event times are exponentially distributed with rates $\lambda_S e^{\mu_S}$ and $\lambda_T e^{\mu_T}$. By taking expectations over the (unknown) effect size vector $\boldsymbol{\mu}$, it is straightforward to compute the expected waiting cost incurred between the start of the study and the interim analysis, W_1^t, and between the interim and final analysis, W_2^t. Similarly, we can obtain expected patient enrollment costs P_1^t and P_2^t. Details of this derivation as well as high-quality approximations that are more analytically and computationally tractable are provided in Appendix C.

2.5. Optimal 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. Our optimal trial design selects patient enrollment n, interim analysis target posterior variance v_1, final analysis target posterior variance v_2, interim analysis threshold m_1, and final analysis threshold m_2 such that the Type I/II error requirements are satisfied at minimum expected cost.

$$\begin{aligned} \min_{n,v_1,m_1,v_2,m_2} \ & W_1^t(n,v_1,m_1) + W_2^t(n,v_1,m_1,v_2,m_2) + P_1^t(n,v_1,m_1) + P_2^t(n,v_1,m_1,v_2,m_2) \\ \text{s.t.} \ & \alpha(v_1,v_2,m_1,m_2) \leq \alpha \\ & \beta(v_1,v_2,m_1,m_2) \leq \beta \end{aligned}$$

Functions W_1^t, W_2^t, P_1^t, and P_2^t are derived in Appendix C.2 and capture the expected cost of the trial design. Functions $\alpha(\cdot)$ and $\beta(\cdot)$ are derived in Appendix C.1 and are used in the formulation for Type I/II error control. As discussed in Appendix C.1, this optimization model can be simplified to a 3-variable search over n, v_1, and m_1, because v_1 and m_1 imply a unique v_2 and m_2 that satisfy the Type I/II constraints at minimal cost. The resulting 3-variable model can be solved to acceptable accuracy using standard nonlinear optimization techniques.

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 cost of each of the three trial designs (subject to Type I/II error constraints) varies 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.

The full trial design does not yield tractable comparative statics, so we consider a simplified setting without an interim analysis.[2] Although trial designs without interim analyses preclude early stopping within a trial, for a given patient enrollment, Type C trials will stop significantly faster than Type A or B trials, because they require less waiting time to achieve the target variance needed for Type I/II error control. Indeed, numerical results on the full trial design match the resulting comparative statics closely (see Section 3.3), suggesting that the benefits of an interim analysis are somewhat orthogonal to the benefits of combining surrogate and true outcomes. We further impose that our final analysis occurs after all n_t^* participants have been enrolled, *i.e.,* we do not conclude the trial before the enrollment target has been reached. Lastly, we restrict our exposition to positive correlations ($\rho_0, \rho_I > 0$) and set the discount factor $r = 0$; this is for algebraic convenience, and it is straightforward to verify that our results remain qualitatively similar otherwise.

Parameters of Interest: We are interested in comparative statics with respect to the study-level and individual-level correlations ρ_0 and ρ_I; the variances of surrogate and true outcome effect sizes σ_{0S}^2 and σ_{0T}^2 across historical clinical trials; the surrogate and true outcome arrival rates in the control group, λ_S and λ_T; and the patient recruitment and waiting costs c_n and c_w.

To evaluate comparative statics, we invoke the envelope theorem on the constrained optimization problem given in Section 2.5 with the above simplifications (proofs are provided in Appendix D).

LEMMA 1 **(Type A Trials)**. *Trials that perform inference on only true outcomes satisfy:*

1. *The cost does not depend on any of the surrogate properties, including the individual-level correlation ρ_I, study-level correlation ρ_0, the prior variance of surrogate and true outcome effect sizes σ_{0S}^2 and σ_{0T}^2, or the surrogate outcome arrival rate in the control arm λ_S:*

$$\frac{dCost_A}{d\rho_I} = \frac{dCost_A}{d\rho_0} = \frac{dCost_A}{d\sigma_{0S}} = \frac{dCost_A}{d\sigma_{0T}} = \frac{dCost_A}{d\lambda_S} = 0\,.$$

2. *The cost is monotonically increasing in the trial recruitment and waiting costs c_n and c_w:*

$$\frac{dCost_A}{dc_n} > 0 \quad \textit{and} \quad \frac{dCost_A}{dc_w} > 0\,.$$

3. *The cost is monotonically decreasing in the true outcome arrival rate in the control arm λ_T:*

$$\frac{dCost_A}{d\lambda_T} < 0\,.$$

[2] This setting is practically relevant, since 43% of cancer RCTs have no interim analyses (Floriani et al. 2008).

These relationships are to be expected. Since Type A trials do not utilize surrogate outcomes, surrogate properties do not play any role in determining the cost of the trial. The Type I/II constraints require the same number of true outcome events to be observed regardless of prior distribution, so σ_{0T}^2 does not play any role in the cost of the trial. A larger λ_T implies that we see patient outcomes faster, thereby decreasing trial costs since we can shorten trials while maintaining the target statistical power.

LEMMA 2 **(Type B Trials)**. *Trials that perform inference on only surrogate outcomes satisfy:*

1. *The cost does not depend on the individual-level correlation ρ_I or on the true outcome arrival rate in the control arm λ_T:*

$$\frac{dCost_B}{d\rho_I} = \frac{dCost_B}{d\lambda_T} = 0\,.$$

2. *The cost is monotonically increasing in the prior variance of the true outcome effect size σ_{0T}^2, as well as the trial recruitment and waiting costs c_n and c_w:*

$$\frac{dCost_B}{d\sigma_{0T}} > 0\,, \quad \frac{dCost_B}{dc_n} > 0 \quad \textit{and} \quad \frac{dCost_B}{dc_w} > 0\,.$$

3. *The cost is monotonically decreasing in the study-level correlation ρ_0, the prior variance of the surrogate effect size σ_{0S}^2, and the surrogate outcome arrival rate in the control arm λ_S:*

$$\frac{dCost_B}{d\rho_0} < 0\,, \quad \frac{dCost_B}{d\sigma_{0S}} < 0 \quad \textit{and} \quad \frac{dCost_B}{d\lambda_S} < 0\,.$$

Since Type B trials do not directly utilize true outcomes, properties like λ_T and ρ_I do not play any role in determining the cost of the trial. A larger surrogate prior variance σ_{0S}^2 means fewer surrogate events must be observed to learn if μ_S is high, medium, or low, which in turn speeds learning if μ_T is high, medium, or low and therefore decreases trial costs. Recalling that $\mu_S|\{\mu_T = x\} \sim \mathcal{N}(\rho_0\sigma_{0S}x/\sigma_{0T}, (1-\rho_0^2)\sigma_{0S}^2)$, we read that a small true outcome prior variance σ_{0T}^2 will yield a large difference between the expected surrogate effect size μ_S under the null hypothesis ($\mu_T = 0$) versus $\mu_T = \pm\delta$, decreasing the effort needed to differentiate these two cases and therefore the overall costs. Larger ρ_0 implies that the surrogate is more predictive of the true outcome, decreasing the number of surrogate events that must be observed to accurately estimate the true outcome effect size and therefore lowering costs. Finally larger λ_S decreases trial costs (similar to λ_T in Type A trials).

LEMMA 3 **(Type C Trials)**. *Trials that perform inference on both outcomes satisfy:*

1. *The cost is monotonically increasing in the trial recruitment and waiting costs c_n and c_w:*

$$\frac{dCost_C}{dc_n} > 0 \quad \textit{and} \quad \frac{dCost_C}{dc_w} > 0\,.$$

2. *The cost is monotonically decreasing in the surrogate and true outcome arrival rates in the control arm λ_S and λ_T:*

$$\frac{dCost_C}{d\lambda_S} < 0 \quad \textit{and} \quad \frac{dCost_C}{d\lambda_T} < 0\,.$$

3. *The cost is non-monotonic in the individual-level correlation ρ_I, study-level correlation ρ_0, as well as the variance of the surrogate and true outcome effect sizes σ_{0S} and σ_{0T}:*

$$\frac{dCost_C}{d\rho_I} \gtreqless 0\,, \quad \frac{dCost_C}{d\rho_0} \gtreqless 0\,, \quad \frac{dCost_C}{d\sigma_{0S}} \gtreqless 0 \quad \textit{and} \quad \frac{dCost_C}{d\sigma_{0T}} \gtreqless 0\,.$$

Unsurprisingly, as with Type A and B trials, larger λ_S and λ_T decrease trial costs. However, a number of comparative statics are surprising. For example, intuitively, a stronger study-level correlation ρ_0 and a stronger individual-level correlation ρ_I would lead to a more informative surrogate — thus, *a priori* one might assume that the cost of a Type C trial would be monotonically decreasing in ρ_0 and ρ_I. Instead, we see that the cost of the trial is non-monotonic in these quantities; we examine this effect in the next subsection.

3.2. Absolute Benefits of Type C Trials

We find all the non-monotone relationships identified in Lemma 3 to be unexpected and relevant. Since Type C trials are especially complex, we will interpret these comparative statics in a regime where the patient enrollment n_C^* is large relative to the other trial parameters.

Specifically, as detailed in Appendix D.2, 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_C^*. This rules out very informative study-level priors (*i.e.*, σ_{0S}, σ_{0T} are very small), which matches practice since the FDA favors uninformative/objective priors to avoid "assuming the result" (see, *e.g.*, Lee and Chu 2012). It also rules out very high study-level correlations (since $\lambda_{\max}(\mathbf{\Sigma}_0^{-1}) \to \infty$ when $|\rho_0| \to 1$), which is an uninteresting regime since decision-makers can already safely rely on Type B trials.

We find that $d\text{Cost}_C/d\sigma_{0T} = 0$ to highest order in n_C^*, *i.e.,* the non-monotonic dependence on σ_{0T} that we found in Lemma 3 is not salient in practice. However, our cost remains non-monotonic in ρ_0, ρ_I and σ_{0S}. We can thus evaluate the sets of exogenous parameters for which $d\text{Cost}_C/dp = 0$ for each of $p \in \{\rho_0, \rho_I, \sigma_{0S}\}$. We find that the surface on which $d\text{Cost}_C/d\rho_0 = 0$ coincides with the surface on which $d\text{Cost}_C/d\sigma_{0S} = 0$. This is because, as discussed in Appendix D.2, these two quantities contribute to the same effect — increasing ρ_0 and decreasing σ_{0S} both increase the weight of the study-level Bayesian prior $\mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$ in our effect size estimates. We observe a trade-off between an informative study-level link between our surrogate and true outcome effect size estimates, and an informative link between paired surrogate and true outcomes at the patient-level. To illustrate this tension, we focus our discussion on ρ_0 and ρ_I for the remainder of this section.

Figure 2 shows that we obtain three regions with different behaviors of comparative statics. The dashed line between Regions 1 and 2 depicts parameter pairs where $d\text{Cost}_C/d\rho_0 = 0$, while the dashed line between Regions 2 and 3 depicts parameter pairs where $d\text{Cost}_C/d\rho_I = 0$. Region 2 captures the intuitive regime, where the cost of a Type C trial improves (decreases) from increasing both correlations ρ_0 and ρ_I. In contrast, Regions 1 and 3 capture regimes where increasing one type of correlation may in fact be detrimental. Specifically, in Region 1, the cost of a Type C trial improves from decreasing ρ_0 and increasing ρ_I; in Region 3, the cost of a Type C trial improves from decreasing ρ_I and increasing ρ_0.

It is worth noting that there are always 3 local minima (for the cost of the trial) at the corners $(\rho_0 = 1, \rho_I = 0)$, $(\rho_0 = 0, \rho_I = 1)$ and $(\rho_0 = 1, \rho_I = 1)$; there is at most one local maxima and no local minima in the interior (see Appendix D.3).

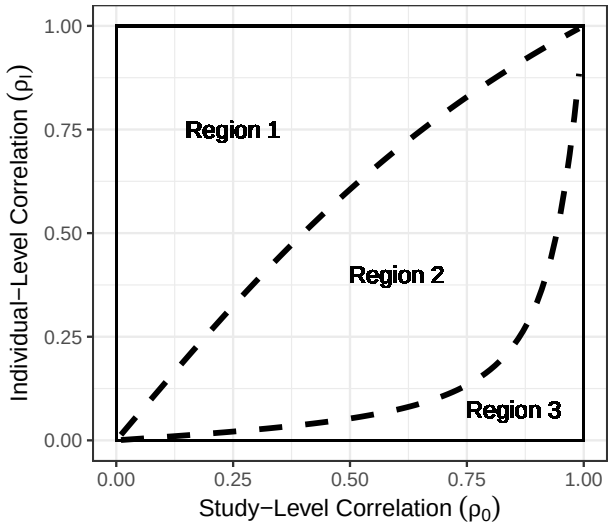


Figure 2 Regions with different relationships between the cost of a Type C trial and the correlations ρ_0, ρ_I.

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 $\rho_0 = 1$, which implies that 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 ρ_I can increase trial costs by removing a source of statistical independence when ρ_0 is high. A similar argument holds vice-versa: when $\rho_I = 1$, increasing ρ_0 can increase trial costs. This trade-off is not present in Region

2, where increasing both ρ_0 and ρ_I is cost-saving; in this regime, the benefit of additional statistical independence is smaller than the benefit of a more predictive surrogate (see Remark 1 in Section 2.1). Thus, the traditional wisdom that more predictive surrogates are always more valuable is not true from a statistical viewpoint. Appendix A posits a disease progression model that can be used to identify situations where we may expect large $|\rho_0|$ paired with small $|\rho_I|$ or vice versa.

It may also be useful to reason about when Type B trials produce true outcome effect size estimates that are more accurate than (have lower posterior variance than) Type A trials in the regime where patient enrollment n is large; these are regions where Type B trials will be preferred to Type A trials due to needing lower cost to reach the target Type I/II error control. Specifically, for a trial with n patients and length w, Type B trials are preferable to Type A trials when

$$\rho_0^2(1 - 4/(\sigma_{0S}^2 \mathbb{E}[q_S|w])) > 1 - 4/(\sigma_{0T}^2 \mathbb{E}[q_T|w])\,.$$

Here, $\mathbb{E}[q_S|w]$ and $\mathbb{E}[q_T|w]$ indicate the expected number of surrogate and true outcome events observed throughout the trial's length w. As expected, Type B trials are preferable to Type A trials when the surrogate is highly predictive (high ρ_0), and the surrogate requires relatively fewer samples to estimate accurately than the true outcome (high σ_{0S}/σ_{0T}).

3.3. Relative Benefits of Type C Trials

By construction, Type C trials always reduce costs relative to Type A and B trials, subject to the same Type I/II constraints. We now use the lemmas from Section 3.1 to characterize parameter regimes where Type C trials are particularly advantageous. For ease of comparison, we fix the patient enrollment to be n across all three trial types, and optimize only over the length of the trial w_t^* (note that this implies $w_C^* < w_A^*, w_B^*$). Once again, we consider the large n regime.

THEOREM 1 **(When to Bother?)**. *The cost improvement relative to Type A trials $Cost_A - Cost_C$ is increasing in c_w; is non-monotone in ρ_0, ρ_I and σ_{0S}; and does not depend on σ_{0T}. The cost improvement relative to Type B trials $Cost_B - Cost_C$ is increasing in σ_{0T} and c_w; is decreasing in σ_{0S} and ρ_0; and it is non-monotone in ρ_I.*

The proof of Theorem 1 is provided in Appendix D.4. Theorem 1 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, there are 3 local minima within the (ρ_0, ρ_I) space where a Type C trial might have particularly high benefit — high ρ_I and low/moderate ρ_0 (the most promising part of Region 1), high ρ_0 and ρ_I (the most promising part of Region 2), and moderate/high ρ_0 and low ρ_I (the most promising part of Region 3). However, from Theorem 1, we know that Type C trials provide little incremental benefit over Type B trials when ρ_0 is large. As a result, we might conclude that two regions of the (ρ_0, ρ_I) space that are most likely to yield

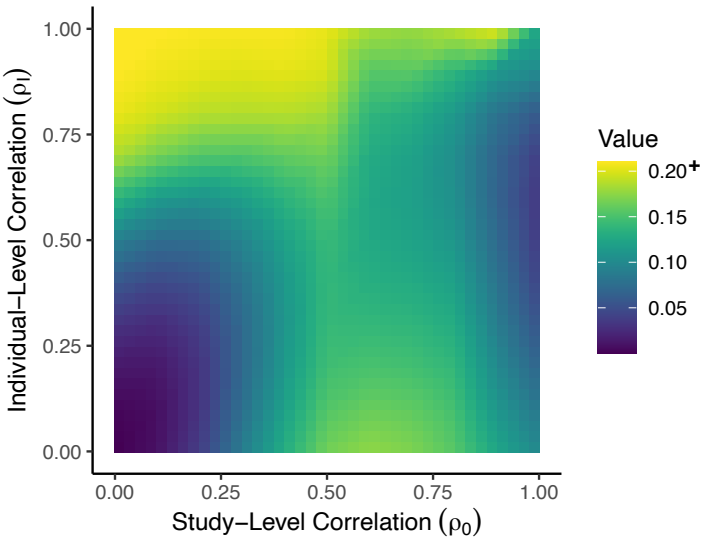


Figure 3 Fractional reduction in cost by Type C trials relative to the cost of the best standard design (Type A or B) as a function of the study-level correlation ρ_0 (x-axis) and the individual-level correlation ρ_I (y-axis). Extreme outliers at the boundaries producing values above 20% are capped at 21% for visual clarity.

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 cost reduction of a Type C trial compared to its best competitor,

$$\text{Value} = \frac{\min\{\text{Cost}_A, \text{Cost}_B\} - \text{Cost}_C}{\min\{\text{Cost}_A, \text{Cost}_B\}},$$

as a function of ρ_0 and ρ_I, averaged over a grid of exogenous parameter values (see Appendix D.5 for details). Importantly, unlike our comparative statics above, these results do not use the various simplifications we assumed in Section 3.1, include an interim analysis, and follow the optimal designs prescribed in Section 2.5. 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/moderate. 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. Appendix A characterizes the types of diseases for which we may expect moderate/large ρ_0 and low ρ_I, and vice-versa.

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

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 (see Appendix D.5 for details). Historically, only pairs with very high ρ_0 values would be considered viable surrogates for use as the primary endpoint, *e.g.*, the Institute for Quality and Efficiency in Health Care suggests a cutoff of $\rho_0 > 0.85$ (Kemp and Prasad 2017). 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.

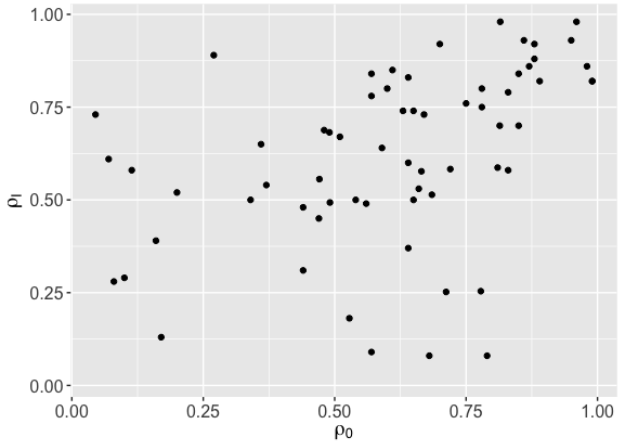


Figure 4 Scatterplot of study-level correlation ρ_0 (x-axis) and individual-level correlation ρ_I (y-axis) for different true and surrogate outcome pairs collected from 65 meta-analyses in the medical literature.

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. Specifically, we find that treatments for colorectal cancer and gastro-oesophogeal cancer, as well as immunotherapy treatments across a range of cancers, fall into the region with low ρ_0 and high ρ_I. Meanwhile, treatments for pancreatic cancer, renal cell carcinoma, lung cancer, colorectal cancer liver metastases, and advanced gastric cancer, as well as treatment with immune checkpoint inhibitors across a range of cancers, fall into the region with moderate ρ_0 and low ρ_I. These are very prevalent diseases. For example, colorectal cancer alone is the third most commonly diagnosed cancer in the United States (American Cancer Society 2020), and the American Cancer Society estimates that there will be over 57,000 new cases of pancreatic cancer, 73,000 new cases of renal cell carcinoma, and 228,000 new cases of lung cancer in the United States in 2020 (American Cancer Society 2020). Thus, our proposed design has the potential to offer significant benefit to many patients by speeding the regulatory approval of effective new treatments.

As captured in Figure 3 and in our case study of metastatic breast cancer in Section 4, our design has the potential to deliver 10% or higher cost savings even for less appealing regions with moderate ρ_0 and ρ_I. Given the large costs of Phase III trials, these represent significant benefits, as well.

4. Simulation Based on Large-Scale Clinical Trial Database

While the clinical trial model in Section 2 was built to resemble real-world trials, we still make some assumptions that may be violated in practice. As described in Section 2.1 and Appendix F, these models assume study effect sizes are drawn from a bivariate normal prior and satisfy the proportional hazards assumption, and that key parameters such as ρ_0, ρ_I, σ_{0S}^2, and σ_{0T}^2 are known with certainty. Furthermore, our cost calculations assume that the event times are exponentially distributed, with control arm rates λ_S and λ_T that are known with certainty. However, these assumptions may not always hold. It is therefore important to confirm that these designs still provide cost benefits and achieve the specified Type I/II error control even when 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 or death).

In the remainder of this section, we use a large-scale database of MBC clinical trial results and perform an additional literature review to estimate the parameters needed for our 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 standard trial designs; as described in Section 4.2, realistic simulation of individual patient outcomes required a significant data collection effort from the MBC clinical trial literature.

4.1. Data Collection and Parameter Estimation

As summarized in Table 2.1, the trial design takes as input six exogenous outcome parameters (σ_{0S}^2, σ_{0T}^2, ρ_0, ρ_I, λ_S, and λ_T), one exogenous trial recruitment parameter (λ_E), three exogenous Type I/II error control parameters (α, β, and δ), and three exogenous economic parameters (c_n, c_w, and r). We additionally have five endogenous trial design inputs (n, v_1, v_2, m_1, and m_2), which are computed by solving the optimization problem from Section 2.5. We now describe how each parameter was chosen to obtain the final trial design.

We chose the individual-level correlation ρ_I based on the individual patient data meta-analysis of Burzykowski et al. (2008), who report the rank correlation coefficient between PFS and OS to be 0.688 with 95% confidence interval $[0.686, 0.690]$. All remaining outcome parameters are estimated using a repository of 1,865 studies of MBC drug therapies collected by Silberholz et al. (2019),

	Average	Range
Publication year	2005	[1984, 2017]
Number of patients	319.7	[51, 1286]
Proportion female	1.00	[0.99, 1.00]
Median age	56.5	[47.5, 71.7]
Mean ECOG performance status	0.59	[0.21, 1.12]
Proportion with visceral disease	0.63	[0.18, 0.87]
Median OS (months)	21.9	[9.0, 45.9]
Log(OS hazard ratio)	-0.05	[-0.80, 1.00]
Median PFS (months)	7.4	[2.2, 17.9]
Log(PFS hazard ratio)	-0.14	[-0.85, 1.24]

Table 4.1 Aggregate patient characteristics of the 93 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.

which is publicly available at `http://www.cancertrials.info`. To calculate our control group surrogate and true outcome arrival rates λ_S and λ_T, we used the median PFS and OS of 7.2 months and 21.3 months respectively across all Phase III studies in this repository.[3] Since an exponential distribution with rate λ has median $(\ln 2)/\lambda$, we infer the surrogate and true outcome exponential distribution event rates to be $\lambda_S = (\ln 2)/7.2 \approx 0.096$ and $\lambda_T = (\ln 2)/21.3 \approx 0.032$.

We perform our trial simulations on a subset of 93 studies from this repository that are two-arm RCTs (one treatment and one control) that report surrogate and true outcome effect size estimates as well as Kaplan-Meier curves for both outcomes. We manually extract the 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 93 RCTs.

As described in more detail in Section 4.2, our simulation approach assumes that the effect size observed in each of the 93 studies is the true effect size (in practice, of course, these studies are finite-sized and the effect size was observed with some sampling error). We fit the prior under this assumption. For study $i \in \{1, \ldots, 93\}$, define $\hat{\mathbf{e}}_i$ to be the surrogate and true outcome effect sizes read from the Kaplan-Meier curves. Assuming each study's effect size estimate is independently drawn from the prior $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_0})$ with no sampling error, the log likelihood function over $\boldsymbol{\Sigma}_0$ is:

$$LL(\boldsymbol{\Sigma}_0) \propto \sum_{i=1}^{93} -0.5(\log\det(\boldsymbol{\Sigma}_0) + (\hat{\mathbf{e}}_i)'(\boldsymbol{\Sigma}_0)^{-1}(\hat{\mathbf{e}}_i)) - log(2\pi)\,.$$

Maximizing the log likelihood using Nelder-Mead simplex yields parameter values $\sigma_{0S} = 0.326$, $\sigma_{0T} = 0.255$, and $\rho_0 = 0.737$. Figure 5 plots the study effect size estimates along with the fitted

[3] Note that these values differ slightly from the values in Table 4.1, since the former is computed over all studies in the repository while the latter looks at the median values across the subset of 93 Phase III trials used in the simulation.

95% confidence ellipse for study true effect sizes $\boldsymbol{\mu}$. Our assumption of a bivariate normal distribution visually appears reasonable, and indeed, the effect sizes fail to reject the null hypothesis of multivariate normality under the Henze-Zirkler test ($p = 0.145$).

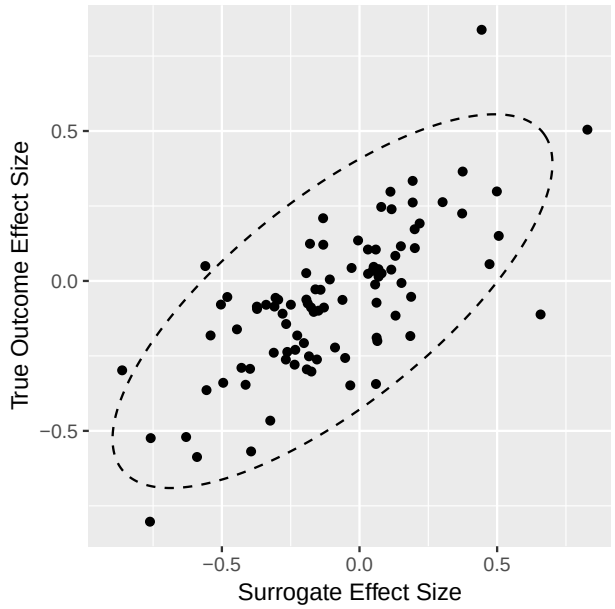


Figure 5 Effect size estimates of the 93 MBC RCTs from Silberholz et al. (2019) used to parameterize the prior, along with the 95% confidence ellipse of the true study effect sizes from the fitted prior.

We estimate the MBC economic and error-control parameters from the perspective of a drug company; we discuss how other decision-makers might make different choices in selecting these parameters in Section 5. We use error control parameters $\alpha = 0.05$, $\beta = 0.2$, and $\delta = 0.5$; these were the most commonly reported values in our 93 studies. We also need to 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 $\lambda_E = 8$. We set c_p based on published estimates that per-patient costs average \$50,000 in Phase III trials (Moore et al. 2018). The parameter c_w captures the expected incremental cost to the drug company of delaying an approval decision by a month. There is no cost if the drug is ultimately rejected for inefficacy, and otherwise the cost is equivalent to missing out on a month of the expected revenue from the drug. Using published information on revenues for oncology drugs (Prasad and Mailankody 2017), we estimate that the average monthly revenue for a recently FDA approved drug is \$45 million. An estimated 45.2% of Phase III oncology trials yield approvals (Hay et al. 2014); therefore, we estimate $c_w = 20,340,000$. Additionally, we use a time-discounting parameter of $r = 0.08/12$ per month, based on analyses in Smith and Gravelle

Param.	Value	Source
σ_{0S}	0.326	Estimated from data in Silberholz et al. (2019)
σ_{0T}	0.255	"
ρ_0	0.737	"
λ_S	0.096	"
λ_T	0.032	"
ρ_I	0.688	Burzykowski et al. (2008)
α	0.05	Standard choice in historical clinical trial data
β	0.2	"
δ	0.5	"
λ_E	8	Silberholz et al. (2019)
c_w	2.034×10^7	Prasad and Mailankody (2017) and Hay et al. (2014)
c_p	5.0×10^4	Moore et al. (2018)
r	0.0067	Spellberg et al. (2012) and Smith and Gravelle (2001)
n^A	336	Optimized based on other parameters (see Section 2.5)
v_1^A	2.54×10^{-2}	"
v_2^A	2.14×10^{-2}	"
m_1^A	0.378	"
m_2^A	0.235	"
n^C	441	"
v_1^C	2.54×10^{-2}	"
v_2^C	2.14×10^{-2}	"
m_1^C	0.378	"
m_2^C	0.235	"

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

(2001) and Spellberg et al. (2012). Using these exogenous parameters, we optimized the endogenous parameters v_1^t, v_2^t, m_1^t, m_2^t, and n^t for each trial type $t \in \{A, B, C\}$, as detailed in Section 2.

Table 4.2 summarizes the full set of parameter values used in the MBC trial design, along with their sources.[4] Note that a Type B trial that performs inference on only surrogate outcomes is infeasible in this context, since it cannot meet the Type I/II error constraint even with infinite patient enrollment. This is because the MBC surrogate of PFS is not sufficiently predictive of the true outcome OS in historical clinical trial data. This is in line with guidance from the Institute for Quality and Efficiency in Health Care, which classifies a surrogate as having proven validity only if $\rho_0 > 0.85$ (Kemp and Prasad 2017), while the MBC study-level correlation is only $\rho_0 = 0.737$.

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 93 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

[4] Note that although the target enrollment $n^A < n^C$ in Table 4.2, since Type C trials stop well before Type A trials, the *actual* patient enrollment for Type C trials is lower than that for Type A trials (as shown in Table 4.3).

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, following the trial procedure from Algorithm 1. When simulating, we assume each patient's enrollment time is drawn uniformly at random between 0 and n/λ_E. 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 $\rho_{exp,i}$ ($\rho_{ctl,i}$).

For each trial type t, we 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 percentage of trials which reject the null hypothesis $prop_t^{reject}$.

In order to verify Type I/II error control, we also need to calculate the true outcome effect size μ_T for each trial based on the reported Kaplan-Meier curves. However, these curves often end before they reach a y-axis value of 0, *i.e.*, we do not see the events of all patients in the trial. We address this by uniformly capping the maximum follow-up (*i.e.*, event times in either arm) at 40 months. We note that, under this approach, our baseline Type A trial has nearly perfect error control. We then simulate a very large trial ($n = 100,000$ and $\lambda_E^{ground} = n/40$) for 40 months to calculate the true outcome effect size estimates μ_T.

4.3. Simulation Results

On average, we found that the discounted cost of running a Type A trial was \$179 million, running a Type B trial was infeasible given the required Type I/II error control, and the discounted cost of running a Type C trial was \$150 million. Thus, a Type C trial resulted in a \$29.2 million (16.3%) reduction in trial costs relative to standard trial designs. Table 4.3 breaks down the costs by patient enrollment and waiting time for each feasible trial type.

Evidently, the cost reductions imbued by Type C trials compared to Type A trials arose from reduced waiting times (we save an average of \$27 million, or 16.5% of waiting costs). Furthermore, we are successfully able to leverage information from a moderately predictive surrogate that cannot be trusted as a primary endpoint alone (as evidenced by the infeasibility of the Type B trial design) to significantly speed up trial decisions.

	Trial Type A: True Outcome Only	Trial Type C: Combined Outcomes
Avg. Number of Patients Enrolled	294	250
Avg. Length of Trial (months)	38.3	31.2
Avg. Overall Cost	\$179M	\$150M

Table 4.3 **Breakdown of different components of the overall cost from our simulation results.**

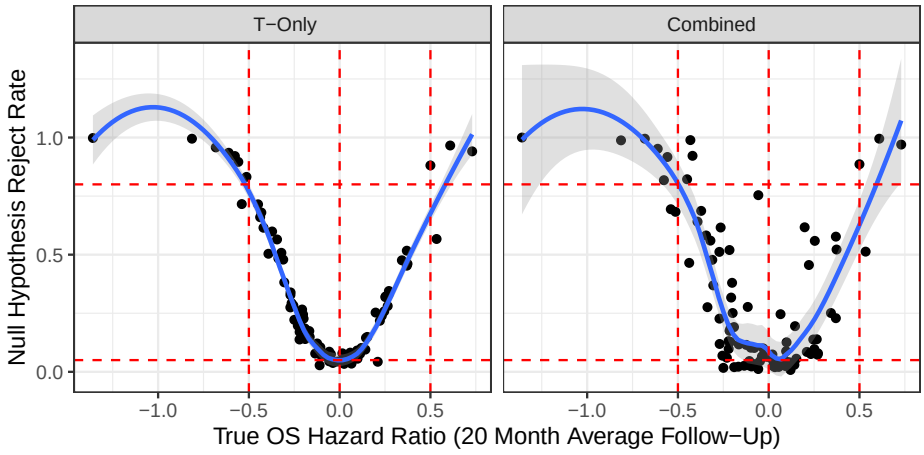


Figure 6 **Proportion of times the trial rejects the null hypothesis as a function of the "ground-truth" effect size for the 93 MBC RCTs from Silberholz et al. (2019), under Type A (left) and Type C (right) trial designs.**

Next, we check to ensure that we are properly controlling Type I/Type II error. This is particularly important for Type C trial designs, since we are relying on a number of modeling assumptions that could plausibly be misspecified. In Figure 6, we plot the proportion of times that a trial rejects the null hypothesis, as a function of the ground-truth true outcome effect size μ_T; points represent each of the 93 trials, and the blue curve represents LOESS smoothing. Further, vertical lines capture $\mu_T = 0$ and $\mu_T = \pm\delta$, while horizontal lines capture rejection probabilities of α and $1-\beta$. We see that both Type A and Type C trials meet the required Type I and Type II error rates on average. Both trial designs are slightly below the required power $1-\beta = 0.8$ to detect a true outcome effect size of 0.5 (*i.e.*, the new treatment is much worse than the control). However, we believe this is not a cause for concern, both because very few of our historical trials had $\mu_T \approx 0.5$ and also because establishing that a treatment is *worse* than the standard of care is not the primary objective of clinical trials in practice. Importantly, we note that despite the likely scenario that our model's assumptions are somewhat misspecified, our Type C trial design meets the target error control as evaluated on actual clinical trial data.

Figure 6 also shows more variability in the null hypothesis reject rate for individual studies under the Type C trial design as compared to the Type A trial design. For instance, the 8 trials with

$-0.6 \leq \mu_T \leq -0.4$ had reject rates in the range $[0.465, 0.989]$ for Type C trials, compared with $[0.615, 0.921]$ for Type A trials. This is to be expected, since trials with unexpectedly high (low) surrogate effect sizes μ_S will reject the null hypothesis at lower (higher) rates in a Type C design with $\mu_T \approx -0.5$, while μ_S plays no role in the Type A design. Indeed, Type C trials guarantee the desired reject rate on average across trials. This average (the blue curve in Figure 6) indicates the Type C design is performing as intended.

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 estimating how two treatments compare on the true outcome (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. 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 estimating parameters $\{\rho_0, \sigma^2_{0S}, \sigma^2_{0T}\}$ from data collected in the presence of publication bias would actually lead to a *conservative* trial design due to an underestimated value of $|\rho_0|$.

Another practical concern we have not explicitly modeled is patient dropouts during the course of a clinical trial. Dropouts occur when patients elect not to finish a full course of a drug therapy (*e.g.*, due to significant toxicity) and are relatively common in cancer clinical trials. Consequently, modern

clinical trials perform intention-to-treat (ITT) analyses (Fisher 1990), which include all patients randomized to an arm, regardless of their adherence or subsequent withdrawal from treatment. To more accurately capture dropouts, our approach could be extended to model time-to-event outcomes as being drawn from a mixture model, with one group for patients who stopped early and another for those who completed the per-protocol treatment. Dropouts are likely to affect our Type A, B, and C designs uniformly, so our qualitative results on "when to bother?" as well as our cost benefits relative to existing trial designs would likely remain similar under this model extension.

Our proposed clinical trial model and resulting design could be extended in a number of ways, which we view as promising future research directions. First, while we focus on time-to-event outcomes, our design can be straightforwardly extended to many other types of outcomes, such as continuous, binary or categorical outcomes. The effect sizes considered could be correspondingly expanded (*e.g.*, odds ratios or risk ratios for binary outcomes), and a fixed delay between the surrogate and true outcome measurements would need to be added. Second, we minimized a cost objective subject to a Type I/II error constraint, but one could consider other objectives such as a value-based Health Technology Assessment objective that balances trial costs and population benefit (Brennan et al. 2006, Hampson and Jennison 2013, NICE 2018). Third, our approach can be numerically extended to allow for multiple interim analyses or continuous monitoring. For multiple interim analyses, one could easily expand the optimization problem in Section 2 to optimize over the target variances and thresholds for each additional interim analysis. The resulting Type I/II error constraints can be addressed through techniques in the alpha-spending literature (O'Brien and Fleming 1979, Demets and Lan 1994). A continuous monitoring approach would require solving for a decision boundary in an optimal stopping problem (see, *e.g.*, the framework in Chick et al. 2017).

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. This research was supported in part by the Wharton Dean's Research Fund.

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.

American Cancer Society. 2020. Cancer facts and figures 2020. Online. URL `https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf`.

Athey, Susan, Raj Chetty, Guido W Imbens, Hyunseung Kang. 2019. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Tech. rep., National Bureau of Economic Research.

Bastani, Hamsa. 2020. Predicting with proxies: Transfer learning in high dimension. *Management Science* .

Berry, Donald A. 1972. A bernoulli two-armed bandit. *The Annals of Mathematical Statistics* 871–897.

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 .

Cox, D.R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B* **34**(2) 187–220.

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

Demets, David L, KK Gordon Lan. 1994. Interim analysis: the alpha spending function approach. *Statistics in medicine* **13**(13-14) 1341–1352.

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`.

Fisher, Lloyd D. 1990. Intention to treat in clinical trials. *Statistical issues in drug research and development* .

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.

Hay, Michael, David W Thomas, John L Craighead, Celia Economides, Jesse Rosenthal. 2014. Clinical development success rates for investigational drugs. *Nature biotechnology* **32**(1) 40–51.

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

Kemp, Robert, Vinay Prasad. 2017. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC medicine* **15**(1) 134.

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, 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.

Machin, David, Yin Bun Cheung, Mahesh KB Parmar. 2006. *Survival Analysis: A Practical Approach*, chap. 3. 2nd ed. John Wiley & Sons, 62–69.

Moore, Thomas J, Hanzhe Zhang, Gerard Anderson, G Caleb Alexander. 2018. Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015-2016. *JAMA internal medicine* **178**(11) 1451–1457.

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`.

O'Brien, Peter C, Thomas R Fleming. 1979. A multiple testing procedure for clinical trials. *Biometrics* 549–556.

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.

Rojas-Cordova, Alba, Ebru K Bish. 2018. Optimal patient enrollment in sequential adaptive clinical trials with binary response. *Available at SSRN 3234590* .

Rojas-Cordova, Alba C, Niyousha Hosseinichimeh. 2018. Trial termination and drug misclassification in sequential adaptive clinical trials. *Service Science* **10**(3) 354–377.

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

Sargent, Daniel J, Harry S Wieand, Daniel G Haller, Richard Gray, Jacqueline K Benedetti, Marc Buyse, Roberto Labianca, Jean Francois Seitz, Christopher J O'Callaghan, Guido Francini, et al. 2005. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology* **23**(34) 8664–8670.

Schoenfeld, David. 1981. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**(1) 316–319.

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* .

Smith, David H, Hugh Gravelle. 2001. The practice of discounting in economic evaluations of healthcare interventions. *International journal of technology assessment in health care* 236–243.

Spellberg, Brad, Priya Sharma, John H Rex. 2012. The critical impact of time discounting on economic incentives to overcome the antibiotic market failure. *Nature Reviews Drug Discovery* **11**(2) 168–168.

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

Sugimoto, Tomoyuki, Takashi Sozu, Toshimitsu Hamasaki, Scott R. Evans. 2013. A logrank test-based method for sizing clinical trials with two co-primary time-to-event endpoints. *Biostatistics* **14**(3) 409–421.

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.

Appendix A: Causal Model for Drug Impact on Disease Progression and Survival

Here, we posit a causal model of drug impact on disease progression and overall survival, with the goal of identifying potential mechanisms for a drug to simultaneously exhibit high study-level correlation ρ_0 and low individual-level correlation ρ_I, or vice versa. The model is purposely simple (*e.g.*, it does not model patient heterogeneity in response to treatment) to highlight simple mechanisms that can drive ρ_0 and ρ_I.

Assume a disease has some continuously varying severity level that tends to worsen through time (*e.g.*, the log of the number of cancer cells in a patient with metastatic cancer). Prior to disease progression, we model the amount the disease state has increased above a baseline value by time t as a Brownian motion with variance d_1^2 and drift $c_1 e^{\delta_1} > 0$, where δ_1 controls how the drug changes the rate of disease progression relative to standard therapy (which has drift $c_1 > 0$). Define the time to progression (TTP), which is our surrogate outcome, to be the time at which the Brownian motion hits some threshold a_1. Then

$$\mathbb{E}[TTP] = \frac{a_1}{c_1 e^{\delta_1}} \quad \text{and} \quad \text{Var}(TTP) = \frac{d_1^2 a_1}{(c_1 e^{\delta_1})^3}.$$

Following disease progression, we use an identical setup to model post-progression changes in disease state at time t after disease progression as a second, independent Brownian motion with variance d_2^2 and drift $c_2 e^{\delta_2} > 0$. By defining survival post-progression (SPP) as the time at which the second Brownian motion reaches threshold a_2, we obtain

$$\mathbb{E}[SPP] = \frac{a_2}{c_2 e^{\delta_2}} \quad \text{and} \quad \text{Var}(SPP) = \frac{d_2^2 a_2}{(c_2 e^{\delta_2})^3}.$$

Let OS represent the patient's overall survival (the true outcome). Then,

$$OS = TTP + SPP\,.$$

We model the control group as having drug effects $\delta_1 = \delta_2 = 0$ and the experimental therapy as having drug effects

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}),$$

uniformly experienced by all patients in the treatment arm. Patients will typically discontinue a drug following disease progression, but we still model a potential drug effect δ_2 on SPP because the drug may impact a patient's ability to tolerate a follow-on therapy or their sensitization to the selected follow-on therapy. Given that δ_2 only captures indirect effects of the drug, we expect $\sigma_2 < \sigma_1$.

Conditioned on a fixed δ_1 and δ_2, a patient's TTP and SPP are independent. Thus, a patient's TTP and OS will have low correlation (yielding small ρ_I) when SPP has a larger variance than TTP, and a high correlation when SPP has a smaller variance than TTP. The experimental drug impacts TTP via δ_1 alone but impacts OS via both δ_1 and δ_2. Consequently, we may expect the average impact on the two outcomes to be highly correlated (yielding large ρ_0) if TTP has a much larger mean than SPP or if δ_1 and δ_2 are highly correlated, while we would obtain a small ρ_0 if δ_1 and δ_2 are uncorrelated and SPP has a larger mean than TTP. Summarizing:

• We can obtain large ρ_0 and small ρ_I with correlated δ_1 and δ_2, coupled with a larger variance for SPP than for TTP. We might expect this to hold in cases where TTP tends to be a small fraction of OS. As a numerical example of this, $\rho_0 \approx 0.98$ and $\rho_I \approx 0.12$ with parameter set $d_1 = 1.5$, $a_1 = 2$, $c_1 = 1$, $d_2 = 5$, $a_2 = 20$, $c_2 = 1$, $\sigma_1 = 0.2$, $\sigma_2 = 0.1$, and $\rho = 0.99$, which implies that $\mathbb{E}[TTP] = 2e^{-\delta_1}$, $\text{Var}(TTP) = 4.5e^{-3\delta_1}$, $\mathbb{E}[SPP] = 20e^{-\delta_2}$, and $\text{Var}(SPP) = 500e^{-3\delta_2}$.

• We can obtain small ρ_0 and large ρ_I with uncorrelated δ_1 and δ_2, coupled with a larger variance for TTP than for SPP and a larger mean for SPP than for TTP. We might expect a simultaneously small mean and large variance for TTP for drugs that have a small number of "exceptional responders" who obtain a very large TTP; examples may include some recent immunotherapy drugs for certain cancers. As a numerical example of this, $\rho_0 \approx 0.15$ and $\rho_I \approx 0.98$ with parameter set $d_1 = 5$, $a_1 = 0.3$, $c_1 = 2$, $d_2 = 0.05$, $a_2 = 7$, $c_2 = 0.5$, $\sigma_1 = 0.2$, $\sigma_2 = 0.1$, and $\rho = 0$, which implies that $\mathbb{E}[TTP] = 0.15e^{-\delta_1}$, $\text{Var}(TTP) = 0.9375e^{-3\delta_1}$, $\mathbb{E}[SPP] = 14e^{-\delta_2}$, and $\text{Var}(SPP) = 0.14e^{-3\delta_2}$.

Appendix B: Bayesian Updates for Time-to-Event Outcomes

B.1. Notation

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

Let period 1 be the time between the start of the study and the interim analysis, let period 2 be the time between the interim and final analyses, and let period 3 be the time following the final analysis (we do not observe any information about period 3 in the trial). Define q_{ij} to be the number of patients with a surrogate event in period $i \in \{1,2,3\}$ and with a true outcome event in period $j \in \{1,2,3\}$. Since we assume that surrogate events precede true outcome events, $q_{21} = q_{31} = q_{32} = 0$. For notational simplicity, at the final analysis define the total number of surrogate outcomes observed as $q_S = \sum_{i=1}^{2}\sum_{j=1}^{3} q_{ij}$, and the total number of true outcomes observed as $q_T = \sum_{i=1}^{2}\sum_{j=1}^{i} q_{ji}$.

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

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

B.2. Effect Size Estimates for Patient Subsets

To perform our Bayesian updates, we require effect size estimate vectors $\hat{\mathbf{e}}_{ij}$ for any group of patients with a surrogate observed in period i and true outcome observed in period j (that is, for each pairing of periods $i, j \in \{1,2,3\}$ such that $q_{ij} > 0$). This requires us to separately estimate effect sizes among different groups of patients depending on when their outcomes are observed. This can be accomplished by modifying the logrank test or Cox proportional hazards, which are the standard procedures used to estimate hazard ratios.

In this section, we briefly describe each procedure, along with the (straightforward) modifications needed to obtain effect size estimates for our different patient subsets.

We begin with some notation. Assume there is a clinical trial in which there are K observed events at unique times $t_1 < t_2 < \cdots < t_K$ after patient enrollment.[5] Let O_k take value 1 if the observed event at time t_k was in the treatment group, and let it take value 0 if the observed event was in the control group. Further define N_{1k} to be the number of treatment group patients at risk at time t_k; that is, N_{1k} treatment group patients did not experience an event and were not censored before t_k time had elapsed since their enrollment. Similarly define N_{0k} to be the number of at risk patients in the control group at time t_k and $N_k = N_{0k} + N_{1k}$ to be the total number of patients at risk. Let $[K] := \{1, \cdots, K\}$ denote the set of all event time indices, and for any pair of periods i and j let $K_{ij} \subseteq [K]$ indicate the subset of those indices corresponding to events for patients who observed a surrogate outcome in period i and a true outcome in period j.

The Mantel-Haenzel estimator (Machin et al. 2006) of the hazard ratio across all patients in the clinical trial, $\hat{\mu}_{MH}$, and its variance $var(\hat{\mu}_{MH})$ can be easily calculated from these primitives:

$$\hat{\mu}_{MH} = \frac{\sum_{k\in[K]} O_k - N_{1k}/N_k}{\sum_{k\in[K]} N_{0k}N_{1k}/N_k^2} \quad \text{and} \quad var(\hat{\mu}_{MH}) = \frac{1}{\sum_{k\in[K]} N_{0k}N_{1k}/N_k^2}.$$

The Cox Proportional Hazards model (Cox 1972) estimate of the hazard ratio can be obtained by solving for $\hat{\mu}_{CPH}$ within the equation

$$\sum_{k\in[K]} O_k - \frac{N_{1k}e^{\hat{\mu}_{CPH}}}{N_{0k} + N_{1k}e^{\hat{\mu}_{CPH}}} = 0,$$

and its variance can then be computed as

$$var(\hat{\mu}_{CPH}) = \frac{1}{\sum_{k\in[K]} N_{0k}N_{1k}e^{\hat{\mu}_{CPH}}/(N_{0k} + N_{1k}e^{\hat{\mu}_{CPH}})^2}.$$

In both cases, the estimator can be modified to estimate the log hazard ratio for periods i and j by replacing all cases of summation over $[K]$ with summation over K_{ij}.

For Bayesian updates we assume that either $q_{ij} = 0$ (in which case we fix $\hat{\mathbf{e}}_{ij} = \mathbf{0}$), or otherwise q_{ij} is sufficiently large that approximate bivariate normality is expected to hold. When performing Bayesian updates using different $\hat{\mathbf{e}}_{ij}$, we will assume that each of these effect size vectors have the same correlation ρ_I between the surrogate and true outcomes for the group of patients. In reality these might somewhat differ if event times have "early dependence" (patients with smaller event times have a larger correlation between the two event times) or "late dependence" (patients with larger event times have a larger correlation).

B.3. 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 interim analysis in a type A trial (q_{11} patients with a true outcome observed), in a type B trial ($q_{11} + q_{12} + q_{13}$ patients with a surrogate outcome observed),

[5] For simplicity we present here a setting with no ties. There are straightforward extensions of all procedures presented here that handle tied event times.

and in a type C trial (q_{11} patients with both outcomes observed and an additional $q_{12}+q_{13}$ patients with only a surrogate outcome observed).

$$
\begin{aligned}
\boldsymbol{\mu}|\mathcal{F}_{1A} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1A}, \boldsymbol{\Sigma}_{1A}) \\
\hat{\boldsymbol{\mu}}_{1A} &:= \boldsymbol{\Sigma}_{1A}(q_{11}\mathbf{T}\hat{\mathbf{e}}_{11}) \\
\boldsymbol{\Sigma}_{1A} &:= (\boldsymbol{\Sigma}_0^{-1} + q_{11}\mathbf{T})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{1B} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1B}, \boldsymbol{\Sigma}_{1B}) \\
\hat{\boldsymbol{\mu}}_{1B} &:= \boldsymbol{\Sigma}_{1B}(q_{11}\mathbf{S}\hat{\mathbf{e}}_{11} + q_{12}\mathbf{S}\hat{\mathbf{e}}_{12} + q_{13}\mathbf{S}\hat{\mathbf{e}}_{13}) \\
\boldsymbol{\Sigma}_{1B} &:= (\boldsymbol{\Sigma}_0^{-1} + (q_{11}+q_{12}+q_{13})\mathbf{S})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{1C} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{1C}, \boldsymbol{\Sigma}_{1C}) \\
\hat{\boldsymbol{\mu}}_{1C} &:= \boldsymbol{\Sigma}_{1C}(q_{11}(\mathbf{S}+\mathbf{U})\hat{\mathbf{e}}_{11} + q_{12}\mathbf{S}\hat{\mathbf{e}}_{12} + q_{13}\mathbf{S}\hat{\mathbf{e}}_{13}) \\
\boldsymbol{\Sigma}_{1C} &:= (\Sigma_0^{-1} + (q_{11}+q_{12}+q_{13})\mathbf{S} + q_{11}\mathbf{U})^{-1}
\end{aligned}
$$

A straightforward application of the same identity yields the posterior distribution at the final analysis for a type A trial (after $q_{12}+q_{22}$ additional true outcome observations) and a type B trial (after $q_{22}+q_{23}$ additional surrogate observations). For type C trials, it can be used to perform the update for the q_{23} new patients with only a surrogate observed and the q_{22} new patients with both outcomes observed. An additional q_{12} 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, $\hat{e}_{12T}|\hat{e}_{12S} = x \sim \mathcal{N}(\mu_T + \rho_I(x-\mu_S), 4(1-\rho_I^2)/q_{12})$, so again the identity from Bishop (2006) can be applied. All together, we obtain the following posterior distributions at the final analysis:

$$
\begin{aligned}
\boldsymbol{\mu}|\mathcal{F}_{2A} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2A}, \boldsymbol{\Sigma}_{2A}) \\
\hat{\boldsymbol{\mu}}_{2A} &:= \boldsymbol{\Sigma}_{2A}(\boldsymbol{\Sigma}_{1A}^{-1}\hat{\boldsymbol{\mu}}_{1A} + q_{12}\mathbf{T}\hat{\mathbf{e}}_{12} + q_{22}\mathbf{T}\hat{\mathbf{e}}_{22}) \\
\boldsymbol{\Sigma}_{2A} &:= (\boldsymbol{\Sigma}_{1A}^{-1} + (q_{12}+q_{22})\mathbf{T})^{-1} \\
&\quad = (\boldsymbol{\Sigma}_0^{-1} + (q_{11}+q_{12}+q_{22})\mathbf{T})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{2B} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2B}, \boldsymbol{\Sigma}_{2B}) \\
\hat{\boldsymbol{\mu}}_{2B} &:= \boldsymbol{\Sigma}_{2B}(\boldsymbol{\Sigma}_{1B}^{-1}\hat{\boldsymbol{\mu}}_{1B} + q_{22}\mathbf{S}\hat{\mathbf{e}}_{22} + q_{23}\mathbf{S}\hat{\mathbf{e}}_{23}) \\
\boldsymbol{\Sigma}_{2B} &:= (\boldsymbol{\Sigma}_{1B}^{-1} + (q_{22}+q_{23})\mathbf{S})^{-1} \\
&\quad = (\boldsymbol{\Sigma}_0^{-1} + (q_{11}+q_{12}+q_{13}+q_{22}+q_{23})\mathbf{S})^{-1} \\
\boldsymbol{\mu}|\mathcal{F}_{2C} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{2C}, \boldsymbol{\Sigma}_{2C}) \\
\hat{\boldsymbol{\mu}}_{2C} &:= \boldsymbol{\Sigma}_{2C}(\boldsymbol{\Sigma}_{1C}^{-1}\hat{\boldsymbol{\mu}}_{1C} + q_{22}(\mathbf{S}+\mathbf{U})\hat{\mathbf{e}}_{22} + q_{23}\mathbf{S}\hat{\mathbf{e}}_{23} + q_{12}\mathbf{U}\hat{\mathbf{e}}_{12}) \\
\boldsymbol{\Sigma}_{2C} &:= (\boldsymbol{\Sigma}_{1C}^{-1} + (q_{22}+q_{23})\mathbf{S} + (q_{12}+q_{22})\mathbf{U})^{-1} \\
&\quad = (\boldsymbol{\Sigma}_0^{-1} + (q_{11}+q_{12}+q_{13}+q_{22}+q_{23})\mathbf{S} + (q_{11}+q_{12}+q_{22})\mathbf{U})^{-1}
\end{aligned}
$$

B.4. Variance of Estimates

From the above distributions, we then are able to extract the following expressions for the true outcome posterior variance at the final analysis point. Recall from Appendix B.1 that at the final analysis, we denote the total number of surrogate outcomes observed as $q_S = \sum_{i=1}^{2}\sum_{j=1}^{3} q_{ij}$, and the total number of true outcomes observed as $q_T = \sum_{i=1}^{2}\sum_{j=1}^{i} q_{ji}$. Then we get the following estimates for the posterior variance at the final analysis:

$$V_A = \boldsymbol{\Sigma}_{2A}[2,2] = \frac{\sigma_{0T}^2}{1+q_T\sigma_{0T}^2/4}\,,$$
$$V_B = \boldsymbol{\Sigma}_{2B}[2,2] = \frac{\sigma_{0T}^2+q_S D_0/4}{1+q_S\sigma_{0S}^2/4}\,,$$
$$V_C = \boldsymbol{\Sigma}_{2C}[2,2] = \frac{1}{\gamma_C}\left(D_I\sigma_{0T}^2 + \frac{D_0 q_T}{4} + \frac{(q_S-q_T)D_0 D_I}{4}\right),$$

where for notational convenience we have defined $D_0 = \sigma_{0S}^2\sigma_{0T}^2(1-\rho_0^2)$, $D_I = 1-\rho_I^2$, and

$$\gamma_C = D_I + \frac{q_T^2 D_0}{16} + \frac{q_T\sigma_{0T}^2}{4} + \frac{q_T\sigma_{0S}^2}{4} + \frac{(q_S-q_T)D_0 q_T}{16} + \frac{(q_S-q_T)D_I\sigma_{0S}^2}{4} - \frac{q_T\rho_0\sigma_{0S}\sigma_{0T}\rho_I}{2}\,.$$

Appendix C: Trial Structure Details

C.1. Type I/II Error Control

As discussed in Appendix B, our Bayesian updates are of the form

$$\boldsymbol{\mu}|\mathcal{F}_{pt} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{pt}, \boldsymbol{\Sigma}_{pt}),$$

where $p \in \{1,2\}$ indicates the analysis (interim or final) and $t \in \{A,B,C\}$ indicates the trial type. The trial rejects the null hypothesis whenever $|\hat{\mu}_{ptT}| > m_p$; that is, whenever the posterior mean of the effect size estimate exceeds the thresholds specified for the analysis p in the trial design. Key to studying Type I/II error control, then, is the distribution of $\hat{\mu}_{ptT}$ under either the null hypothesis ($\mu_T = 0$, which implies $\mu_S \sim \mathcal{N}(0, (1-\rho_0^2)\sigma_{0S}^2)$) or when the effect size magnitude is large ($\mu_T = \pm\delta$, which implies $\mu_S \sim \mathcal{N}(\pm\sigma_{0S}\rho_0\delta/\sigma_{0T}, (1-\rho_0^2)\sigma_{0S}^2)$).

From the distribution of a Bayesian update's posterior mean when the true value of interest is actually fixed instead of being drawn from the prior distribution,

$$\hat{\mu}_{1tT}|\{\mu_T = x\} \sim \mathcal{N}((1-v_1\sigma_{0T}^2)x, v_1(1-v_1/\sigma_{0T}^2))$$

for each trial type $t \in \{A,B,C\}$. Let α_1 be the probability of rejecting the null hypothesis at the interim analysis when it in fact holds, and let $1-\beta_1$ be the probability of rejecting the null hypothesis at the interim analysis when $|\mu_T| = \delta$. Defining

$$m_1 = z_1\sqrt{v_1(1-v_1/\sigma_{0T}^2)},$$

then the tail probabilities of the normal distribution yield

$$\alpha_1(v_1,z_1) = 2\Phi(-z_1) \quad \text{and} \quad 1-\beta_1(v_1,z_1) = 1+\Phi(-z_1-\delta\sqrt{1/v_1-1/\sigma_{0T}^2}) - \Phi(z_1-\delta\sqrt{1/v_1-1/\sigma_{0T}^2}).$$

Similar logic can be used to study rejecting the null hypothesis at the final analysis. First,

$$\hat{\mu}_{2tT}|\{\mathcal{F}_{1t}, \mu_T = x\} \sim \mathcal{N}((v_2/v_1)\hat{\mu}_{1tT} + (1-v_2/v_1)x, v_2(1-v_2/v_1)).$$

As before, define α_2 and $1-\beta_2$ as the probabilities of rejecting the null hypothesis at the final analysis (but not at the interim analysis) under our two scenarios. Defining

$$m_2 = z_2\sqrt{v_2(1-v_2/\sigma_{0T}^2)},$$

integration over the z-score of $\hat{\mu}_{1tT}$ yields

$$\alpha_2(v_1,v_2,z_1,z_2) = \int_{-z_1}^{z_1}\phi(x)\left(1+\Phi(\frac{-\frac{m_2}{v_2}-x\sqrt{\frac{1}{v_1}-\frac{1}{\sigma_{0T}^2}}}{\sqrt{\frac{1}{v_2}-\frac{1}{v_1}}})-\Phi(\frac{\frac{m_2}{v_2}-x\sqrt{\frac{1}{v_1}-\frac{1}{\sigma_{0T}^2}}}{\sqrt{\frac{1}{v_2}-\frac{1}{v_1}}})\right)dx \text{ and}$$

$$1-\beta_2(v_1,v_2,z_1,z_2) = \int_{-z_1}^{z_1}\phi(x-\delta\sqrt{\frac{1}{v_1}-\frac{1}{\sigma_{0T}^2}})\left(1+\Phi(\frac{-\frac{m_2}{v_2}-x\sqrt{\frac{1}{v_1}-\frac{1}{\sigma_{0T}^2}}}{\sqrt{\frac{1}{v_2}-\frac{1}{v_1}}}-\delta\sqrt{\frac{1}{v_2}-\frac{1}{v_1}})-\Phi(\frac{\frac{m_2}{v_2}-x\sqrt{\frac{1}{v_1}-\frac{1}{\sigma_{0T}^2}}}{\sqrt{\frac{1}{v_2}-\frac{1}{v_1}}}-\delta\sqrt{\frac{1}{v_2}-\frac{1}{v_1}})\right)dx.$$

With these preliminaries in place, it is straightforward to obtain trial designs that provide the desired Type I/II error control. We begin with a trial design with no interim analysis (which is equivalent to setting $m_1=\infty$). This setting collapses to the rejection probabilities for the interim analysis shown above. Type I error control requires

$$z_2 = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2),$$

and Type II error control implies that v_2 must be selected as the unique solution to

$$\xi(v_2) = \Phi(\Phi^{-1}(1-\alpha/2)-\delta\sqrt{1/v_2-1/\sigma_{0T}^2})-\Phi(\Phi^{-1}(\alpha/2)-\delta\sqrt{1/v_2-1/\sigma_{0T}^2}) = \beta.$$

Since m_2 and v_2 are uniquely determined by the Type I/II error control requirements, finding a minimal-cost trial design with no interim analysis can be accomplished via a 1-dimensional line search to find the minimal-cost patient enrollment target n.

With an interim analysis and final analysis, define

$$\alpha(v_1,v_2,z_1,z_2) = \alpha_1(v_1,z_1)+\alpha_2(v_1,v_2,z_1,z_2) \text{ and}$$
$$1-\beta(v_1,v_2,z_1,z_2) = (1-\beta_1(v_1,z_1))+(1-\beta_2(v_1,v_2,z_1,z_2)).$$

We seek the minimum-cost design such that $\alpha(v_1,v_2,z_1,z_2)=\alpha$ and $\beta(v_1,v_2,z_1,z_2)=\beta$. Define

$$z^\alpha(v_1,v_2,z_1) = \min\{z_2 \mid \alpha_1(v_1,z_1)+\alpha_2(v_1,v_2,z_1,z_2)\le\alpha\}$$

to be the minimum final analysis z-score z_2 that would still satisfy the Type I error requirements, assuming an interim analysis run with parameters v_1 and z_1 and a final analysis run once the posterior variance reaches v_2. Similarly define

$$z^\beta(v_1,v_2,z_1) = \max\{z_2 \mid (1-\beta_1(v_1,z_1))+(1-\beta_2(v_1,v_2,z_1,z_2))\ge 1-\beta\}$$

to be the maximum z_2 that satisfies the Type II error requirements under parameters v_1, z_1, and v_2. For a fixed v_1 and z_1, any v_2 such that $z^\alpha(v_1,v_2,z_1)\le z^\beta(v_1,v_2,z_1)$ can be used to simultaneously meet the Type I and II error requirements, by selecting any $z_2\in[z^\alpha(v_1,v_2,z_1),z^\beta(v_1,v_2,z_1)]$. The minimum-cost such design is the one with final analysis variance

$$v_2(v_1,z_1) = \max\{v_2 \mid z^\alpha(v_1,v_2,z_1)\le z^\beta(v_1,v_2,z_1)\}$$

and final analysis z-score

$$z_2(v_1, z_1) = z^\beta(v_1, v_2(v_1, z_1), z_1).$$

$v_2(v_1, z_1)$ and $z_2(v_1, z_1)$ can easily be computed via one-dimensional line searching. Since the Type I/II error constraints imply fixed v_2 and z_2 values given specified v_1 and z_1 values, optimizing the trial design can be accomplished by searching for a minimal-cost design within a 3-dimensional space of n, v_1, and z_1 values. In our experience, high-quality trial designs can be found quickly via standard non-linear optimization techniques.

C.2. Trial Costs Under Exponentially Distributed Outcomes

Consider a patient whose selected trial arm (control or treatment) experiences an event (surrogate or true outcome) following an exponential distribution with rate λ. Our trial's n patients enroll uniformly at random between times 0 and n/λ_E, so $R(w, \lambda)$ — the probability that the patient will have been enrolled and observed the event within w time of the start of the trial — can be obtained via the convolution of a uniform and exponential random variable:

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

Our trial allocates half of patients to the control arm and half to the treatment arm. Thus, for effect sizes μ_S and μ_T, the expected probability that an arbitrary patient (regardless of the treatment arm they were allocated to) experiences the surrogate or true outcome by time w respectively are

$$A_S(w, \mu_S) = (R(w, \lambda_S) + R(w, e^{\mu_S}\lambda_S))/2 \qquad \text{and} \qquad A_T(w, \mu_T) = (R(w, \lambda_T) + R(w, e^{\mu_T}\lambda_T))/2\,.$$

Recall that our trials perform interim and final analyses after reaching posterior variances v_1 and v_2. Let T_1^t and T_2^t be the random time of these two analyses for a trial of type $t \in \{A, B, C\}$. As discussed earlier, these stopping times are reached when some proportions p_1 and p_2 of the patients experience an event (for Type C trials, this is an effective sample proportion that can be achieved by different combinations of surrogate and true outcome events). This random variable is an order statistic, and the asymptotics of its sample quantiles are well-studied. We consider the large n limit, while ensuring that the patient enrollment rate λ_E also grows at a similar rate (otherwise the proportion of patients enrolling by any fixed time w will approach 0), *i.e.*, we fix the ratio n/λ_E to some positive constant c. Leveraging the multivariate central limit theorem and the multivariate delta-method, we can establish that

$$\begin{aligned} \lim_{\substack{n, \lambda_E \to \infty \\ n/\lambda_E = c}} \mathbb{E}[T_i^A \mid \mu_T] &= A_T^{-1}(p_i, \mu_T) \\ \lim_{\substack{n, \lambda_E \to \infty \\ n/\lambda_E = c}} \mathbb{E}[T_i^B \mid \mu_S] &= A_S^{-1}(p_i, \mu_S) \\ \lim_{\substack{n, \lambda_E \to \infty \\ n/\lambda_E = c}} \mathbb{E}[T_i^C \mid \mu_S, \mu_T] &= s^{-1}(p_i, \mu_S, \mu_T) \end{aligned}$$

where we have defined the function

$$s(t, \mu_S, \mu_T) = \frac{A_S(t, \mu_S)A_T(t, \mu_T)}{(1 - \rho_I^2)A_S(t, \mu_S) + \rho_I^2 A_T(t, \mu_T)}\,.$$

Denote P_1^t and P_2^t to be the proportions of patients enrolled by the interim and final analyses. Using the same logic yields

$$\lim_{\substack{n,\lambda_E\to\infty \\ n/\lambda_E=c}} \mathbb{E}[P_i^A \mid \mu_T] = \min\{A_T^{-1}(p_i,\mu_T)\lambda_E/n, 1\}$$
$$\lim_{\substack{n,\lambda_E\to\infty \\ n/\lambda_E=c}} \mathbb{E}[P_i^B \mid \mu_S] = \min\{A_S^{-1}(p_i,\mu_S)\lambda_E/n, 1\}$$
$$\lim_{\substack{n,\lambda_E\to\infty \\ n/\lambda_E=c}} \mathbb{E}[P_i^C \mid \mu_S,\mu_T] = \min\{s^{-1}(p_i,\mu_S,\mu_T), 1\}\,.$$

The overall expected discounted cost of the trial length can be obtained by taking the expectation over effect sizes μ_S and μ_T, noting that the discounted cost of a trial of length W is $c_w/r(1-e^{-rW})$. The total expected discounted costs of waiting until the interim analysis are

$$W_1^A = c_w/r\left(1-\int_{-\infty}^{\infty} e^{-rA_T^{-1}(p_1,t)}\phi(t,0,\sigma_{0T}^2)dt\right),$$
$$W_1^B = c_w/r\left(1-\int_{-\infty}^{\infty} e^{-rA_S^{-1}(p_1,s)}\phi(s,0,\sigma_{0S}^2)ds\right), \text{ and}$$
$$W_1^C = c_w/r\left(1-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-rs^{-1}(p_1,s,t)}\phi([s\ t]',\mathbf{0},\mathbf{\Sigma}_0)dsdt\right),$$

where $\phi(x,\mu,\sigma^2)$ is the normal pdf under mean μ and variance σ^2. The total discounted costs of waiting between the interim and final analyses can be similarly computed, though here we additionally integrate over the posterior mean values obtained at the interim analysis:

$$W_2^A = c_w/r\int_{-r_1}^{r_1}\int_{-\infty}^{\infty}(e^{-rA_T^{-1}(p_1,t)}-e^{-rA_T^{-1}(p_2,t)})\phi(t,x,v_1)\phi(x,0,\sigma_{0T}^2-v_1)dtdx,$$
$$W_2^B = c_w/r\int_{-r_1}^{r_1}\int_{-\infty}^{\infty}(e^{-rA_S^{-1}(p_1,s)}-e^{-rA_S^{-1}(p_2,t)})\phi(s,\frac{\sigma_{0S}}{\sigma_{0T}\rho_0}x,\sigma_{0S}^2(1-\frac{\sigma_{0T}^2-v_1}{\sigma_{0T}^2\rho_0^2}))\phi(x,0,\sigma_{0T}^2-v_1)dtdx, \text{ and}$$
$$W_2^C = c_w/r\int_{-r_1}^{r_1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(e^{-rs^{-1}(p_1,s,t)}-e^{-rs^{-1}(p_2,s,t)})\phi([s\ t]',[w\ x]',\mathbf{\Sigma}_{1C})\phi([w\ x]',\mathbf{0},\mathbf{\Sigma}_0-\mathbf{\Sigma}_{1C})dsdtdwdx.$$

Patient enrollment costs P_i^t can similarly be computed, noting that the discounted cost to enroll p proportion of the n patients uniformly through time is $c_p\lambda_E/r(1-e^{-rnp/\lambda_E})$.

While conceptually straightforward, these expected discounted costs do not have closed-form solutions under normally distributed priors for μ_S and μ_T, and nested integrals can be computationally burdensome. As we result, we derive the following approximation that can be more easily integrated into our analytic models.

PROPOSITION 1. *In the regime where λ_S and λ_T are small (events occur slowly) and the final analysis time occurs after all patients are enrolled $w_t^* > n_t^*/\lambda_E$, the proportion of patients who experience surrogate and true outcomes respectively in a type t trial of length w_t^* concentrates to*

$$\mathbb{E}_{\mu_S\sim\mathcal{N}(0,\sigma_{0S}^2)}[A_S^{-1}(p,\mu_S)] \approx A_S^{-1}(p,0)\,,$$
$$\mathbb{E}_{\mu_T\sim\mathcal{N}(0,\sigma_{0T}^2)}[A_T^{-1}(p,\mu_T)] \approx A_T^{-1}(p,0)\,,$$
$$\mathbb{E}_{\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_0)}[s^{-1}(p,\mu_S,\mu_T)] \approx s^{-1}(p,0,0)\,.$$

Proof of Proposition 1 We begin by using our assumption that our hazard rate λ is small (for both outcomes) and $t > n_t^*/\lambda_E$ to simplify and approximate R, the probability that the patient will have been enrolled and observed an event within time t —

$$\begin{aligned} R(t,\lambda) &= 1 - \frac{e^{-\lambda(t-n_t^*/\lambda_E)} - e^{-\lambda t}}{\lambda n_t^*/\lambda_E} \\ &= 1 - \frac{(1-\lambda(t-n_t^*/\lambda_E) + \lambda^2(t-n_t^*/\lambda_E)^2/2 + O(\lambda^3)) - (1-\lambda t + \lambda^2 t^2/2 + O(\lambda^3))}{\lambda n_t^*/\lambda_E} \\ &= \lambda(t - n_t^*/(2\lambda_E)) + O(\lambda^2)\,. \end{aligned}$$

This implies that the proportion of patients that experience a surrogate and true outcome respectively at time t is approximately

$$A_S(\mu_S, w_t^*) \approx \lambda_S(t - n_t^*/(2\lambda_E))\frac{1+e^{\mu_S}}{2} \quad \text{and} \quad A_T(\mu_T, w_t^*) \approx \lambda_T(t - n_t^*/(2\lambda_E))\frac{1+e^{\mu_T}}{2}\,.$$

We can then invert the above expressions, yielding

$$\begin{aligned} A_S^{-1}(p,\mu_S) &\approx \frac{2p/\lambda_S}{1+e^{\mu_S}} + n_t^*/(2\lambda_E)\,, \\ A_T^{-1}(p,\mu_T) &\approx \frac{2p/\lambda_T}{1+e^{\mu_T}} + n_t^*/(2\lambda_E)\,, \\ s^{-1}(p,\mu_S,\mu_T) &\approx \frac{2p(1-\rho_I^2)/\lambda_T}{1+e^{\mu_T}} + \frac{2p\rho_I^2/\lambda_S}{1+e^{\mu_S}} + n_t^*/(2\lambda_E)\,. \end{aligned}$$

Applying the identity that

$$\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}\left[\frac{2}{1+e^X}\right] = 1\,,$$

we find that $\mathbb{E}[A_S^{-1}(p,\mu_S)] \approx A_S^{-1}(p,0)$, $\mathbb{E}[A_T^{-1}(p,\mu_T)] \approx A_T^{-1}(p,0)$, and $\mathbb{E}[s^{-1}(p,\mu_S,\mu_T)] \approx s^{-1}(p,0,0)$, as desired. In other words, substituting $\mu_S = \mu_T = 0$ yields a good approximation for the expectations of A_S, A_T over the random variables μ_S, μ_T as desired. □

While Proposition 1 above is stated for evaluating the expectations of A_S, A_T over distributions of μ_S, μ_T centered at 0, it suggests the following reasonable approximation $\mathbb{E}[A_S^{-1}(p,\mu_S)] \approx A_S^{-1}(p,\mathbb{E}\mu_S)$, $\mathbb{E}[A_T^{-1}(p,\mu_T)] \approx A_T^{-1}(p,\mathbb{E}\mu_T)$ for sufficiently small $\mathbb{E}\mu_S, \mathbb{E}\mu_T$. This is relevant since the center of our means may shift after conditioning on observations collected thus far at the interim analysis; we use the resulting approximation for computational tractability to optimize our trial designs in Sections 3.3 and 4.

Appendix D: Comparative Statics

This section collects results and details presented in Section 3. Appendix D.1 provides the proofs of Lemmas 1-3, Appendix D.2 takes a large n_C^* approximation to interpret promising regions where Type C trials perform well, Appendix D.3 characterizes local cost minima for Type C trials in (ρ_0, ρ_I) space, Appendix D.4 provides the proof of Theorem 1, and Appendix D.5 describes parameter selection and data collection from meta-analyses presented in Section 3.

D.1. Proof of Lemmas 1, 2 and 3

We can recast our constrained optimization problem in Section 2.5 under the simplified setting described in Section 3. Recall that we are considering the undiscounted setting where the final analysis time occurs after all patients are enrolled, and with no interim analysis ($m_1 = \infty$). Then, for a type t trial, the expected number of patients simplifies to the target enrollment n_t, and we denote the expected waiting time to reach the target posterior variance v as w_t. Proposition 1 in Appendix C.2 shows that the expected proportion of patients who experience surrogate and true outcomes in a trial concentrates to $A_S^{-1}(0, w_t)$ and $A_T^{-1}(0, w_t)$ respectively; we can thus take a similar approach to derive the following expressions for the expected proportion of surrogate and true outcomes A_S and A_T respectively for a trial of expected length w_t:

$$\mathbb{E}_{\mu_S \sim \mathcal{N}(0, \sigma_{0S}^2)}[A_S(\mu_S, w_t)] \approx A_S(0, w_t) = 1 + \frac{1}{n_t \lambda_S / \lambda_E} e^{-\lambda_S w_t} \left(1 - e^{n_t \lambda_S / \lambda_E}\right),$$
$$\mathbb{E}_{\mu_T \sim \mathcal{N}(0, \sigma_{0T}^2)}[A_T(\mu_T, w_t)] \approx A_T(0, w_t) = 1 + \frac{1}{n_t \lambda_T / \lambda_E} e^{-\lambda_T w_t} \left(1 - e^{n_t \lambda_T / \lambda_E}\right).$$

With some abuse of notation, we will drop the first argument of $A_S(0, w_t)$ and $A_T(0, w_t)$ for the remainder of this appendix. Then, we can denote the number of observed surrogate and true outcomes at the final analysis as $q_S = n_t A_S(w_t)$ and $q_T = n_t A_T(w_t)$ respectively. We can compute the resulting variance of our posterior estimate $V_t(n_t, w_t)$ of $\hat{\mu}_T$ (expressions given in Appendix B.4), and impose that we meet our target variance at the final analysis, *i.e.*, $V_t(n_t, w_t) \leq v$.

To simplify the algebra, we further perform a variable transformation from v to $\kappa = 4/v - 4/\sigma_{0T}^2$; this is useful to make the Type I/II error constraint (given by the function ξ) independent of all exogenous parameters of interest. Specifically, recall from Appendix C that Type I error control without an interim analysis requires $z_2 = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1 - \alpha/2)$. Then Type II error control implies that v (or our transformed variable κ) must be selected as the unique solution to $\xi(\kappa) = \Phi(\Phi^{-1}(1-\alpha/2) - \delta\sqrt{\kappa/4}) - \Phi(\Phi^{-1}(\alpha/2) - \delta\sqrt{\kappa/4}) = \beta$.

Applying the above, the constrained optimization problem for finding the optimal parameters for the trial design in Section 2.5 simplifies to:

$$\begin{aligned} \min_{n,\kappa,w} \; & c_n n + c_w w \\ \text{s.t.} \; & V_t(n, w) \leq \frac{\sigma_{0T}^2}{1 + \kappa \sigma_{0T}^2 / 4} \\ & \xi(\kappa) \leq \beta \,. \end{aligned}$$

For trial type t, denote the solutions to the optimization problem above as n_t^*, κ_t^* and w_t^*. To evaluate the comparative statics with respect to any exogenous parameter p, we can invoke the envelope theorem

$$\frac{d\text{Cost}_t}{dp} = \frac{\partial c_n n_t^*}{\partial p} + \frac{\partial c_w w_t^*}{\partial p} + \eta_t^* \frac{\partial}{\partial p}\left(\frac{\sigma_{0T}^2}{1 + \kappa_t^* \sigma_{0T}^2/4} - V_t(n_t^*, w_t^*, p)\right) + \zeta_t^* \frac{\partial(\beta - \xi(\kappa_t^*))}{\partial p}, \tag{2}$$

where $\eta_t^*, \zeta_t^* < 0$ are Lagrange multipliers.

Preliminaries. As previously discussed, $q_S = n_t^* A_S(w_t^*)$ and $q_T = n_t^* A_T(w_t^*)$. Clearly, among our parameters of interest, these quantities depend only on λ_S and λ_T respectively. We can then examine the surrogate outcome arrival rate in the control arm:

$$\frac{\partial A_S(w_t^*)}{\partial \lambda_S} = \underbrace{\frac{e^{-\lambda_S w_t^*}}{n_t^* \lambda_S^2 / \lambda_E}}_{+} \underbrace{(\lambda_S(w_t^* - n_t^*/\lambda_E) + 1)}_{+} \underbrace{\left(e^{n_t^* \lambda_S / \lambda_E} - 1 - \frac{n_t^* \lambda_S / \lambda_E}{\lambda_S(w_t^* - n_t^*/\lambda_E) + 1}\right)}_{+} > 0 \,. \tag{3}$$

The above computation holds analogously for $A_T(w_T^*)$ and λ_T.

D.1.1. Type A Trials. Recall from Appendix B.4 that

$$V_A = \frac{\sigma_{0T}^2}{1 + n_A^* A_T(w_A^*)\sigma_{0T}^2/4}\,.$$

Clearly, the objective in Eq. (2) does not depend on any of the surrogate properties, including ρ_I, ρ_0, σ_{0S}, and λ_S. It follows that

$$\frac{d\text{Cost}_A}{d\rho_I} = \frac{d\text{Cost}_A}{d\rho_0} = \frac{d\text{Cost}_A}{d\sigma_{0S}} = \frac{d\text{Cost}_A}{d\lambda_S} = 0\,.$$

We also can trivially observe that

$$\frac{d\text{Cost}_A}{dc_n} = n_A^* > 0 \quad \text{and} \quad \frac{d\text{Cost}_A}{dc_w} = w_A^* > 0\,.$$

Through algebraic manipulation and our expression for V_A, we can rewrite our constraint of $V_A(n_A^*, w_A^*) \leq \frac{\sigma_{0T}^2}{1+\kappa_A^*\sigma_{0T}^2/4}$ as $\kappa_A^* \leq n_A^* A_T(w_A^*)$. Substituting our rewritten constraint into Eq. (2), we observe that there is no dependence on σ_{0T} and therefore,

$$\frac{d\text{Cost}_A}{d\sigma_{0T}} = 0\,.$$

Using Eq. (3), we can verify

$$\frac{d\text{Cost}_A}{d\lambda_T} = \underbrace{\eta_A^*}_{-}\underbrace{\frac{\partial A_T(w_A^*)}{\partial \lambda_T}}_{+}\underbrace{\frac{n_A^*\sigma_{0T}^4/4}{(1 + n_A^* A_T(w_A^*)\sigma_{0T}^2/4)^2}}_{+} < 0\,.$$

D.1.2. Type B Trials. Recall from Appendix B.4 that

$$V_B = \frac{\sigma_{0T}^2 + n_B^* A_S(w_B^*) D_0/4}{1 + n_B^* A_S(w_B^*)\sigma_{0S}^2/4}\,,$$

where $D_0 = \sigma_{0S}^2\sigma_{0T}^2(1-\rho_0^2) > 0$. Note that $\partial D_0/\partial\rho_0 < 0$.

Clearly, the objective in Eq. (2) does not depend on ρ_I and λ_T. It follows that

$$\frac{d\text{Cost}_B}{d\rho_I} = \frac{d\text{Cost}_A}{d\lambda_T} = 0\,.$$

We can also trivially observe that

$$\frac{d\text{Cost}_B}{dc_n} = n_B^* > 0 \quad \text{and} \quad \frac{d\text{Cost}_B}{dc_w} = w_B^* > 0\,.$$

Next, we can verify

$$\frac{d\text{Cost}_B}{d\sigma_{0S}} = \underbrace{\eta_B^*}_{-}\left(\underbrace{\frac{\rho_0^2 n_B^* A_S(w_B^*)\sigma_{0S}\sigma_{0T}^2}{2(1 + n_B^* A_S(w_B^*)\sigma_{0S}^2/4)^2}}_{+}\right) < 0\,.$$

$$\frac{d\text{Cost}_B}{d\rho_0} = \underbrace{-\eta_B^*}_{+}\underbrace{\frac{\partial D_0}{\partial\rho_0}}_{-}\underbrace{\frac{n_B^* A_S(w_B^*)/4}{1 + n_B^* A_S(w_B^*)\sigma_{0S}^2/4}}_{+} < 0\,.$$

Through algebraic manipulation and our expression for V_B, we can rewrite our constraint of $V_B(n_B^*, w_B^*) \leq \frac{\sigma_{0T}^2}{1+\kappa_B^*\sigma_{0T}^2/4}$ as $\kappa_B^* \leq \frac{n_B^* A_S(w_B^*)\sigma_{0S}^2\rho_0^2}{\sigma_{0T}^2 + D_0 n_B^* A_S(w_B^*)/4}$. Then, with a modified Lagrange multiplier $\eta_B^* < 0$, we have

$$\frac{d\text{Cost}_B}{d\sigma_{0T}} = \eta_B^*\frac{d}{d\sigma_{0T}}\left(\frac{n_B^* A_S(w_B^*)\sigma_{0S}^2\rho_0^2}{\sigma_{0T}^2(1 + \sigma_{0S}^2(1-\rho_0^2)n_B^* A_S(w_B^*)/4)}\right) = \underbrace{\frac{-2\eta_B^* n_B^* A_S(w_B^*)\sigma_{0S}^2\rho_0^2}{\sigma_{0T}^3(1 + \sigma_{0S}^2(1-\rho_0^2)n_B^* A_S(w_B^*)/4)}}_{+} > 0\,.$$

Using Eq. (3) and the fact that $\sigma_{0T}^2\sigma_{0S}^2 - D_0 = \sigma_{0S}^2\sigma_{0T}^2\rho_0^2 > 0$, we can verify

$$\frac{d\text{Cost}_B}{d\lambda_S} = \underbrace{\eta_B^*}_{-}\underbrace{\frac{\partial A_S(w_B^*)}{\partial\lambda_S}}_{+}\underbrace{\frac{n_B^*(\sigma_{0T}^2\sigma_{0S}^2 - D_0)}{4(1 + n_B^* A_S(w_B^*)\sigma_{0S}^2/4)^2}}_{+} < 0\,.$$

D.1.3. Type C Trials. For notational convenience, we leave number of surrogate and true outcomes at the final analysis be denoted by $q_S = n_C^* A_S(w_C^*, \mu_{0S})$ and $q_T = n_C^* A_T(w_C^*, \mu_{0T})$ respectively. Recall from Appendix B.4 that

$$V_C = \frac{1}{\gamma_C}\left(D_I\sigma_{0T}^2 + \frac{D_0 A_T n_C^*}{4} + \frac{(q_S - q_T)D_0 D_I}{4}\right),$$

where $D_0 = \sigma_{0S}^2\sigma_{0T}^2(1-\rho_0^2)$, $D_I = 1-\rho_I^2$, and

$$\gamma_C = D_I + \frac{q_T^2 D_0}{16} + \frac{q_T\sigma_{0T}^2}{4} + \frac{q_T\sigma_{0S}^2}{4} + \frac{(q_S-q_T)D_0 q_T}{16} + \frac{(q_S-q_T)D_I\sigma_{0S}^2}{4} - \frac{q_T\rho_0\sigma_{0S}\sigma_{0T}\rho_I}{2}.$$

Note that

$$\begin{aligned}\gamma_C &> \frac{q_T\sigma_{0T}^2}{4} + \frac{q_T\sigma_{0S}^2}{4} - \frac{q_T\rho_0\sigma_{0S}\sigma_{0T}\rho_I}{2} \geq \frac{q_T\sigma_{0T}^2}{4} + \frac{q_T\sigma_{0S}^2}{4} - \frac{q_T\sigma_{0S}\sigma_{0T}}{2}\\ &= \frac{q_T}{4}(\sigma_{0T}-\sigma_{0S})^2 \;>\; 0\,.\end{aligned}$$

Additionally,

$$\begin{aligned}\frac{\partial(1/\gamma_C)}{\partial\lambda_S} &= \underbrace{-\frac{1}{\gamma_C^2}}_{-}\underbrace{\frac{\partial q_S}{\partial\lambda_S}}_{+}\underbrace{\left(\frac{D_0 q_T}{16} + \frac{D_I\sigma_{0S}^2}{4}\right)}_{+} < 0\,,\\ \frac{\partial(1/\gamma_C)}{\partial\lambda_T} &= \underbrace{-\frac{1}{\gamma_C^2}}_{-}\underbrace{\frac{\partial q_T}{\partial\lambda_T}}_{+}\underbrace{\left(\frac{q_T D_0}{16} + \frac{(q_S-q_T)D_0}{16} + \frac{\sigma_{0T}^2}{4} + \frac{\rho_I^2\sigma_{0S}^2}{4} + \frac{\sigma_{0S}\sigma_{0T}\rho_0\rho_I}{2}\right)}_{+} < 0\,.\end{aligned}$$

The latter follows from the fact that $\frac{\sigma_{0T}^2}{4} + \frac{\rho_I^2\sigma_{0S}^2}{4} - \frac{\sigma_{0S}\sigma_{0T}\rho_0\rho_I}{2} \geq \frac{1}{4}(\sigma_{0T} - \sigma_{0S}\rho_I)^2 > 0$. Lastly, we note that

$$\frac{dD_0}{d\rho_0}, \frac{dD_I}{d\rho_I} < 0 \quad \text{and} \quad \frac{dD_0}{d\sigma_{0S}}, \frac{dD_0}{d\sigma_{0T}} > 0 \quad \text{and} \quad \frac{dD_0}{d\rho_I}, \frac{dD_I}{d\rho_0}, \frac{dD_I}{d\sigma_{0S}}, \frac{dD_I}{d\sigma_{0T}} = 0\,.$$

As with Type A and B trials, we can trivially observe that

$$\frac{d\text{Cost}_C}{dc_n} = n_C^* > 0 \quad \text{and} \quad \frac{d\text{Cost}_C}{dc_w} = w_C^* > 0\,.$$

Using Eq. (3) and further noting that

$$\begin{aligned}&\frac{1}{\gamma_C}\left(\frac{D_0 q_T}{16} + \frac{D_I\sigma_{0S}^2}{4}\right)\left(D_I\sigma_{0T^2} + \frac{D_0 q_T}{4} + \frac{(q_S-q_T)D_0 D_I}{4}\right) - \frac{D_0 D_I}{4}\\ &\quad= \frac{1}{\gamma_C}\left(\frac{D_0^2 q_T^2\rho_I^2}{64} + \frac{D_I^2\sigma_{0S}^2\sigma_{0T}^2\rho_0^2}{4} + \frac{q_T\rho_0\sigma_{0S}\sigma_{0T}\rho_I D_0 D_I}{8}\right) \;>\; 0\,,\end{aligned}$$

we can verify

$$\frac{d\text{Cost}_C}{d\lambda_S} = \underbrace{\eta_C^*}_{-}\left(\underbrace{\frac{\partial q_S}{\partial\lambda_S}}_{+}\underbrace{\left(\frac{1}{\gamma_C^2}\left(\frac{D_0 q_T}{16} + \frac{D_I\sigma_{0S}^2}{4}\right)\left(D_I\sigma_{0T^2} + \frac{D_0 q_T}{4} + \frac{(q_S-q_T)D_0 D_I}{4}\right) - \frac{1}{\gamma_C}\frac{D_0 D_I}{4}\right)}_{+}\right) < 0\,.$$

Using Eq. (3) and simplifying also yields

$$\begin{aligned}\frac{d\text{Cost}_C}{d\lambda_T} = \underbrace{\eta_C^*}_{-}\underbrace{\frac{\partial q_T}{\partial\lambda_T}}_{+}\underbrace{\frac{1}{\gamma_C^2}}_{+}\Bigg[&\underbrace{\left(\frac{D_0 D_I q_T}{4} + \frac{(q_S-q_T)D_0 D_I}{4} + D_I\sigma_{0T^2}\right)}_{+}\\ &\left(\underbrace{\frac{q_T D_0}{16} + \frac{\sigma_{0T}^2}{4} + \frac{(q_S-q_T)D_0}{16} - \frac{\rho_0\sigma_{0S}\sigma_{0T}\rho_I}{2}}_{+}\right) + \underbrace{\frac{\rho_0^2\rho_I^2\sigma_{0S}^2\sigma_{0T}^2}{4}}_{+}\Bigg] < 0\,.\end{aligned}$$

We now examine our non-monotonic comparative statics based on Eq. (2). In all these cases, we identify a positive and a negative term — we verified numerically that these relationships are indeed non-monotone, and the sign depends on which term dominates.

$$\frac{d\text{Cost}_C}{d\rho_I} = \underbrace{\eta_C^*}_{-}\Bigg[\underbrace{\frac{1}{\gamma_C^2}}_{+}\underbrace{\left(\frac{\partial D_I}{\partial \rho_I}+\frac{(q_S-q_T)\sigma_{0S}^2}{4}\frac{\partial D_I}{\partial \rho_I}-\frac{q_T\rho_0\sigma_{0S}\sigma_{0T}}{2}\right)}_{-}\underbrace{\left(D_I\sigma_{0T}^2+\frac{D_0q_T}{4}+\frac{(q_S-q_T)D_0D_I}{4}\right)}_{+}$$
$$\underbrace{-\frac{1}{\gamma_C}\left(\sigma_{0T}^2+\frac{(q_S-q_T)D_0}{4}\right)}_{-}\underbrace{\frac{\partial D_I}{\partial \rho_I}}_{-}\Bigg] \gtreqless 0\,.$$

$$\frac{d\text{Cost}_C}{d\rho_0} = \underbrace{\eta_C^*}_{-}\Bigg[\underbrace{\frac{1}{\gamma_C^2}}_{+}\underbrace{\left(\frac{q_T^2}{16}\frac{\partial D_0}{\partial \rho_0}+\frac{(q_S-q_T)q_T}{16}\frac{\partial D_0}{\partial \rho_0}-\frac{q_T\sigma_{0S}\sigma_{0T}\rho_I}{2}\right)}_{-}\underbrace{\left(D_I\sigma_{0T}^2+\frac{D_0q_T}{4}+\frac{(q_S-q_T)D_0D_I}{4}\right)}_{+}$$
$$\underbrace{-\frac{1}{\gamma_C}\left(\frac{q_T}{4}+\frac{(q_S-q_T)D_I}{4}\right)}_{-}\underbrace{\frac{\partial D_0}{\partial \rho_0}}_{-}\Bigg] \gtreqless 0\,.$$

Noting that $\partial\gamma_C/\partial\sigma_{0S} \gtreqless 0$,

$$\frac{d\text{Cost}_C}{d\sigma_{0S}} = \underbrace{\eta_C^*}_{}\left(\underbrace{\frac{1}{\gamma_C^2}}_{+}\underbrace{\frac{\partial\gamma_C}{\partial\sigma_{0S}}}_{(+/-)}\underbrace{\left(D_I\sigma_{0T}^2+\frac{D_0o_T}{4}+\frac{(q_S-q_T)D_0D_I}{4}\right)}_{+}\underbrace{-\frac{1}{\gamma_C}}_{-}\underbrace{\frac{\partial D_0}{\partial\sigma_{0S}}}_{+}\underbrace{\left(\frac{o_T}{4}+\frac{(q_S-q_T)D_I}{4}\right)}_{+}\right) \gtreqless 0\,.$$

Similarly,

$$\frac{d\text{Cost}_C}{d\sigma_{0T}} = \underbrace{\frac{\eta_C^*}{\gamma_C^2}}_{-}\underbrace{\frac{\partial\gamma_C}{\partial\sigma_{0T}}}_{(+/-)}\underbrace{\left(D_I\sigma_{0T}^2+\frac{D_0o_T}{4}+\frac{(q_S-q_T)D_0D_I}{4}\right)}_{+}\underbrace{-\frac{-\eta_C^*}{\gamma_C}}_{+}\underbrace{\left(2D_I\sigma_{0T}+\frac{\partial D_0}{\partial\sigma_{0T}}\frac{q_T}{4}+\frac{\partial D_0}{\partial\sigma_{0T}}\frac{(q_S-q_T)D_I}{4}\right)}_{+}$$
$$+\underbrace{\eta_C^*}_{-}\underbrace{\frac{2\sigma_{0T}}{(1+\kappa_C^*\sigma_{0T}^2/4)^2}}_{+}\,.$$

D.2. Type C: Large n Regime

We start with the following lemma that will be useful for deriving our large n approximation.

LEMMA 4. *Given a symmetric matrix $X\in\mathbb{R}^{2\times2}$ 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

$$\begin{aligned}(I+X)^{-1} &= (I+U\Sigma U')^{-1} = U'(I+\Sigma)^{-1}U\\ &= U'\left(\begin{bmatrix}\frac{1}{1+\sigma_0} & 0\\ 0 & \frac{1}{1+\sigma_1}\end{bmatrix}\right)U\end{aligned}$$

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

□

Once again, we seek to use the envelope theorem in Eq. (2) to compute comparative statics for each exogenous parameter p. However, to get analytical expressions, we now approximate $V_C(n_C^*, w_C^*)$ to second order in n_C^*. Recall from Appendix B.3 that $V_C = \Sigma_C[2,2]$, where $\Sigma_C := (\Sigma_0^{-1} + (q_{11}+q_{12}+q_{13})\mathbf{S} + q_{11}\mathbf{U})^{-1}$. We can rewrite this as

$$
\Sigma_C = \left(\Sigma_0^{-1} + n_C^* L\right)^{-1} \quad \text{and} \quad L = \left[\frac{A_T(w_C^*)}{4(1-\rho_I^2)} \begin{bmatrix} \rho_I^2 & -\rho_I \\ -\rho_I & 1 \end{bmatrix} + \frac{A_S(w_C^*)}{4} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right].
$$

Since n_C^* 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.*,

$$
\begin{aligned}
\Sigma_C &= \frac{1}{n_C^*}\left(\frac{1}{n_C^*}L^{-1}\Sigma_0^{-1} + I\right)^{-1} L^{-1} \\
&= \frac{1}{n_C^*}\left(I - \frac{1}{n_C^*}L^{-1}\Sigma_0^{-1} + \frac{1}{{n_C^*}^2}L^{-1}\Sigma_0^{-1}L^{-1}\Sigma_0^{-1} + \mathcal{O}\left(\frac{1}{{n_C^*}^3}\right)\right) L^{-1} \\
&= \frac{1}{n_C^*}L^{-1} - \frac{1}{{n_C^*}^2}L^{-1}\Sigma_0^{-1}L^{-1} + \mathcal{O}\left(\frac{1}{{n_C^*}^3}\right). \qquad (4)
\end{aligned}
$$

Comparative statics in the large n_C^* regime follow from using the above expression in Eq. (2) and taking the highest-order term in n_C^*.

We first examine σ_{0T}. As we did for Types A and B, we can rewrite our constraint of $V_C(n_C^*, w_C^*) \leq \frac{\sigma_{0T}^2}{1+\kappa_C^*\sigma_{0T}^2/4}$ as $\kappa_C^* \leq \frac{n_C^* A_S(w_C^*)A_T(w_C^*)}{A_S(w_C^*)-A_S(w_C^*)\rho_I^2+A_T(w_C^*)\rho_I^2}$ to highest order in n_C^*. Substituting this constraint into Eq. (2), we observe that there is no dependence on σ_{0T} (similar to Type A trials) and therefore, $d\text{Cost}_C/d\sigma_{0T} = 0$, *i.e.,* the non-monotonic dependence on σ_{0T} that we found in Lemma 3 is not salient in practice.

Next, we find that our cost remains non-monotonic in ρ_0, σ_{0S} and ρ_I. Moreover, the surface on which $d\text{Cost}_C/d\rho_0 = 0$ coincides with the surface on which $d\text{Cost}_C/d\sigma_{0S} = 0$. This is because increasing ρ_0 and decreasing σ_{0S} contribute to the same effect — they increase the weight of the study-level prior Σ_0^{-1} relative to the weight of trial outcome observations L^{-1} when evaluating the variance of our true outcome posterior estimate in Eq. (4). Specifically, if the eigenvalues of Σ_0^{-1} are much larger than those of L^{-1}, then the second term of Eq. (4) dominates when computing $(\Sigma_C)_{2,2}$. In particular, note that $(\Sigma_C)_{2,2} = x^\top \Sigma_C x$, where $x = \begin{bmatrix}0 & 1\end{bmatrix}^\top$. Then, we have $x^\top(L^{-1}/n_C^*)x \leq \lambda_{\max}(L^{-1})/n_C^*$, whereas $x^\top L^{-1}\Sigma_0^{-1}L^{-1} \geq \lambda_{\min}(\Sigma_0^{-1})\lambda_{\min}(L^{-1})^2/n_C^*$, so the second term dominates as long as $\lambda_{\min}(\Sigma_0^{-1}) \gg \lambda_{\max}(L^{-1})$, for a fixed n_C^* and $\lambda_{\min}(L^{-1})$. Conversely, the L^{-1} term dominates when ρ_I is sufficiently high, and in these cases, incorporating a stronger study-level prior can increase the variance of our estimates.

These relationships are evident in the second-order expansion of $-V_C(n_C^*, w_C^*)$:

$$
\begin{aligned}
&\frac{-4((\frac{A_T(w_C^*)}{A_S(w_C^*)}-1)\rho_I^2+1)(A_S(w_C^*))^2}{A_T(w_C^*)n_C^*} + \\
&\frac{16A_T(w_C^*)^4\left(-\left(\sigma_{0S}+(\frac{A_T(w_C^*)}{A_S(w_C^*)}-1)\rho_I^2\sigma_{0S}^2\right)^2 + 2\frac{A_T(w_C^*)}{A_S(w_C^*)}\rho_0\rho_I\left(1+(\frac{A_T(w_C^*)}{A_S(w_C^*)}-1)\rho_I^2\right)\sigma_{0S}\sigma_{0T} - (\frac{A_T(w_C^*)}{A_S(w_C^*)})^2\rho_I^2\sigma_{0T}^2\right)}{(A_T(w_C^*))^2(n_C^*)^2(\rho_0^2-1)\sigma_{0S}^2\sigma_{0T}^2} \\
&+ \mathcal{O}\left(\frac{A_S(w_C^*)^3}{{n_C^*}^3}\right).
\end{aligned}
$$

Using the above approximation, we can evaluate the surface on which $d\text{Cost}_C/d\rho_0 = 0$, yielding solutions

$$\rho_0 = \frac{\sigma_{0S} + \left(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1\right)\rho_I^2\sigma_{0S}}{\frac{A_T(w_C^*)}{A_S(w_C^*)}\rho_I\sigma_{0S}} \qquad \text{or} \qquad \rho_0 = \frac{\frac{A_T(w_C^*)}{A_S(w_C^*)}\rho_I\sigma_{0T}}{\sigma_{0S} + \left(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1\right)\rho_I^2\sigma_{0S}}\,. \tag{5}$$

As discussed above, the surface on which $d\text{Cost}_C/d\sigma_{0S} = 0$ is identical, and thus yields the same two solutions as above. The surface on which $d\text{Cost}_C/d\rho_I = 0$ yields solutions

$$\rho_0 = \frac{1}{\left(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1\right)\frac{A_T(w_C^*)}{A_S(w_C^*)}\rho_I\sigma_{0S}\sigma_{0T}}\left(2\frac{A_T(w_C^*)}{n_C^*} - 6\frac{A_T(w_C^*)}{n_C^*}\rho_I^2 + 6\frac{(A_T(w_C^*))^2}{A_S(w_C^*)n_C^*}\rho_I^2 \pm \sqrt{\aleph}\right), \tag{6}$$

where we have defined

$$\aleph = \frac{A_T(w_C^*)}{A_S(w_C^*)}\left(\frac{4A_T(w_C^*)}{A_S(w_C^*)}\left(\frac{A_S(w_C^*)}{n_C^*} + 3(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1)\right)^2 - \left(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1\right)\rho_I^2 \times \right.$$
$$\left.\left(8(\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1)\frac{A_S(w_C^*)}{n_C^*}(1 + (\frac{A_T(w_C^*)}{A_S(w_C^*)} - 1)\rho_I^2)\sigma_{0S}^2 + \frac{A_T(w_C^*)}{A_S(w_C^*)}\left(4\frac{A_T(w_C^*)}{n_C^*} + (1 - \frac{A_T(w_C^*)}{A_S(w_C^*)})\sigma_{0S}^2\right)\sigma_{0T}^2\right)\right).$$

We plot these surfaces in Figure 2 in Section 3.

D.3. Local Optima

First, we examine each corner region of Figure 2 in (ρ_0, ρ_I) space in detail. We do this without a large n approximation, since the large n approximation prescribed in Lemma 4 breaks down when $\rho_0 = 1$.

1. Low study-level correlation and high individual-level correlation $(\rho_0 = 0, \rho_I = 1)$:

$$\left.\frac{d\text{Cost}_C}{d\rho_0}\right|_{\substack{\rho_0=0\\ \rho_I=1}} = -\frac{32\eta_C^*\sigma_{0S}^3\sigma_{0T}^3}{(4\sigma_{0T}^2 + 4\sigma_{0S}^2 + \sigma_{0S}^2\sigma_{0T}^2 A_S n_C^*)^2} > 0\,,$$
$$\left.\frac{d\text{Cost}_C}{d\rho_I}\right|_{\substack{\rho_0=0\\ \rho_I=1}} = \frac{8\eta_C^*\sigma_{0T}^4(4 + \sigma_{0S}^2(q_S - q_T))(4 + \sigma_{0S}^2 A_S n_C^*)}{A_T n_C^*(4\sigma_{0S}^2 + 4\sigma_{0T}^2 + \sigma_{0S}^2\sigma_{0T}^2 A_S n_C^*)^2} < 0\,.$$

This shows our local minimum in the upper left hand corner, since cost is increasing as we move away (increase ρ_0 or decrease ρ_I).

2. High study-level correlation and high individual-level correlation $(\rho_0 = 1, \rho_I = 1)$:

$$\left.\frac{d\text{Cost}_C}{d\rho_0}\right|_{\substack{\rho_0=1\\ \rho_I=1}} = \frac{2\eta_C^*\sigma_{0S}^2\sigma_{0T}^2}{(\sigma_{0S} - \sigma_{0T})^2} < 0\,,$$
$$\left.\frac{d\text{Cost}_C}{d\rho_I}\right|_{\substack{\rho_0=1\\ \rho_I=1}} = \frac{8\eta_C^*\sigma_{0T}^2}{A_T n_C^*(\sigma_{0S} - \sigma_{0T})^2} < 0\,.$$

This shows our local mimimum in the upper right hand corner, since cost is increasing as we move away (decrease ρ_0 or decrease ρ_I).

3. High study-level correlation and low individual-level correlation $(\rho_0 = 1, \rho_I = 0)$:

$$\left.\frac{d\text{Cost}_C}{d\rho_0}\right|_{\substack{\rho_0=1\\ \rho_I=0}} = \frac{2\eta_C^*\sigma_{0S}^2\sigma_{0T}^2 A_S n_C^*(4 + \sigma_{0S}^2 A_S n_C^*)}{(4 + A_T n_C^*(\sigma_{0S}^2 + \sigma_{0T}^2) + \sigma_{0S}^2(A_S - A_T)n_C^*)^2} < 0\,,$$
$$\left.\frac{d\text{Cost}_C}{d\rho_I}\right|_{\substack{\rho_0=1\\ \rho_I=0}} = -\frac{8\eta_C^* A_T n_C^*\sigma_{0S}\sigma_{0T}^3}{(4 + A_T n_C^*(\sigma_{0S}^2 + \sigma_{0T}^2) + \sigma_{0S}^2(A_S - A_T)n_C^*)^2} > 0\,.$$

This shows our local minimum in the upper left hand corner, since cost is increasing as we move away (decrease ρ_0 or increase ρ_I).

4. Low study-level correlation and low individual-level correlation ($\rho_0 = 0, \rho_I = 0$):

$$\left.\frac{d\text{Cost}_C}{d\rho_0}\right|_{\substack{\rho_0=0\\ \rho_I=0}} = 0\,,$$
$$\left.\frac{d\text{Cost}_C}{d\rho_I}\right|_{\substack{\rho_0=0\\ \rho_I=0}} = 0\,.$$

This shows a saddle point in the bottom left hand corner, since cost is unaffected as we move away (increase ρ_0 or increase ρ_I).

Thus, we have established that there are always 3 local minima for the cost of the trial occurring on the corners of Figure 2 except $(0,0)$. Next, we examine the interior. To simplify the analysis, we use our large n approximation described in the previous subsection to solve for

$$\frac{d\text{Cost}_C}{d\rho_0} = \frac{d\text{Cost}_C}{d\rho_I} = 0\,.$$

When we substitute the first solution for ρ_0 from Eq. (5) into Eq. (6), we obtain only one solution $\rho_I = 0$, which is clearly not in the interior. When we substitute the second solution for ρ_0, we obtain the following three solutions:

$$\rho_I = 0 \quad \text{and} \quad \rho_I = \pm\frac{\sqrt{8\frac{A_S(w_C^*)}{n_C^*} - 8\frac{A_T(w_C^*)}{n_C^*} - \frac{A_T(w_C^*)}{A_S(w_C^*)}\sigma_{0T}^2} + \left(\frac{A_T(w_C^*)}{A_S(w_C^*)}\right)^2\sigma_{0T}^2}{2\sqrt{2\left(\frac{A_S(w_C^*)}{n_C^*} - 2\frac{A_T(w_C^*)}{n_C^*} + \frac{(A_T(w_C^*))^2}{n_C^* A_S(w_C^*)}\right)}}\,.$$

The first is again not in the interior. It can be verified that the other solution (only one is feasible, *i.e.*, satisfies $0 \le \rho_I \le 1$) is a local cost maximum. Thus, there is at most one local maximum and no local minima in the interior.

D.4. Proof of Theorem 1

Recall that, for ease of comparison, we fix the patient enrollment n across all three trial types, and further consider a large n approximation.

First, consider Type A vs. Type C trials. Since the cost of Type A trials do not depend on $\rho_I, \rho_0, \sigma_{0S}$ and σ_{0T}, the dependence of $\text{Cost}_A - \text{Cost}_C$ on these parameters is simply the opposite of those in Lemma 3 — *i.e.*, $d(\text{Cost}_A - \text{Cost}_C)/dp$ remains non-monotonic for $p \in \{\rho_I, \rho_0, \sigma_{0S}\}$ and does not depend on $p = \sigma_{0T}$. Since we have fixed patient enrollments n and Type C trials are strictly less constrained than Type A trials, it follows that $w_A^* - w_C^* > 0$, which then implies

$$\frac{d(\text{Cost}_A - \text{Cost}_C)}{dc_w^*} = w_A^* - w_C^* > 0\,.$$

Second, consider Type B vs. Type C trials. Since the cost of Type B trials do not depend on ρ_I, the dependence of $\text{Cost}_B - \text{Cost}_C$ on this parameter is simply the opposite of that in Lemma 3 — *i.e.*, $d(\text{Cost}_B - \text{Cost}_C)/dp$ is non-monotonic for $p = \rho_I$. Again, since we have fixed patient enrollments n and Type C trials are strictly less constrained than Type B trials, it follows that $w_B^* - w_C^* > 0$, which then implies

$$\frac{d(\text{Cost}_B - \text{Cost}_C)}{dc_w^*} = w_B^* - w_C^* > 0\,.$$

It remains to examine parameters relating to our study-level prior Σ_0, *i.e.*, $p \in \{\sigma_{0S}, \sigma_{0T}, \rho_0\}$. Unlike Type A or Type C trials, Type B trials rely critically on the informativeness of the surrogate through the study-level

prior to meet the Type I/II error constraints; as we observe in Section 4, it may not even be feasible to run a Type B trial if the surrogate isn't sufficiently predictive. Consequently, Type B trials rely on Σ_0 parameters to a higher order in n than Type C trials, and thus the dependence of $\text{Cost}_B - \text{Cost}_C$ on these parameters simply inherits those in Lemma 2. Specifically,

$$\begin{aligned}
\frac{d(\text{Cost}_B - \text{Cost}_C)}{d\sigma_{0S}} &= \frac{8\eta_B^*\rho_0^2\sigma_{0T}^2}{nA_S(w_B^*)\sigma_{0S}^3} + O\left(\frac{1}{n^2}\right) < 0\,, \\
\frac{d(\text{Cost}_B - \text{Cost}_C)}{d\sigma_{0T}} &= \frac{-2\eta_B^* nA_S(w_B^*)\sigma_{0S}^2\rho_0^2}{\sigma_{0T}^3(1+\sigma_{0S}^2(1-\rho_0^2)nA_S(w_B^*)/4)} + O\left(\frac{1}{n}\right) > 0\,, \\
\frac{d(\text{Cost}_B - \text{Cost}_C)}{d\rho_0} &= 2\eta_B^*\rho_0\sigma_{0T}^2\left(1 - \frac{4}{nA_S(w_B^*)\sigma_{0S}^2}\right) + O\left(\frac{1}{n^2}\right) < 0\,.
\end{aligned}$$

In other words, while both Type B and Type C trials may benefit from a surrogate that is informative at the study-level, it is more beneficial for a Type B trial than for a Type C trial since Type C trials can also rely on information from true outcomes as well as the informativeness of the surrogate at the individual-level.

D.5. Parameters

Simulation Parameters. The parameters we swept are: $\lambda_T \in \{0.02, 0.04, 0.06\}$, $\sigma_{0S} \in \{0.1, 0.2, 0.4\}$, $\sigma_{0T} \in \{0.1, 0.2, 0.4\}$, and $c_w \in \{10^3, 10^4, 10^5\}$. Keeping with the motivation that the surrogate is easier to measure, we restricted to instances where $\sigma_{0S} \geq \sigma_{0T}$. Other parameters were fixed to be $c_n = 1, \lambda_S = 0.1, \delta = 0.5, \alpha = 0.05, \beta = 0.3, \lambda_E = 1/20$.

Meta-Analyses Data Collection. We obtained the true and surrogate outcome pairs in Figure 4 from 65 meta-analyses in the medical literature. We searched PubMed using the terms "surrogate outcome meta-analysis", which gave us 537 results, and then manually extracted all studies which listed *both* study-level and individual-level correlations to get these final 65 data points. As the literature on meta-analyses and surrogate outcomes expands, one will be able to collect this data on a much larger range of diseases.

Appendix E: Expected Successes

While we focus on clinical trial designs that minimize cost subject to a Type I/II error constraint (which matches the standard FDA decision-making process), a trial planner may alternatively be interested in maximizing *expected successes* (see, *e.g.*, Berry 1972). These designs dynamically select the treatment arm of each patient when they enter the trial, with the goal of maximizing the proportion of patients with a successful outcome. The typical setting is that a binary-valued outcome is observed immediately upon patient enrollment, and a multi-armed bandit approach is applied to determine dynamic treatment allocation that balances between exploration and exploitation.

E.1. Model

We now briefly lay out a strategy for incorporating surrogate and true outcome observations for inference in this setting. To define a success, we must first model binary outcomes. Extending the approach of Renard et al. (2002) to the case with both true and surrogate outcomes, we model binary outcomes using underlying latent continuous values $\tilde{S}_i$ underlying the binary surrogate outcome S_i and a latent continuous value $\tilde{T}_i$ underlying the binary true outcome T_i. In this surrogacy model, we assume known, fixed constants $\boldsymbol{c} := [c_S\ c_T]'$ determine the event rate in the control arm, while treatment effects $\boldsymbol{\mu} := [\mu_S\ \mu_T]'$ control the event rate in the treatment

arm. Similar to our approach in our main model in Section 2, we assume these treatment effects follow a bivariate normal prior distribution $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. The correlation between the latent variables, ρ_I, which we assume is a known constant, induces an individual-level correlation between the observed binary outcomes. We can then posit the following relationship between the latent values, where Z_i is an indicator for whether patient i is in the treatment arm (1) or not (0).

$$\begin{bmatrix} \tilde{S}_i \\ \tilde{T}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} c_S + \mu_S Z_i \\ c_T + \mu_T Z_i \end{bmatrix}, \Sigma_I \right) \qquad \Sigma_I = \begin{bmatrix} 1 & \rho_I \\ \rho_I & 1 \end{bmatrix}$$

Then we can model the probability that each binary outcome takes value one as being equal to the probability that each latent variable is greater than 0:

$$\Pr(S_i = 1) = \Pr(\tilde{S}_i > 0) = \Phi(c_S + \mu_S Z_i) \qquad \Pr(T_i = 1) = \Pr(\tilde{T}_i > 0) = \Phi(c_T + \mu_T Z_i)$$

Keeping with our motivation that the surrogate outcome is faster to measure than the true outcome, we assume that the surrogate outcome is observed immediately while the true outcome is observed with some fixed delay of g time periods. Note that if the true outcome has not yet been measured, we don't yet know its value. Denote this as $T_i = *$.

So at any point in time after they are enrolled, we have observed patient i's surrogate outcome $S_i \in \{0, 1\}$, and true outcome $T_i \in \{*, 0, 1\}$. If we define $n_{ab}^{(i)} := \mathbb{I}[S_i = a, T_i = b]$ for each $a \in \{0, 1\}, b \in \{*, 0, 1\}$, then for each patient exactly one of values is set to 1 and all others are set to 0. Therefore, patient i's current results can be given by

$$\boldsymbol{x}_i = \left[n_{0*}^{(i)} \; n_{1*}^{(i)} \; n_{00}^{(i)} \; n_{01}^{(i)} \; n_{10}^{(i)} \; n_{11}^{(i)} \right]',$$

Furthermore, we can denote the event that a patient experiences a surrogate outcome failure as $s_0^{(i)} = n_{0*}^{(i)} + n_{00}^{(i)} + n_{01}^{(i)}$ and that a patient experiences a surrogate outcome successes as $s_1^{(i)} = n_{1*}^{(i)} + n_{10}^{(i)} + n_{11}^{(i)}$. Likewise, we can denote the event that a patient experiences a true outcome failure as $t_0^{(i)} = n_{00}^{(i)} + n_{10}^{(i)}$ and that a patient experiences a surrogate outcome successes as $t_1^{(i)} = n_{01}^{(i)} + n_{11}^{(i)}$. We also define $p_{ab}(\boldsymbol{\mu})$ as the probability that a treatment group patient experiences a particular outcome under a given value of $\boldsymbol{\mu}$.

We then can calculate an update to the prior distribution based on the results $\boldsymbol{x}$ of multiple patients in the treatment group (since c_S and c_T are assumed to be known, control group patients provide no new information). We define $n_{ab} = \sum_i n_{ab}^{(i)}$, where the sum is over all the treatment group patients i. Then our posterior distribution is

$$posterior(\boldsymbol{\mu} \mid \boldsymbol{x}) \propto \left(\prod_{a \in \{0,1\}, b \in \{0,1,*\}} [p_{ab}(\boldsymbol{\mu})]^{n_{ab}} \right) \phi(\boldsymbol{\mu}, \boldsymbol{\mu}_0, \Sigma_0)$$

Similarly to the main body of this work, we can compare our design to inference using only true outcomes or only surrogate outcomes. For the surrogate outcome only model, we define the count of patients with surrogates outcome a to be $s_a = \sum_i s_a^{(i)}$. Then, our inference is:

$$posterior(\boldsymbol{\mu} \mid \boldsymbol{x}) \propto \left(\prod_{a \in \{0,1\}} [p_{a*}(\boldsymbol{\mu})]^{s_a} \right) \phi(\boldsymbol{\mu}, \boldsymbol{\mu}_0, \Sigma_0)$$

For the true outcome only model, we define the count of patients with true outcome b to be $t_b = \sum_i t_b^{(i)}$. Then, our inference is:

$$posterior(\boldsymbol{\mu} \mid \boldsymbol{x}) \propto \left(\prod_{a \in \{0,1\}} [p_{*a}(\boldsymbol{\mu})]^{t_a} \right) \phi(\boldsymbol{\mu}, \boldsymbol{\mu}_0, \Sigma_0)$$

Now, each time a new patient arrives, we need to determine whether to allocate them to the control arm (known expected reward; no learning) or the treatment arm (unknown expected reward; learning). Here we employ the traditional UCB1 algorithm, using our posterior distribution of $\boldsymbol{\mu}$ to guide whether to allocate to the treatment arm or not. From this posterior distribution, we can find true outcome posterior mean μ_T and variance σ_T^2. We then allocate patient number n to the treatment arm whenever $\mu_T + c\sigma_T\sqrt{\ln n} \geq 0$; otherwise we allocate to the control arm. Once all patients have arrived and all outcomes have been observed, we can measure the number of successes in the trial using the final value of t_1. This algorithm is outlined below in Algorithm 2.

Algorithm 2 Pseudocode for UCB1 algorithm used for expected successes design

Input: prior mean $\boldsymbol{\mu}_0$, prior covariance matrix Σ_0, known event rates in the control arm $\boldsymbol{c}$, and trial design type *trial*, total number of patients N
Initialize: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\Sigma = \Sigma_0$
for $i = 1, 2, \ldots, N$ **do**
    if $\mu_T + c(\Sigma[2,2])\sqrt{\ln i} \geq 0$ **then**
        $Z_i = 1$, *i.e.*, patient is assigned to the treatment arm
        Observe $\boldsymbol{x}_i$ (note that we will only have observed the Surrogate at this time)
        if $i > g$ **then**
            Observe updated $\boldsymbol{x}_{i-g}$ (we will observe the true outcome for patient $i-g$ now)
        end if
        Update posterior mean and covariance matrix $\boldsymbol{\mu}$, Σ with the update for trial type *trial*
    else
        $Z_i = 0$
        Observe $\boldsymbol{x}_i$ (surrogate only)
        if $i > g$ **then**
            Observe updated $\boldsymbol{x}_{i-g}$
            Update $\boldsymbol{\mu}$ and Σ with the update for trial type *trial*
        end if
    end if
end for
for $i = N+1, \ldots N+g$ **do**
    Observe updated $\boldsymbol{x}_{i-g}$
end for
return the number of successful true outcomes observed $\Sigma_{i=1}^{N} t_1^{(i)}$.

E.2. Numerical Simulation

As in the main paper, natural competitor models would be ones for which only the surrogate outcomes or only the true outcomes are used for inference. Since we have already worked out the inference for these two models, it is trivial to use their values for μ_T, σ_T, and $\tilde{\sigma}_T$ (true outcome only) within the bandit algorithm.

It is worthwhile mentioning that the surrogate-only model will eventually end up playing the treatment arm in perpetuity regardless of the data, since σ_T tends to some positive constant as the number of enrolled patients gets large. However, for finite trial size N it can outperform true outcome only in some situations.

We can then numerically demonstrate that taking into account both surrogate and true outcome information can yield better outcomes than simply making decisions based on the surrogate alone or the true outcome alone in this framework as well.

Simulation parameters. Here we set the total number of people to enroll in the trial as $N = 200$ patients. We also set the control arm mean $\boldsymbol{c} = [0\ 0]'$, and the prior mean $\boldsymbol{\mu}_0 = [0\ 0]'$. The parameters we swept are: prior surrogate variance $\sigma_S^2 \in \{0.1, 1, 10\}$ with prior true outcome variance $\sigma_T^2 = \sigma_S^2$, study-level correlation $\rho_0 \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$, individual-level correlation $\rho_I \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$, and true outcome delay $g \in \{10, 25, 50, 75, 100, 125, 150\}$.

Results. Following standard practice, we choose the best bandit tuning parameter value c for each parameter set and trial type. We find, averaged over this set of exogenous parameters, that our proposed trial design provides a 2.7% improvement in the number of expected successes over the next best competitor.

Appendix F: 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 time-to-event true outcomes and their corresponding surrogate outcomes 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 time-to-event 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, we gathered information on the effect sizes of 30 true and surrogate outcome pairs across 19 different cancer subtypes. In all cases, the true outcome was time-to-event, the log overall survival hazard ratio. The surrogates varied; time-to-event surrogate outcomes included log hazard ratios for progression-free survival, time to progression, disease-free survival, and failure-free survival, while binary surrogate outcomes included log odds ratios for overall response rate and disease control rate.

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.

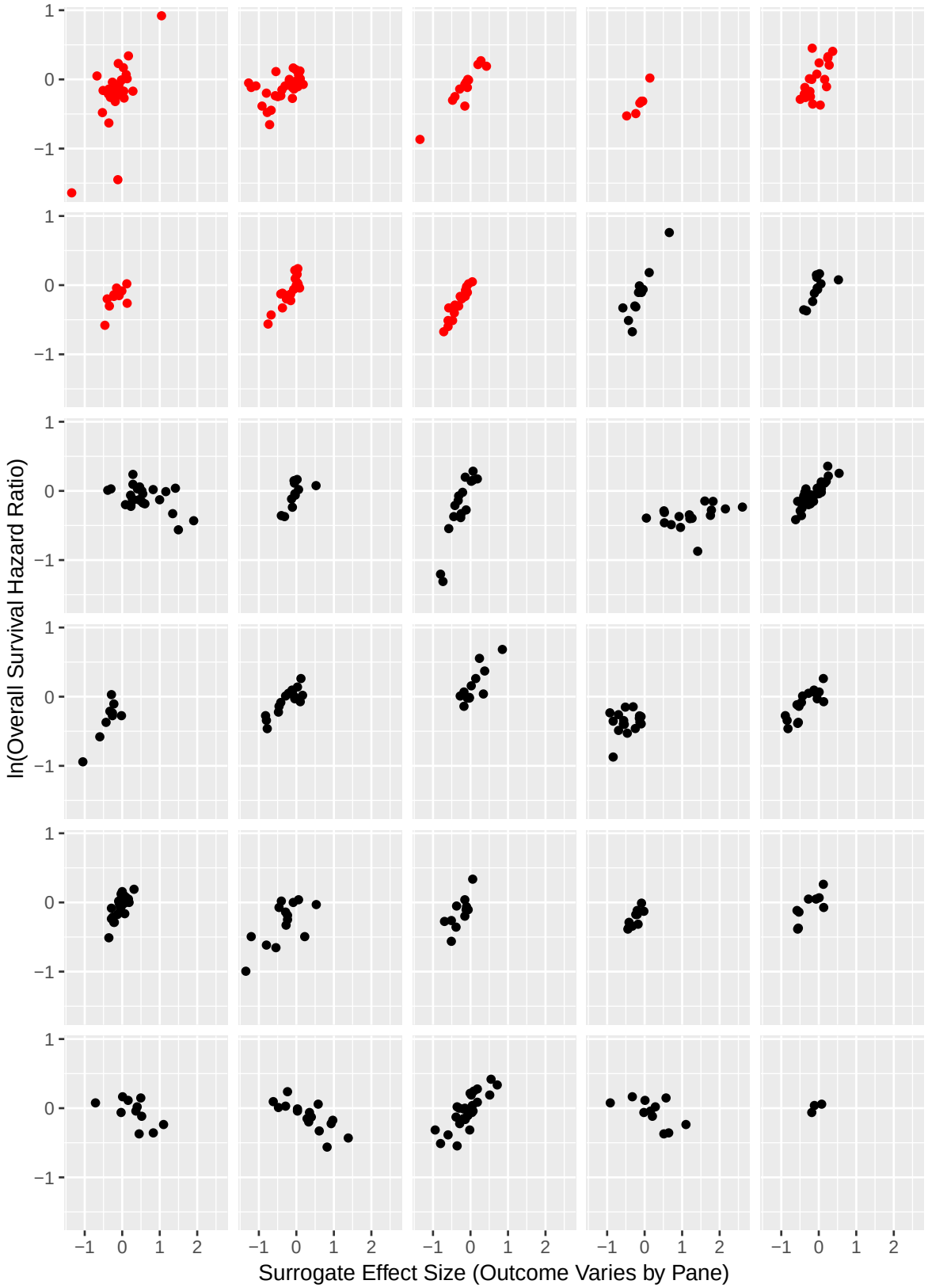


Figure 7 **Surrogate and true outcome effect sizes of cancer studies across 30 surrogate/true outcome pairs from 19 cancer subtypes. The eight surrogate/true outcome pairs with red points reject the null hypothesis of bivariate normality ($p \leq 0.05$), while the remaining 22 with black points do not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